UniProt: Exploring protein sequence and functional information

Michele Magrane [1], Sangya Pundir [2]

- Proteins
- Beginner
- 3 hours

This course is a guide to the UniProt resource, including what kinds of information it provides and how to access the data using the UniProt website.

Some knowledge of biology and proteins is required. Some bioinformatics knowledge is useful but not essential.

This course was updated in April 2018.

Learning objectives:

- Evaluate what UniProt is and what you can do with it
- Explore the wide range of information provided by UniProt and understand where it comes from
- Be able to access, navigate and search on the UniProt website
- Evaluate the tools provided on the UniProt website
- Comprehend how to access completely sequenced proteomes
- Learn how to download data and know what formats are provided

What is UniProt?

The Universal Protein Resource (UniProt [3]) is a comprehensive resource for protein sequence and functional data (Figure 1).
An overview of the databases that comprise UniProt.

---

**The UniProt databases**

There are three UniProt databases:

- The UniProt Knowledgebase (UniProtKB);
- The UniProt Reference Clusters (UniRef);
- The UniProt Archive (UniParc).

**UniProtKB**

UniProtKB is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. It consists of two sections:

- **Reviewed (Swiss-Prot)** - contains manually annotated records;
- **Unreviewed (TrEMBL)** - contains computationally analysed records.

A subset of UniProtKB entries also form the **Proteomes** dataset. This consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

**UniRef**

UniRef provides clustered sets of sequences from UniProtKB (including additional isoforms contained within UniProtKB/Swiss-Prot records) and selected UniParc records. UniRef offers complete coverage of the sequence space at three resolutions:
- The **UniRef100** database combines identical sequences from any organism into a single UniRef entry, displaying the sequence of a representative protein, the accession numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records.
- **UniRef90** is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to the longest sequence (the seed sequence) of the cluster.
- **UniRef50** is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to the longest sequence in the cluster.

**UniParc**

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. UniParc stores each unique sequence only once, giving it a stable and unique identifier (UPI).

Find about more about the databases on the UniProt help page [4].

**Where does the data come from?**

UniProt provides both sequence data and associated functional information, derived from a range of sources (Figure 2).
Sequence data

UniProtKB sequences

More than 95% of the protein sequences provided by UniProtKB come from the translations of coding sequences (CDS) submitted to the ENA/GenBank/DDBJ nucleotide sequence resources of the International Nucleotide Sequence Database Collaboration (INSDC).

These CDS are either generated by gene prediction programs or are experimentally proven. The translated CDS sequences are automatically transferred to the TrEMBL section of UniProtKB. The TrEMBL records can be selected for further manual annotation and then integrated into the UniProtKB/Swiss-Prot section.

In addition to translated CDS, UniProtKB protein sequences may come from:

- The PDB database;
- Sequences experimentally obtained by direct protein sequencing and submitted to UniProt;
- Sequences scanned from the literature;
- Sequences derived from gene prediction but which have not been submitted to ENA/GenBank/DDBJ.

These are imported from resources such as Ensembl and RefSeq.

Importing and combining sequences from a range of sources means that UniProt provides a complete collection of protein sequences and contributes to consistency of protein sets across various sequence resources (Figure 3).

![Figure 3. UniProt imports sequences from a range of sources to ensure that you have access to a complete collection of protein sequences.](image)

UniParc sequences
UniParc is designed to capture all publicly available protein sequence data and contains all the protein sequences from the main publicly available protein sequence databases. A complete list of the source databases is available on the [UniProt website](https://www.ebi.ac.uk/training/online) [12].

**UniRef sequences**

UniRef provides clustered sets of sequences from UniProtKB and selected UniParc records.

**Functional information**

Functional information is found in the UniProt Knowledgebase. UniProt attempts to attach as much functional information as possible to each protein sequence to provide users with an overview of the available information for a given protein. This information is added manually by the UniProt biocurators who are all trained biologists or added automatically through various annotation systems which have been developed within the group (Figure 4).

**Figure 4.** Relationship between manual curation and automatic annotation in the UniProt Knowledgebase (UniProtKB).

**Manual curation**

Manual curation consists of a critical review of experimental and predicted data for each protein as well as manual verification of each protein sequence.

Curation methods applied include:

- manual extraction and structuring of information from the literature
- manual verification of results from computational analyses
- mining and integration of large-scale data sets
- continuous updating as new information becomes available

You can find more information about the manual curation process on the [UniProt website](https://www.ebi.ac.uk/training/online) [13].
Automatic annotation

UniProt has developed two prediction systems to automatically annotate UniProtKB/TrEMBL in a scalable manner with a high degree of accuracy:

1. UniRule is a collection of manually curated annotation rules which define annotations that can be propagated based on specific conditions
2. The Statistical Automatic Annotation System (SAAS) is an automatic decision-tree based rule-generating system

You can find out more information about the automatic annotation pipeline on the [UniProt website](https://www.uniprot.org) [14].

Data evidence

Indicating data origin

The information in a UniProt Knowledgebase (UniProtKB) record comes from a range of different sources. To make it easy to tell where the data have come from, the origin of each piece of information presented in an entry is provided. UniProt makes use of a subset of evidence codes from the Evidence Code Ontology [15] (ECO [16]) to indicate data origin. These ECO codes are shown directly in the text version (also known as flat file version) of the entries. On the UniProt website, they are transformed into user-friendly, easy to understand labels and evidences that are used in manual assertions are coloured gold while those that are used in automatic assertions are coloured blue.

Some evidence examples

1. Experimental data
   If a piece of information has been experimentally shown in a paper, this will be indicated with the details of the paper used (Figure 5).

   ![Figure 5](image.png)

   **Figure 5.** Function section of UniProtKB entry D9N129 (*Caenorhabditis elegans* wdr-20) showing the paper from which the data have been extracted.

2. Data copied from an experimentally characterised protein
   For information which has been transferred from a related experimentally characterised protein, the accession [17] number of the characterised protein is provided (Figure 6).
Figure 6. UniProtKB entry D3DJ41 (Hydrogenobacter thermophilus cfiA) showing the accession number of the entry from which the modified residue has been transferred.

3. Imported data

If information has been imported from another database, the database name and identifier of the entry from which the information has been imported are provided (Figure 7).

Figure 7. UniProtKB entry E9Q4Z2 (mouse Acacb) showing that the gene name has been imported from the Mouse Genome Informatics (MGI) resource.

4. Predicted data

Information which has been predicted by the UniProtKB automatic annotation system or by the sequence analysis programs that are used during the manual curation process are linked to their original source (Figure 8).

