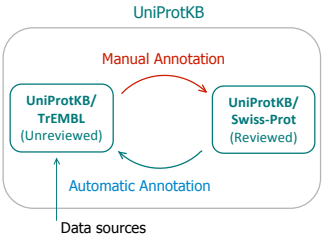
Manual Curation of Vertebrate Proteins in the UniProt Knowledgebase

Michael Gardner[1], Lionel Breuza[2] and UniProt Consortium[1,2,3]

[1]EMBL-European Bioinformatics Institute, Cambridge, UK
[2]Swiss Institute of Bioinformatics, Geneva, Switzerland
[3]Protein Information Resource, Washington DC, USA

1 Introduction

The UniProt Knowledgebase (UniProtKB) aims to provide the scientific community with a consistent and authoritative resource for protein sequence and functional information. UniProtKB consists of two sections – Swiss-Prot and TrEMBL.



2 UniProt Knowledgebase

Swiss-Prot contains manually annotated records with information extracted from the literature and curator-evaluated computational analysis.

TrEMBL contains computationally generated records enriched with automatic annotation and classification.

3 Vertebrate Proteins

Given the importance of human and vertebrate model data in biomedical research, a major focus is the high-quality manual curation of human proteins and their orthologs in other vertebrate species.

The Human Proteome

UniProtKB/Swiss-Prot contains the complete manually reviewed human proteome, comprising approximately 20'300 records. An additional 14'748 isoforms are derived from alternative splicing, alternative initiation and alternative promoter usage. The canonical sequences and additional isoforms are available for download and additional isoforms are also included for Blast searches. The entries are continually updated and reviewed by:

(1) Addition of functional data: Experimental data, including function, regulation, expression, structure, interactions, subcellular location and post-translational modifications are continuously extracted from the literature and used to update entries. This information contributes to a Protein Evidence (PE) score which indicates the level of confidence in the existence of a protein (Figure 1A).

(2) Sequence variation and refinement: All sequences representing individual proteins are collated and variations such as SNPs, isoforms and disease-associated mutations verified (Figure 1B). Dubious isoforms, sequences based on experimental artefacts and protein products derived from erroneous gene model predictions are revisited. This is done in collaboration with HAVANA, Ensembl and HGNC through the Hinxton Sequencing Forum (HSF) as well as through active collaboration with the RefSeq group at the NCBI.

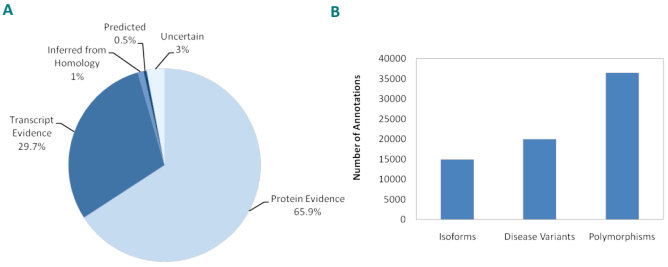


Figure 1: (A) The majority of human protein entries in UniProtKB/Swiss-Prot are supported by evidence at the protein level. (B) UniProtKB/Swiss-Prot captures a large amount of variation in the human proteome.

(3) Standardisation: UniProt is also a member of the Consensus CDS Project (CCDS) and works with other CCDS members to review entries and support convergence across resources.

Vertebrate Proteomes

In addition to the review of the human proteome, other mammalian and non-mammalian vertebrate proteins are increasingly being manually annotated.

(1) Taxonomic coverage: UniProtKB/Swiss-Prot includes an additional 61'000 reviewed entries from model vertebrates such as mouse, rat, apes, cow, pig, chicken, zebrafish and *Xenopus*. In total, UniProtKB/Swiss-Prot contains reviewed entries from 3'086 vertebrate species, making up 50.2% of all eukaryotic proteins in UniProtKB/ Swiss-Prot.

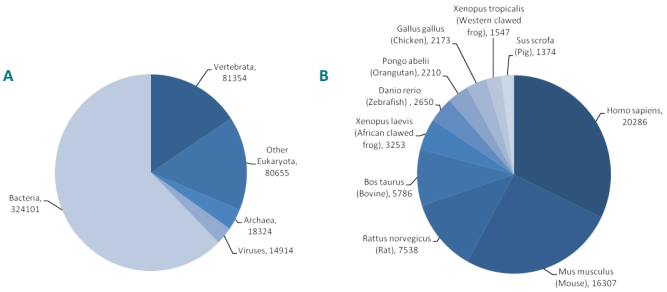


Figure 2: (A) Representation of taxonomic categories in UniProtKB/Swiss-Prot. (B) Proportions of UniProtKB/Swiss-Prot entries from the ten most highly represented vertebrate species.

(2) Manual annotation: The vertebrate dataset contains 187'000 experimentally-derived annotations for topics such as function, regulation, tissue specificity, catalytic activity, pathway, subcellular location and alternative products.

(3) An expanding resource: UniProtKB imports additional sequences from Ensembl in order to expand the vertebrate dataset. Active collaborations with organism-specific databases such as MGI, RGD, Zfin and Xenbase continue to improve the quality of vertebrate sequences.

Summary

Manual curation of vertebrate proteins plays a vital role in providing users with a comprehensive overview of sequence and functional information. Ongoing efforts continue to improve the range and quality of functional annotation and sequences. All data are freely available from www.uniprot.org.

Acknowledgements
UniProt is funded by the European Molecular Biology Laboratory, the US National Institutes of Health, the European Union and the Swiss Federal Government.

European Bioinformatics Institute (EMBL-EBI)
Swiss Institute of Bioinformatics (SIB)
Protein Information Resource (PIR)
Email: help@uniprot.org
URL: www.uniprot.org