Introduction to Protein Databases

一月 26, 2013, 3:03 a.m. by Rosalind Team **Topics**: Bioinformatics Tools, Proteomics

Four Commonly Used Protein Databases

Figure 1. Screenshot of the UniProt homepage.

Proteins are identified in different labs around the world, and data about them is gathered into freely accessible databases. A central repository for protein data is **UniProt**, a comprehensive high-quality database established by an international consortium. UniProt provides detailed protein annotation, including function description, domain structure, and post-translational modifications. It also supports protein similarity search, taxonomy analysis, and literature citations. See **Figure 1** for a screenshot of the UniProt homepage.

UniProt is part of ExPASy, the resource portal of Swiss Institute of Bioinformatics. ExPASy provides access to numerous databases and software tools in different areas of the life sciences.

The most important component of UniProt is the UniProt Knowledgebase, or UniProtKB (see the help page here), a protein database partially curated by experts. UniProtKB comprises two major parts:

- Swiss-Prot: a manually annotated and reviewed, non-redundant protein sequence database. Each Swiss-Prot entry holds all relevant information about a particular protein, including data taken from scientific literature data and the results of computational analysis of the protein.
- TrEMBL (stands for "Translated EMBL"): an automatically annotated database that is not reviewed. TrEMBL is not a true protein database; instead, it holds translated versions of nucleic acid sequences taken from multiple sources, including the European Molecular Biology Laboratory (EMBL) database, worldwide genome sequencing projects, and the coding regions of genes accessed from GenBank.

The National Center for Biotechnology Information (NCBI) also maintains protein data. Data equivalent to that of Swiss-Prot is part of the NCBI's **RefSeq** database (http://www.ncbi.nlm.nih.gov/RefSeq/), a curated collection of genetic sequences. RefSeq records are annotated by NCBI personnel, and they provide reliable information for genomic DNA along with RNA transcribed from DNA and the corresponding translated proteins. NCBI's equivalent project to TrEmbl is the NCBI Protein database, which is part of the larger GenBank database and holds unannotated protein sequences. If a nucleotide sequence contained in GenBank codes for protein, then the corresponding amino acid sequence is automatically annotated and included in the NCBI database with its own protein ID. The NCBI databases also recognize UniProt IDs as search terms.

Because Swiss-Prot annotation provides so much information, NCBI protein records usually provide links to corresponding Swiss-Prot entries whenever possible.

The NCBI Protein Database can be found here.

Problem

The UniProt Knowledgebase can be found here.

You can see a complete description of a protein by entering its UniProt access ID into the site's query field. Equivalently, you may simply insert its ID (`uniprot_id`) directly into a UniProt hyperlink as follows:

```
http://www.uniprot.org/uniprot/uniprot_id
```

For example, the data for protein B5ZC00 can be found at http://www.uniprot.org/uniprot/B5ZC00.

Swiss-Prot holds protein data as a structured .txt file. You can obtain it by simply adding `.txt` to the link:

```
http://www.uniprot.org/uniprot/uniprot_id.txt
```

Given: The UniProt ID of a protein.

Return: A list of biological processes in which the protein is involved (biological processes are found in a subsection of the protein's "Gene Ontology" (GO) section).

Sample Dataset

```
Q5SLP9
```

Sample Output

```
DNA recombination
DNA repair
DNA replication
```

Programming Shortcut

ExPASy databases can be accessed automatically via Biopython's `Bio.ExPASy` module. The function `.get_sprot_raw` will find a target protein by its ID.

We can obtain data from an entry by using the `SwissProt` module. The `read()` function will handle one SwissProt record and `parse` will allow you to read multiple records at a time.

Let's get the data for the B5ZC00 protein:

```
>>>from Bio import ExPASy
>>>from Bio import SwissProt
>>>handle = ExPASy.get_sprot_raw('B5ZC00') #you can give several IDs separate
d by commas
>>>record = SwissProt.read(handle) # use SwissProt.parse for multiple protein
s
```

We now can check the list of attributes for the obtained Swiss-Prot record:

```
 >>> dir(record)
 [..., 'accessions', 'annotation_update', 'comments', 'created', 'cross_refer
ences',
 'data_class', 'description', 'entry_name', 'features', 'gene_name', 'host_or
ganism', 'host_taxonomy_id', 'keywords',
 'molecule_type', 'organelle', 'organism', 'organism_classification', 'refere
nces', 'seqinfo', 'sequence',
 'sequence_length', 'sequence_update', 'taxonomy_id']
```

To see the list of references to other databases, we can check the `.cross_references` attribute of our record:

```
>>>record.cross_references[0]
('EMBL', 'CP001184', 'ACI60310.1', '-', 'Genomic_DNA')
```

Congratulations You solved this problem (attempt #2). Now you may like to try problem "New Motif Discovery".