Figure 8. UniProtKB entry Q9VIC9 (Drosophila melanogaster CG8665) showing that the protein family information has been predicted by automatic annotation.

You can find out more about evidence codes in UniProtKB on the UniProt website [18].
Why do we need UniProt?

Understanding protein function is critical to research in many areas of science such as biology, medicine and biotechnology. As the number of completely sequenced genomes continues to increase, huge efforts are being made in the research community to understand as much as possible about the proteins encoded by these genomes. This work is generating large amounts of data which are spread across multiple locations including scientific literature and many biological databases.

Keeping up with all of this information is a daunting task for most researchers. UniProt helps with this in the following ways:

- it provides an up-to-date, comprehensive body of protein information at a single site;
- it aids scientific discovery by collecting, interpreting and organising this information so that it is easy to access and use;
- it saves researchers countless hours of work in monitoring and collecting this information themselves;
- it provides tools to help with protein sequence analysis;
- it provides links to related information in more than 150 other biological databases to help you access additional information in more specialised collections.

Frequent updates

UniProt is continuously updated for new data and is released every four weeks.

When to use UniProt

You can perform many different tasks using UniProt including the following (Figure 9):
When not to use UniProt

UniProt cannot help you if you:

- Want to buy any kind of biological product. For this, you need to access the online catalogues of the many suppliers of biochemicals and reagents for the life sciences.
- Need medical advice. UniProt staff are biologists and biochemists who are not trained to give medical advice.

How to access and navigate the UniProt website

The UniProt website

On the UniProt website [3], you can navigate to and search various datasets and also navigate to and use various tools.

Navigating the UniProt datasets
UniProt has a number of datasets that you can navigate to and search within. You can click on the dropdown menu to the left of the search box to see these datasets and select the one you are interested in (Figure 10A). UniProtKB is selected by default.

Figure 10. Accessing the UniProt datasets via (A) the drop down menu or (B) the tiles on the homepage

You can also see these datasets as tiles on the home page (Figure 10B). Clicking on one of these tiles will take you to the entire dataset, where you can explore its contents or search within them.

Navigating the UniProt tools

UniProt provides four main tools:

- The Basic Local Alignment Search Tool (BLAST) for sequence search;
- the 'Align' multiple sequence alignment tool;
- the 'Retrieve/ID Mapping' tool where you can submit a list of identifiers to retrieve the corresponding UniProt entries, or map them from or to an external database;
- the 'Peptide search' tool which allows you to submit short peptide sequences of at least 3 residues and find all UniProtKB sequences which have an exact match to the query sequence.
You can navigate to these tools by clicking on their corresponding links in the header or in the footer (Figure 11).

Figure 11. UniProt tools can be accessed from links on the header (A) and footer (B) of every page.

**How to search UniProt**

You can use the search bar in the UniProt banner at the top of all pages to search the various UniProt datasets.

There is a drop-down list to the left of the search bar that allows you to select a data set. There is also an advanced search option on the right hand side of the search bar (Figure 12).

Figure 12. The UniProt search bar.

To search one of the UniProt datasets, proceed as follows:

1) Select the appropriate data set (the default selection is UniProtKB).

2) Type in your query.

3) Hit the search button.

Note that the background color around the search field changes depending on the data set, to remind you which data set is selected.

**Advanced search**
Advanced search

You can also access advanced search options by clicking on ‘Advanced’, e.g. to restrict terms to specific fields in advance or to combine multiple terms using boolean [19] logic. Then click the search button (the “looking glass” icon) to run the query.

If you click on ‘Advanced’ and then on the first dropdown, you will see all field options (Figure 13):

![Advanced search field options](image13)

**Figure 13.** Advanced search field options.

You can select the field of your choice. Some fields have further options under them. Depending on the chosen data set and field, you can enter some text or choose values from a drop-down list. You can enter many parameters through the panel (Figure 14):

![Advanced search example with two field selections](image14)

**Figure 14.** Advanced search example with two field selections.

To delete an entire row, click on the bin icon next to it. To add a third row of fields, click on the ‘+’ button on the right hand side (Figure 15).
Figure 15. Advanced search example with three field selections.

Once you have completed entering your search parameters, click on the search button.

Exploring the UniProtKB results page

When you do a search within UniProtKB, you will see a page showing all your results (Figure 16).

Figure 16. The UniProtKB results page for insulin.

The UniProtKB results page is also the template for all datasets within UniProt and the following features can be found across all results pages.

Table columns

The results table shows information such as Entry accession, Entry name, Reviewed or Unreviewed status, Protein names, Gene names, Organism and Sequence length. You can customise your results table by editing the
COLUMNS to add or remove information.

Filters and views

You can filter your results by Reviewed Swiss-Prot [20] or Unreviewed TrEMBL status, popular organisms and other criteria depending on your search.

You can explore alternative views of your data. For example, the 'view by taxonomy' shows you a taxonomy tree of all organisms found in your results set. The number of results per organism is shown in brackets next to the organism name. Clicking on the number takes you to that subset of your search results in the UniProtKB results view. Clicking on the organism name takes you to a taxonomy page describing the organism and its lineage.

You can also see your results mapped to UniRef sequence clusters with various identity levels.

Action buttons

You can select an entry to run a direct BLAST sequence similarity search or select multiple entries to do an alignment. You can download entries in various formats and also save them for later by adding them to your basket. The selections in your basket remain until you delete them or clear your cookies. If you have customised your results, you can share a URL of your customised table using the Share action button.

Using the UniProt basket

You can add proteins to your basket from search results pages for the UniProtKB, UniRef or UniParc datasets. Simply select them in the results table and click on the 'Add to basket' button (Figure 17).

Figure 17. UniProtKB results page with two entries selected.

You can also add proteins to your basket from the protein entry pages, using the 'Add to basket' button towards the top of the page. Once you add entries to your basket, you will see the number on the basket changing to show that new proteins have been added. Clicking on the basket button will show your saved proteins (Figure 18).
Figure 18. Basket with the 'UniRef' tab selected.

There is a limit of 400 entries in the basket, so be careful when trying to add large datasets. The contents of the basket will remain there until you delete your browser’s cookies or clear the basket yourself.

Exploring a UniProtKB entry

Navigating within a UniProtKB entry

Information in a UniProtKB entry is structured into a number of sections to make it easy to find the data that you are looking for.

When viewing a UniProtKB entry, the menu bar on the left-hand side of the screen lists display options in the top left-hand corner, allowing you to move easily between the full entry, publications, a graphical view of sequence features (Feature viewer) or sequence features in a table (Feature table). When using the entry view, the left-hand bar below the display options lists the entry sections so you can easily jump to specific sections. You can select or unselect the sections that you are interested in (Figure 19).
Display options

The following display options are provided for each UniProtKB entry.

