Caenorhabditis Annotation in UniProtKB

Overview

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.

Manually annotated species include:

- Caenorhabditis elegans
- Mus musculus
- Drosophila melanogaster
- Danio rerio
- Xenopus laevis
- Saccharomyces cerevisiae
- Rattus norvegicus

Caenorhabditis Protein Annotation project

The Caenorhabditis protein annotation project mainly focuses on the manual annotation of experimentally characterised Caenorhabditis elegans (C.elegans), proteins, but also contains entries from other Caenorhabditis species including briggsae, drosophilae, remanei and vulgaris.

Focus: Caenorhabditis elegans

The nematode worm, C.elegans, is a transparent soil dwelling roundworm. It is approximately 1mm in length and has a relatively short life cycle and thus it is a versatile model organism for studying gene and protein function in complex biological processes. The use of this model is adopted by thousands of scientists worldwide. It was the first multicellular organism to be sequenced and was also the organism in which RNA interference was first discovered. Such scientific breakthroughs have paved the way for numerous other large scale sequencing and knockout projects in other multicellular organisms. Nevertheless, C. elegans still remains an essential model organism and practical genetic tool in which to study gene and protein function.

Expert Annotation of Caenorhabditis elegans Proteins

To date, there are 3,679 expertly curated protein entries for C. elegans and this number is continuously increasing through on-going annotation. In particular, UniProtKB works closely with both the nematode worm research community and with WormBase, the database of the biology and genome of C. elegans and related nematode species, to ensure that UniProtKB presents detailed and current proteomes, sequences and functional annotation of nematode proteins in a clear, informative and organised manner in order to facilitate and promote further nematode research.

Each comprehensive UniProtKB entry presents primary protein data (including amino acid sequence, gene and protein names and taxonomic data), amalgamated data from a range of sources (including scientific literature, model organism databases and sequence analysis tools) as well as biological ontologies, classifications and cross-references in order to provide an accurate overview:

- **Protein name.** Each entry always provides a recommended protein name, and short names and synonyms are added if relevant. The current gene name and ORF are imported from Wormbase.

Sequences. The displayed amino acid sequence is the canonical protein sequence. Sequences of isoforms are also available. Positional annotation such as data referring to mutagenesis, post-translational modifications and active sites are based on the canonical sequence.

Evidence attribution. The origin of data is indicated by evidence tags which link to the source of the information.

Literature citations. Publications are selected for curation from literature databases.

Cross-references. Links are provided to sources including Wormbase, INSIDC and PRIDE.

Gene ontology: GO. Assigned manually based on experimental data from the literature and are also added based on automatic methods. GO terms are also imported from other GO consortium members.

Annotation: Experimental data from the literature is used to functionally annotate and characterise proteins. Data regarding function, subcellular location, tissue and developmental stage expression, interactions and disruption phenotype are captured.

UniProt is updated every four weeks and can be freely accessed or downloaded from http://www.uniprot.org

Email: help@uniprot.org
URL: www.uniprot.org