**Entry**
- Full entry showing all available sections

**Publications**
- Publications related to the entry sourced from UniProtKB curation as well as computationally mapped from other sources

**Feature viewer**
- Graphical representation of all UniProtKB sequence features for the entry in an interactive display plus additional sequence features from large-scale projects

**Feature table**
- Tabular representation of all UniProtKB sequence features for the entry

Entry sections

The sections of a UniProtKB entry and the kinds of data they contain are listed below. Some of these will be described in more detail in the following pages.

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Function</strong></td>
<td>Any useful functional information about the protein</td>
</tr>
<tr>
<td><strong>Names &amp; Taxonomy</strong></td>
<td>Information about the protein and gene name(s) and synonym(s) and about the source organism</td>
</tr>
<tr>
<td><strong>Subcellular location</strong></td>
<td>Information on the location of the mature protein in the cell.</td>
</tr>
<tr>
<td><strong>Pathology &amp; Biotech</strong></td>
<td>Information on the disease(s) and phenotype(s) associated with the deficiency of a protein</td>
</tr>
<tr>
<td><strong>PTM / Processing</strong></td>
<td>Describes post-translational modifications (PTMs) and/or processing events.</td>
</tr>
<tr>
<td><strong>Expression</strong></td>
<td>Information on the expression of a gene at the mRNA or protein level in cells or in tissues of multicellular organisms</td>
</tr>
<tr>
<td><strong>Interaction</strong></td>
<td>Information on the quaternary structure of a protein and on interaction(s) with other proteins or protein complexes</td>
</tr>
</tbody>
</table>
Section | Description
--- | ---
**Structure** | Information on the tertiary structure of a protein
**Family & Domains** | Information on sequence similarities with other proteins and the domain(s) present in a protein
**Sequence** | Displays by default the canonical protein sequence and upon request all isoforms described in the entry. Also includes information pertinent to the sequence(s), including length and molecular weight
**Cross-references** | Provides links to information related to entries and found in data collections other than UniProtKB
**Publications** | Contains the literature citations that are the sources of data used to annotate the entry
**Entry information** | Provides some general information about the entry such as the accession number
**Miscellaneous** | Information that does not fit in any other defined sections
**Similar proteins** | Provides links to the UniProt Reference Clusters (UniRef)

### Annotation score

**What is the annotation score?**

The annotation score provides a measure of the annotation content of a UniProtKB entry or proteome [21]. It is a 5-point heuristic [22] scoring system where a score of 5 is associated with the best-annotated entries, and a score of 1 with entries with basic annotation.

**Where can I find the annotation score?**

The annotation score of an entry can be found in the 'Status' line near the top of each entry (Figure 20).

**Figure 20.** Annotation score of UniProtKB entry Q99741 (human CDC6).

### Adding annotation scores to search results

You can add annotation scores to your search results table through the ‘Columns’ button (Figure 21).
How are annotation scores used?

There are several contexts in which annotation scores can be used:

- **UniProtKB** - the annotation scores can help you to get a quick idea of the relative level of annotation of the entries in your search results.
- **UniRef** - UniProt uses annotation scores to select the representative member of a UniRef cluster.
- **Reference proteomes** - UniProt uses annotation scores to assist with the selection of reference proteomes [23].

How are annotation scores calculated?

- Different UniProtKB annotation types are scored either by presence or by number of occurrences. Annotations with experimental evidence score more highly than equivalent predicted annotations, thereby favoring expert literature-based curation over automatic annotation.
- The score of an individual entry is the sum of the scores of its annotations.
- The score of a proteome is the sum of the scores of the entries that are part of the proteome.

Function

Information provided in the Function section
This section of a UniProt entry provides a wide range of functional information about the protein including a general description of its function as well as other related information such as Gene Ontology [24] terms and functional sites or regions within the protein sequence (Figure 22).

Figure 22. Function section of UniProtKB entry Q93560.

Function subsections

The information in the function section is divided into a number of different subsections. The current subsections and their content can be found here [25].

Subcellular location

The Subcellular location section provides information on the location and the topology of the mature protein in the cell. The section presents a visualisation with one tab containing protein location data from UniProt (both expertly curated data from literature and rule-based automatic annotation) and another one with imported data from Gene Ontology (GO) annotation (Figure 23). The section also provides any UniProt keywords related to subcellular location, if available.
Figure 23. Subcellular location section of UniProtKB entry P35670 for the human copper-transporting ATPase 2 protein.

Pathology and Biotech

Information provided in the Pathology and Biotech section

The Pathology and Biotech section of a UniProt entry provides information on diseases and phenotypes associated with protein deficiencies or disruption of protein-coding genes (Figure 24). It also includes a range of other information including naturally occurring amino acid variants, effects of mutagenesis studies, and uses of proteins in the biotechnology industry or as pharmaceutical drugs. Links to organism-specific databases are also included here.
Pathology & Biotech

Involvement in disease

Inflammatory bowel disease 13 (IBD13) [MIM:612244]: A chronic, relapsing inflammation of the gastrointestinal tract with a complex etiology. It is subdivided into Crohn disease and ulcerative colitis phenotypes. Crohn disease may affect any part of the gastrointestinal tract from the mouth to the anus, but most frequently it involves the terminal ileum and colon. Bowel inflammation is transmural and discontinuous; it may contain granulomas or be associated with intestinal or perianal fistulas. In contrast, in ulcerative colitis, the inflammation is continuous and limited to rectal and colonic mucosal layers; fistulas and granulomas are not observed. Both diseases include extraintestinal inflammation of the skin, eyes, or joints. Note: Disease susceptibility is associated with variations affecting the gene represented in this entry.

PTM / Processing

Information provided in the PTM / Processing section

This section of a UniProt entry describes post-translational modifications (PTMs) and/or processing events which affect a protein (Figure 25). It also provides links to specialised proteomics [27] and PTM databases where you can access additional information.
PTM / Processing subsections

The information is contained in a number of different subsections. The current subsections and their content can be found here [28].

Interaction

The Interaction section of a UniProt entry is divided into a number of subsections:

- the ‘Subunit structure’ section describes the quaternary structure of a protein as well as interactions with other proteins or protein complexes;
- the ‘Binary interactions’ section provides a quality-filtered set of binary interactions automatically derived from the IntAct [29] database. These are presented in a visual matrix, which shows blue dots for interactions. The matrix shows interactions between the protein and its interactors as well as those between the interactors. The depth of the blue colour shows the number of experiments providing evidence for the interaction. The interactions can be filtered by subcellular [30] location or disease involvement of one of the interactors;
- the ‘GO - Molecular function’ section shows any interaction-related Gene Ontology [24] (GO) terms available;
- The ‘Protein-protein interaction databases’ section contains links to a number of specialised protein-protein interaction databases where you can find more detailed interaction information (Figure 26).
Expression

In this section of a UniProt entry you will find information on expression at the mRNA or protein level in cells or tissues of multicellular organisms (Figure 27). It details the tissues where expression has been detected and describes how expression varies according to the stage of cell, tissue or organism development. It describes the effects of environmental factors on expression, and also provides links to a number of gene expression databases.

Figure 27. Expression data from UniProtKB entry Q15004 (human PCNA-associated factor).
Family and Domains

Information provided in the Family and Domains section

This section of a UniProt entry describes which family a protein belongs to as well as providing details of domains present within the protein (Figure 28). It also contains links to family and domain databases and phylogenomic resources.

**Figure 28.** Family & Domains section of UniProtKB entry Q52107.

Family and Domains subsections

The information in the Family and Domains section is divided into a number of different subsections. The current subsections and their content can be found [here](#) [31].

Cross-references

Linking to other resources
Cross-references are provided to more than 150 databases to allow you to access related information in other data collections. A single entry can have cross-references to many different databases and have several hundred individual links (Figure 29). The full list of databases cross-referenced in UniProtKB is available here [32].

Figure 29. Subset of the cross-references available in UniProtKB record Q26261 (unc-5 from Caenorhabditis elegans).

Cross-reference categories

The databases are categorized as follows to group related database types together so that it is easy to find the kind of data you are looking for and to understand how the different databases relate to UniProtKB and to each other:
• 2D gel databases;
• 3D structure databases;
• Chemistry;
• Enzyme and pathway databases;
• Family and domain databases;
• Gene expression databases;
• Genome annotation databases;
• Ontologies;
• Organism-specific databases;
• Phylogenomic databases;
• Polymorphism [33] databases;
• Proteomic databases;
• Protein-protein interaction databases;
• Protein family [34] databases;
• Protocols and materials databases;
• PTM databases;
• Sequence databases;
• Other.

Publications

The 'Publications' link under the 'Display' section on the top left-hand side of the page provides access to all references that have been used as sources of data to curate an entry, along with information about what the article has been cited for. It also provides access to additional computationally mapped references, which have not been curated by UniProt but have been imported from other resources. Next to each imported article, you will find the name of the resource from where the reference was imported.

There are filters provided on the left-hand side of the page to view only references that are from UniProtKB or only those that are computationally mapped. There are also filters based on biological categories as well as large or small-scale studies.

Figure 30. Citations listed in UniProtKB entry P21796 (human voltage-dependent anion-selective channel protein 1).
Sequence Features

Sequence annotation (Features)

Sequence annotations describe regions or sites of interest in the protein sequence, such as post-translational modifications, binding sites, enzyme active sites, local secondary structure \[35\] or other characteristics.

UniProtKB sequence features

UniProtKB provides a number of sequence features. The following table shows the features in UniProtKB and also in which entry subsection they can be found (Figure 31).
Features imported from large-scale studies

UniProt imports some sequence features from large-scale studies. Variants are imported from the 1000 Genomes project, the Exome Aggregation Consortium (ExAC), the Exome Sequencing Project (ESP) and ClinVar. Proteomics peptides and modifications are imported from the Encyclopedia of Proteome Dynamics (EPD), PeptideAtlas and MaxQuant. These are not present in the default UniProtKB entry view but can be accessed via the Feature Viewer and also can be downloaded from the UniProt FTP site.

Exploring the Protein Feature Viewer

Data visualisation and integration have become increasingly important as biological data continues to grow at exponential rates. The protein feature viewer in UniProt presents protein sequence features in one compact view using a highly interactive BioJS component. It allows you to see different types of protein sequence features such as domains, sites, post-translational modifications, topology, mutagenesis, proteomics-derived peptides, antigenic sequences and natural variants from multiple sources in a single view.

Similar to genome viewers, the UniProt feature viewer displays information on tracks. Each track presents a...
biological category of features, and can be expanded into sub-tracks for a more detailed overview. The UniProt feature viewer is available for every UniProtKB protein entry. Under the "Display" heading, follow the "Feature viewer" link (Figure 32).

Figure 32. Feature Viewer link in UniProtKB entries.

Keep it in sight

Showing all colocalised sequence features in one view makes it easier to spot patterns and make inferences. Figure 33 shows an example of how you might use the feature viewer. To investigate the involvement of human lipoprotein lipase (UniProt P06858) in lipoprotein lipase deficiency, a rare autosomal recessive lipid [36] disorder, you can click on the "Domain & sites" label and see that there are three active sites. If you want to investigate the potential effect of variants on these sites, click on the "Variants" label to open up the track. You can zoom into the active sites and click on them to see all their positions highlighted across all protein features. This unique functionality lets you view possible variant [37] implications at a glance (Figure 33).
Figure 33. Feature viewer example with Domain and Sites section expanded.

**A view on variants**

Figure 34 shows how you might see whether a variant plays a role in disease. Clicking on the second active site [38] at position 183 highlights three disease variants in the same position. This shows that the variants are impacting the active site, indicating that they play a role in the disease. Clicking on the variant dots opens up an information box with links to relevant publications (Figure 34).
Figure 34. Feature Viewer example highlighting natural variants colocated with an active site.

**Want to integrate the feature viewer into your website?**

The UniProt feature viewer can be integrated into any website. You can choose to keep all available tracks or only those most relevant to you.

If you would like to include the feature viewer in your own website or resource, you can find instructions in our technical documentation here: [http](http://ebi-uniprot.github.io/ProtVista/developerGuide.html) [40].

**Exploring the feature table**

In the UniProtKB entry view, sequence features are split between the different entry sections by topic. Sections also contain relevant comments and cross-references. However, if you would like to see all sequence features together in one table, you can do so using the "Feature Table" option. This option is available for every UniProtKB protein entry. Under the "Display" heading, follow the "Feature table" link (Figure 35).
Figure 35. The Feature Table option in UniProtKB protein entries displaying all positional features in a table format.

**How to use UniProt tools**

UniProt provides four tools to analyse protein data. All four can be accessed from links in the header (Figure 36). They are:

1) **'BLAST'** (Basic Local Alignment Search Tool) for sequence similarity searching;

2) ‘Align’ for multiple sequence alignment;

3) ‘Retrieve/ID Mapping’ for using a list of identifiers to retrieve batches of UniProtKB entries and to convert database identifiers from UniProt to external databases or vice versa;

4) ‘Peptide search’ tool for submitting short peptide sequences of at least 3 residues and finding all UniProtKB sequences which have an exact match to the query sequence.

Figure 36. The UniProt tools can be accessed from the header on each page.

'BLAST' sequence similarity searching
Performing a Blast search...

1. Select the 'Blast' tab of the toolbar at the top of the page to run a sequence similarity search with the Blast program.

2. Enter either a protein or nucleotide sequence or a UniProt identifier into the form field (Figure 37).

3. Click the 'Run Blast' button.

Figure 37. The Blast input page.

Blast searches can be run directly from the 'Blast' button in UniProt entry pages. All relevant results pages (such as UniProtKB, UniRef, UniParc and tool results) allow you to run a Blast search directly by selecting an entry using a checkbox. You can also run Blast searches from within the 'Basket'.

The following kinds of UniProt identifiers are supported:

- **P00750**: UniProtKB entry
- **P00750-2**: UniProtKB entry isoform sequence
- **P00750[1-20]**: Part of UniProtKB entry sequence, from its 1st to 20th amino acid residue (inclusive)
- **A4_HUMAN**: UniProtKB entry name
- **UPI0000000001**: UniParc entry
- **UniRef100_P00750**: UniRef entry

If you select the 'Blast' tab of the toolbar from a UniProtKB, UniRef or UniParc entry page, the current sequence is prefilled in the form.

Jobs have unique identifiers, which (depending on the job type) can be used in queries (e.g. to get the intersection...
of two sequence similarity searches). Job identifiers and the related data are kept for seven days, and are then deleted.

Optional settings

A list of optional settings for your Blast search can be found in the help section [43].

Blast results

The Blast result page shows an overview of the results, colour coded by sequence identity in descending order by score (Figure 38). You can re-order the results by E-Value, Score or Identity.

Figure 38. The Blast results page.

Filters and views

You can use filters on the left hand side to narrow down your search results, e.g. to limit the results to a particular species. You can also map the results to UniProt protein databases UniProtKB, UniRef and UniParc. You can view results by taxonomy or in plain text format.

Alignments table

More detail about each result can be seen in the 'Alignments' table under the 'Overview' section, showing the query sequence aligned to each subject sequence.

You can view each alignment in more detail by clicking on the graphic or on the 'view alignment' link. You can also see information about E-Value, Score and Identity. To add more information, click on the 'edit columns' button.

You can run another Blast search from here by selecting an entry using the checkboxes and clicking on the 'Blast' button. You can also align your Blast results by selecting them using checkboxes and then clicking on the 'Align' button.
button. The results can be downloaded in various formats or be added to your Basket for later.

Aligning multiple protein sequences

Sequence alignments

Carrying out sequence alignments...

1. Click on the Align link in the header bar to align two or more protein sequences with the Clustal Omega program [44].

2. Enter either protein sequences in FASTA [45] format or UniProt identifiers into the form field (Figure 39).

3. Click the 'Run Align' button.

The following kinds of UniProt identifiers are supported:

- P00750: UniProtKB entry
- P00750-2: UniProtKB entry isoform sequence
- A4_HUMAN: UniProtKB entry name
- UP1000000001: UniParc entry
- UniRef100_P00750: UniRef entry

Figure 39. Using the alignment tool.
UniProtKB entries provide an align button to align the canonical sequence with its isoforms. All relevant results pages (such as UniProtKB, UniRef, UniParc and tool results) provide an ‘Align’ button to run alignments directly by selecting entries with checkboxes. You can also run Alignment from within the Basket.

You can learn more about sequence alignments on the UniProt help page [46].

Alignment results

The Alignment results page has a navigation panel on the left, a download button on top and an ‘edit and resubmit’ button on top. The navigation panel allows you to show or hide sections using the checkboxes and jump to sections by clicking on their name. Alignment results are made up of three main sections (Figure 40):

Figure 40. The three main sections of the Alignment results page.

The ’Alignment’ section shows the complete alignment along with the ’Highlight’ panel on the left. This panel allows you to highlight various annotations and amino acid properties in the alignment (Figure 41).
Figure 41. The ‘Alignment’ section.

The ‘Tree’ section shows evolutionary relationships between the sequences by representing the species on branches with varying distances (Figure 42).

Figure 42. The ‘Tree’ section.

The ‘Result info’ section shows information such as the query sequences, date of job execution, job identifier, running time, identical positions, identity, similar positions and program of alignment. Jobs are stored for seven days and the job identifier can be used to access the alignment through UniProt for that duration.

**Batch retrieval & ID mapping**

You can use the ‘Retrieve/ID Mapping’ feature in UniProt to download UniProt entries corresponding to a list of UniProt accessions. You can also use this feature to convert database identifiers from UniProt to external databases or vice versa.

**Batch retrieval of UniProtKB entries**
1. Click on ‘Retrieve/ID Mapping’ in the header bar.

2. Provide your list of UniProtKB identifiers in the box titled ‘1. Provide your identifiers’. You can separate the identifiers with a space or enter each one on a new line. You can also upload an input file from your machine (Figure 43).

3. For the default batch download of UniProtKB entries, leave the ‘From’ and ‘To’ dropdowns on their default selection of ‘UniProtKB’ and click on the ‘Go’ button.

You get a results page where you can further filter the results, edit the table columns to add or remove information and download the information in various formats.

![Retrieve/ID Mapping tool](Image)

**Converting UniProt identifiers to external identifiers (or vice versa)**

Try if for yourself...

Let's assume that we have a list of UniProtKB identifiers that we would like to convert to PDB [8] identifiers.
1. Click on 'Retrieve/ID Mapping' in the header bar.

2. Provide your list of UniProtKB identifiers in the box titled '1. Provide your identifiers'. You can separate the identifiers with a space or enter each one on a new line. You can also upload an input file from your machine (Figure 43).

3. Select the 'From' database as UniProtKB (default selection) and the 'To' database as 'PDB'. As you can see from the 'From' and 'To' dropdowns, you can convert various types of identifiers to UniProtKB and vice versa (Figure 44, below).

You will get a results page with your converted identifiers ready for download.

![Retrieve/ID mapping](image)

Figure 44. Converting UniProtKB identifiers to PDB identifiers.

**Results**

**When the 'to' target is UniProtKB**

If you are mapping external IDs to UniProtKB or retrieving a batch of UniProtKB IDs, you will see a results page similar to the UniProtKB results (Figure 45).
Figure 45. The results page seen when mapping external IDs to UniProtKB.

When the 'to' target is an external database

If you are mapping UniProtKB IDs to an external database, you will get a list of mapped IDs (Figure 46).

Figure 46. The results page seen when mapping UniProtKB IDs to an external database.

Peptide search

The peptide search tool allows you to submit peptide sequences of at least 3 residues and to find all UniProtKB sequences which have an exact match to the query sequence.

To access the tool, click on the 'Peptide search' link in the header which is at the top of every page on the UniProt website.
All you need to provide as input is a peptide sequence at least three amino acids long. You have the option to choose a taxonomy restriction and you can also specify whether you would like the tool to treat isoleucine and leucine as equivalent.
Peptide search results

The peptide search results provide UniProtKB [48] entries with sequences that contain a match for the peptide(s) that you searched for. If you searched for multiple peptides, results are shown if any one of them matched a sequence (Figure 49).

**Figure 49. Peptide search results page**

**Overview**

The results page provides a reporting sentence with a link to the job ID and a summary of the number of sequence matches found for the peptides you submitted. You can use the job ID in the URL to access the results for up to 7 days after you have performed the search. After this, the job is deleted from our servers.

**Filters and views**

You can use filters on the left hand side to narrow down your search results, e.g. to limit the results to a particular species. You can also map the results to the UniProt [9] protein databases UniProtKB, UniRef [49] and UniParc [50]. You can view results by the default results table format but also by taxonomy and keywords.

**How to get data from UniProt**

There are a number of different ways to access UniProt data including:

- downloading data from the UniProt FTP site;
downloading small data sets from the UniProt web site;

accessing the resource programmatically.

Data downloads

Downloading complete data sets

You can download the entire UniProtKB [48], UniRef [49] and UniParc [50] databases. For downloading these complete data sets, we recommend that you use the UniProt FTP site [51]. If you are located in Europe, the Middle East or Africa, you may find it faster to download data from our mirror sites in the United Kingdom [52] or Switzerland [53] instead.

Downloading subsets of data

You can download small data sets and subsets directly from the UniProt web site by following the download link on any search result page. This will allow you to download data in a range of formats including FASTA [45], tab-delimited, text, Excel, GFF [54], XML [55] and RDF/XML [56] (Figure 50).

Figure 50. Use the 'Download' button to download data sets directly from the web site.

Accessing UniProt data programmatically

UniProt [9] provides several application programming interfaces (APIs [57]) to query and access its data programmatically.

Ways to access UniProt programmatically

1. RESTful UniProt website

What: RESTful URLs that can be bookmarked, linked and used in programs for all entries, queries and tools available through the UniProt website. Data is available in all formats provided on the website, e.g. text, XML [55], RDF, FASTA [45], GFF [54], tab-separated for UniProtKB [48] protein data.

Why: Access data and tools from the UniProt website with any programming language.

Documentation: https://www.uniprot.org/help/api [58]

2. Proteins REST [59] API
What: Extended REST API with a service providing genomic coordinates of UniProtKB sequences, and other services providing annotations imported and mapped from large-scale data sources such as 1000 Genomes, ExAC, PeptideAtlas, MaxQB and HPA via the variation, proteomics [27], and antigen services.
Why: Access UniProt data as well as large-scale data sets (like variants and proteomics data) imported and mapped to UniProt through a single service.
Documentation: https://www.ebi.ac.uk/proteins/api/doc/ [60]

3. UniProt SPARQL API
What: SPARQL API for all UniProt data, stored in Resource Description Framework (RDF) format (Help). An SQL-like graph query language that allows you to perform complex queries across all UniProt data, as well as across other resources that provide a SPARQL endpoint, such as Ensembl [10] or Wikidata.
Why: Access data from UniProt, and other resources, using a low-cost alternative to importing the data into e.g. a relational database [61] and building a local data warehouse [62].
Documentation: https://sparql.uniprot.org [63]

4. UniProt Java API
What: A Java library that provides a stable remote API for programmatically accessing UniProt data.
Why: Access data and tools from UniProt using Java.
Documentation: https://www.ebi.ac.uk/uniprot/japi/ [64]

Tasks which can be performed programmatically

- Retrieval of individual entries;
- Batch retrieval of entries;
- Retrieval of entries via queries;
- Release number and date;
- Convert between different data formats;
- Query variation data from UniProt as well as large-scale sources (1000 Genomes, EXAC, etc);
- Query by genome coordinates to find proteins;
- Retrieve all positional sequence features for an entry;

How to submit data to UniProt

Submitting protein sequence data

UniProt accepts submissions of directly sequenced protein sequences obtained by Edman degradation [65] or by MS/MS if the spectra obtained have been studied manually and can be provided with the amino acids and ions. These sequences and any associated biological information can be submitted using SPIN [66], a web-based submission tool (Figure 51). All of the information required to create a database entry is collected during the submission process. Accession [17] numbers are provided for submitted sequences which can be used in
Figure 51. Directly sequenced protein sequences can be submitted to UniProt using the online submission tool SPIN.

Submitting updates and corrections

We welcome any feedback regarding updates or corrections of existing data which can be sent to the help [at] uniprot.org (UniProt help desk).

publications and data can be kept confidential until publication.

Summary

What is UniProt? [67] The Universal Protein Resource (UniProt [3]) is a comprehensive resource for protein sequence and annotation data.

Why use UniProt? [68] UniProt provides a regularly updated, comprehensive body of protein information at a single site with links to more than 150 other resources providing access to additional related information. It aids scientific discovery by collecting, interpreting and organising this information so that it is easy to access and use.

What are the UniProt databases? [69] There are three UniProt databases:

- UniProt Knowledgebase (UniProtKB) - the central hub for the collection of functional information on proteins;
- UniProt Reference Clusters (UniRef) - provides clustered sets of sequences;
- UniProt Archive (UniParc) - a comprehensive and non-redundant database that contains most of the publicly available protein sequences.

How to access UniProt? [70] UniProt can be accessed via the UniProt website [3].

What data does UniProt provide? [71] UniProt provides sequence data as well as a huge range of protein information covering areas such as function, subcellular location, involvement in disease, subunit structure, interactions with other proteins, post-translational modifications and much more.

Where does UniProt data come from? [71] The data comes from a range of sources including scientific literature, sequence analysis tools, in-house automatic annotation systems and other biological databases.

How can I download UniProt data? [72] You can download the entire UniProtKB, UniRef, UniParc and UniMES databases from the UniProt FTP site. Small data sets can be downloaded directly from the UniProt web site by following the download link on any search result page. You can also access the data via the UniProt JAPI or the UniProt BioMart or programatically using REST.

What tools are provided? [73] UniProt provides four tools for protein sequence analysis:

- BLAST is provided for sequence similarity searching;
- The ‘Align’ tool allows you to align two or more protein sequences using the Clustal Omega program;
- The ‘Retrieve/ID mapping’ tool allows you to submit a list of identifiers to retrieve the corresponding UniProt
entries or to map them from or to an external database;
- The ‘Peptide search’ tool allows you to submit short peptide sequences of at least 3 residues and find all UniProtKB sequences which have an exact match to the query sequence.

When to use UniProt: guided examples

The section allows you to revise the knowledge you have gained on this course by providing guided examples of how UniProt can be used. The following examples demonstrate some uses of UniProt:

1. Finding protein function
2. From disease to protein to variant
3. Download the proteome set for E.coli (strain K12)

Finding protein function

UniProt is the primary source for high quality and expertly curated functional data on proteins. This example demonstrates how to use UniProtKB to find out the function of CDC7.

- On the UniProt website [3], select ‘UniProtKB’ (this is the default selection) from the drop-down menu next to the search box. Now enter the gene name ‘cdc7’ in the search box and click on the ‘Search’ icon (Figure 52).

![Figure 52. UniProtKB search with ‘cdc7’.](image)

- You get a results page with an entry named CDC7_HUMAN being the top hit (Figure 53).

![Figure 53. UniProtKB results for ‘cdc7’.](image)

- If you click on the top hit, you are taken to the protein entry page. The page title area describes the protein name, gene name, organism and status. Further down, the first information section is ‘Function’ (Figure 54). As the title area shows, this is a Reviewed (Swiss-Prot) entry - the ‘Function’ section provides expertly curated information. Evidence for annotation can be accessed through the ‘1 publication’ evidence tag.
From disease to protein to variant

UniProt makes it easy to identify and retrieve disease-related proteins and the disease-causing variants that they contain. One way to do this is to follow the steps below. In this example, we are interested in finding out about spinal muscular atrophy (SMA2).

- On the [UniProt website](https://www.uniprot.org), select 'Diseases' from the drop-down menu next to the search box, add the disease name 'sma2' in the search box and click on the 'Search' button (Figure 55).

  ![Figure 55. Diseases search for 'sma2'.](https://www.ebi.ac.uk/training/online)

  - 'sma2' is an abbreviation of 'spinal muscular atrophy 2' which is the disease that is retrieved by the search. The search returns a page which gives an overview of the disease and a link to a UniProtKB protein (Figure 56).
Figure 56. Diseases results for 'sma2'.

- You can click on the disease name to find some more details about this disease as well as links to other resources which also provide information about the disease (Figure 57).

Figure 57. Diseases entry for 'spinal muscular atrophy 2'.

- Clicking on the UniProtKB link in the original results page will take you to a table with a summary of the protein which has been implicated in this disease, survival motor neuron protein (Figure 58).

Figure 58. 'spinal muscular atrophy 2' mapped to UniProtKB results.

- Click on the accession number (Q16637) to view the entry. Disease-related information can be found in the 'Pathology & Biotech' section of the entry. This provides a summary of the disease as well as a list of variants which are known to be involved in the disease and links to the publications where these were originally reported (Figure 59).
As well as accessing variant information from the website, UniProt also provides variation information through FTP downloads [74]. The humsavar.txt file is an index of manually curated human polymorphisms and disease mutations in UniProtKB/Swiss-Prot. Additional files are provided for a number of species which list variants from sources - such as 1000 Genomes [75], Ensembl [76] and COSMIC [77] - that are not in UniProtKB and that modify the protein sequence, including nonsynonymous or missense variants, stop lost, stop gained and initiator codon variants.

**Download the proteome set for E.coli (strain K12)**

You can use UniProt to download protein sets for completely sequenced organisms (also known as 'proteomes'). For example, let's try and download the proteome for Escherichia coli strain K12.

- Go to the [UniProt website](https://www.uniprot.org) and click on the search selection dropdown (Figure 60).
Figure 60. Dataset selection dropdown.

- Select 'Proteomes', type *Escherichia coli* and click on the looking search icon (Figure 61).

Figure 61. Proteomes search for 'Escherichia Coli'.

- You get a results page with 'Escherichia coli (strain K12)' being the top hit. You can also see an icon next to the name showing that this is a Reference Proteome. Reference Proteomes have been selected to cover well-studied model organisms and other proteomes of interest for biomedical research (Figure 62).
Figure 62. Proteomes results for 'Escherichia Coli'.

- You can click on the Proteome ID to see the full proteome entry (Figure 63).

Figure 63. Proteome entry for 'Escherichia Coli (strain K12)'.

- If you scroll down to the 'Components' section, you can see a download button (Figure 64). This will allow you to download the full proteome or individual components such as chromosomes for those species with multiple chromosomes or organelles such as the mitochondrial genome.
Figure 64. Components in the entry for 'Escherichia Coli (strain K12)'.

Exercises

If you need help to complete this section you can look in the ‘Need some help?’ and ‘Want to know how we did it?’ sections.

Finding entries with 3D structures

Scenario

You were flipping through a past issue of Science and came across the following article:

The Protein Kinase Complement of the Human Genome

G. Manning¹,², D. B. Whyte¹, R. Martinez¹, T. Hunter² and S. Sudarsanam¹,²

Abstract

We have catalogued the protein kinase complement of the human genome (the "kinome") using public and proprietary genomic, complementary DNA, and expressed sequence tag (EST) sequences. This provides a starting point for comprehensive analysis of protein phosphorylation in normal and disease states, as well as a detailed view of the current state of human genome analysis through a focus on one large gene family. We identify 518 putative protein kinase genes, of which 71 have pseudogenes. Chromosomal mapping revealed several small clusters of kinase genes and revealed that 244 kinases map to disease loci or cancer amplicons.
Exercise

You want to find all human protein kinases in UniProt that have a 3D structure associated with them. How would you do this?

Need some help?

Try using the advanced search..

Want to know how we did it?

1. Click on 'Advanced' in the search box;
2. Select 'Protein name' from the field dropdown and enter 'kinase' in the input box;
3. Select 'Structure' in the next row, then '3D structure' and then 'yes' in the consecutive field dropdowns;
4. Select 'Organism' in the next row and enter 'Homo Sapiens';
5. Hit the search button;
6. You will come to the results page with all matching results displayed.

Mapping other database identifiers to UniProt

Scenario

You are helping colleagues analyse the protein products of the gene tp53 in different organisms. They have sent you a list of gene accessions that they are interested in. They are interested in analysing the corresponding protein products for these genes.

<table>
<thead>
<tr>
<th>Genes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X99952</td>
<td>[78]</td>
</tr>
<tr>
<td>AP002032</td>
<td>[79]</td>
</tr>
<tr>
<td>CP002688</td>
<td>[80]</td>
</tr>
<tr>
<td>AY056186</td>
<td>[81]</td>
</tr>
</tbody>
</table>
Exercise
How would you download the corresponding UniProt protein sequences for these gene accessions?

Need some help?
Try using the Batch retrieval and ID mapping tool.

Want to know how we did it?

1. Click on 'Retrieve/ID Mapping ' in the header;
2. Copy and paste the list of gene accessions into the input box;
3. Select 'EMBL/Genbank/DDBJ' in the 'From' dropdown and 'UniProtKB' in the 'To' dropdown;
4. Hit the Go button;  
5. You will come to the results page with all matching results displayed.

Quiz: Exploring UniProt

Questions: 10  
Attempts allowed: Unlimited  
Available: Always  
Pass rate: 75%  
Backwards navigation: Allowed

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

References

UniProt overview

These papers give an overview of the UniProt project:


Read more

You can find additional papers about the project on the UniProt website [101].

Acknowledgements
Contributors

UniProt is produced by staff at the European Bioinformatics Institute (EMBL-EBI) in Hinxton, Cambridge, UK, the SIB Swiss Institute of Bioinformatics in Geneva, Switzerland and the Protein Information Resource at Georgetown University, Washington DC and University of Delaware, USA (Figure 65).

Figure 65. The UniProt Consortium.

Funding

UniProt is supported by the National Institutes of Health [4U41HG006104-04, 5U41HG002273-13, 5R01GM080646-08, 5G08LM010720-03, 5P20GM103446-13]; British Heart Foundation [SP/07/007/23671]; Swiss Federal Government through the State Secretariat for Education, Research and Innovation; National Science Foundation [DBI-1062520]; and European Molecular Biology Laboratory core funds.

Contributors

[1]
Michele Magrane

EMBL-EBI
Annotation Coordinator, Protein Function Content team

Michele Magrane is the UniProt annotation coordinator within the Protein Function Content team at EMBL-EBI, managing a team of biocurators who contribute to the expert curation of the UniProt Knowledgebase. She studied at University College Dublin and has more than 15 years’ experience in the field of protein curation. She was a founding member of the International Society for Biocuration and has served on the Nominating and Award Committees of the society.

Sangya Pundir

EMBL-EBI
User Experience Manager - Martin team: UniProt development

Sangya Pundir is a User Experience (UX) Manager in EMBL-EBI’s UniProt team, where she established a user-centred process for the redesign of the world’s leading protein resource. To make UniProt easy for researchers to explore, Sangya conducts usability testing and information-gathering methods such as card sorting, contextual studies and workshops. Before she came to EMBL-EBI, Sangya worked at a healthcare consultancy, designing bespoke management systems. She holds an MSc in Biotechnology, Bioprocessing and Business Management from the University of Warwick.

Source URL: https://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional

Links
[1] https://www.ebi.ac.uk/training/online/trainers/magrane
[2] https://www.ebi.ac.uk/training/online/trainers/sangya.pundir
[5] https://www.ebi.ac.uk/training/online/glossary/cds
[6] https://www.ebi.ac.uk/training/online/glossary/manual-annotation
[7] https://www.ebi.ac.uk/training/online/glossary/uniprotkbswiss-prot
[8] https://www.ebi.ac.uk/training/online/glossary/pdb
[9] https://www.ebi.ac.uk/training/online/glossary/uniprot
[10] https://www.ebi.ac.uk/training/online/glossary/ensembl
[11] https://www.ebi.ac.uk/training/online/glossary/refseq
[15] https://www.ebi.ac.uk/training/online/glossary/ontology
[16] https://www.ebi.ac.uk/training/online/glossary/eco
[17] https://www.ebi.ac.uk/training/online/glossary/accession
[19] https://www.ebi.ac.uk/training/online/glossary/boolean
[20] https://www.ebi.ac.uk/training/online/glossary/swiss-prot
[21] https://www.ebi.ac.uk/training/online/glossary/proteome
[22] https://www.ebi.ac.uk/training/online/glossary/heuristic
[24] https://www.ebi.ac.uk/training/online/glossary/gene-ontology
[27] https://www.ebi.ac.uk/training/online/glossary/proteomics
[29] https://www.ebi.ac.uk/training/online/glossary/intact
[30] https://www.ebi.ac.uk/training/online/glossary/subcellular
[33] https://www.ebi.ac.uk/training/online/glossary/polymorphism
[34] https://www.ebi.ac.uk/training/online/glossary/protein-family
[35] https://www.ebi.ac.uk/training/online/glossary/secondary-structure
[36] https://www.ebi.ac.uk/training/online/glossary/lipid
[37] https://www.ebi.ac.uk/training/online/glossary/variant
[38] https://www.ebi.ac.uk/training/online/glossary/active-site
[39] https://www.ebi.ac.uk/training/online/glossary/http
[40] https://www.ebi.ac.uk/training/online/glossary/html
[41] https://www.ebi.ac.uk/training/online/glossary/blast
[42] https://www.ebi.ac.uk/training/online/glossary/active-site
[45] https://www.ebi.ac.uk/training/online/glossary/fasta
[48] https://www.ebi.ac.uk/training/online/glossary/uniprotkb
[49] https://www.ebi.ac.uk/training/online/glossary/uniref
[50] https://www.ebi.ac.uk/training/online/glossary/uniparc
[51] https://www.ebi.ac.uk/training/online/ftp.uniprot.org
[52] ftp://ftp.ebi.ac.uk/pub/databases/uniprot/
[54] https://www.ebi.ac.uk/training/online/glossary/gff
[55] https://www.ebi.ac.uk/training/online/glossary/xml
[56] https://www.ebi.ac.uk/training/online/glossary/rdfxml
[57] https://www.ebi.ac.uk/training/online/glossary/api
[58] https://www.uniprot.org/help/api
[59] https://www.ebi.ac.uk/training/online/glossary/rest
[60] https://www.ebi.ac.uk/proteins/api/doc/
[61] https://www.ebi.ac.uk/training/online/glossary/relational-database
[62] https://www.ebi.ac.uk/training/online/glossary/data-warehouse
[63] https://sparql.uniprot.org
[64] https://www.ebi.ac.uk/uniprot/japi/
[65] https://www.ebi.ac.uk/training/online/glossary/edman-degradation
[66] https://www.ebi.ac.uk/swissprot/Submissions/spin/
[67] https://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional/what-uniprot
[70] https://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional/how-access-and-navigate-uniprot-website