Published in final edited form as:

Curr Protoc Bioinformatics. ; 50: 1.27.1–1.27.10. doi:10.1002/0471250953.bi0127s50.

Searching and navigating UniProt databases

Sangya Pundir^{1,*}, Michele Magrane¹, Maria J. Martin¹, Claire O'Donovan¹, and The UniProt Consortium^{1,2,3}

¹European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, U.K ²Swiss Institute of Bioinformatics, Geneva, Switzerland ³Protein Information Resource, Washington D.C., U.S.A

Abstract

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt website receives approximately 400,000 unique visitors per month and is the primary means to access UniProt. It provides ten searchable datasets and three main tools. The key UniProt datasets are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef) and the UniProt Archive (UniParc) and protein sets for completely sequenced genomes (Proteomes). Other supporting datasets include information about proteins that is present in UniProtKB protein entries like literature citations, taxonomy, subcellular locations, etc. This paper focuses on how to use UniProt datasets. The basic protocol describes navigation and searching mechanisms for the UniProt datasets while two alternative protocols build on the basic protocol to describe advanced search and query building.

Keywords

UniProt; search; navigation; tutorial

INTRODUCTION

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. UniProt, or the Universal Protein Resource provides an up-to-date, comprehensive body of protein information at a single site.

The UniProt website can be accessed at the URL <http://www.uniprot.org/>. The following three basic protocols describe how you can navigate the site to access datasets and how you can make the most of the search functionality to find your data of interest within these datasets.

*spundir@ebi.ac.uk.

BASIC PROTOCOL 1: SEARCHING UNIPROT DATASETS

The UniProt website provides ten main datasets and three main tools. The key UniProt datasets are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), the UniProt Archive (UniParc) and protein sets for completely sequenced genomes (Proteomes). Supporting datasets include information about proteins that is present in UniProtKB protein entries like literature citations, taxonomy, subcellular locations, keywords, cross-referenced databases and diseases. Additional searchable sets include annotation programs used in UniProt and UniProt help. The three tools that UniProt provides are the 'Blast' sequence search tool, the 'align' multiple sequence alignment tool and the 'Retrieve/ ID Mapping' tool where you can upload lists of identifiers to download corresponding UniProt entries or map them to/ from external databases. The following steps describe how to explore and search the different datasets within UniProt.

Necessary Software

An up-to-date web browser and computer.

Text search and filtering within a UniProt dataset

1. Go the UniProt home page at <http://www.uniprot.org/>.
2. Click on the dropdown to the left of the search box to see all UniProt datasets and select the one you are interested in. The dropdown panel to the left of the search box is shown in Figure 1

You can also select the main datasets as tiles on the home page. Note that the background color around the search field changes depending on the dataset, in order to help identify the selected dataset.

Enter your query in the search box and hit the search button. For example, select UniProtKB (the default option) and enter 'human insulin'. For the full UniProt query syntax, see Table 1.
3. You will see the search results page for your dataset. The results page for the example query is shown in Figure 2.
4. Click on the filters on the left to narrow down your results. For example, click on 'human' under the 'popular organisms' filter to narrow down the term 'human' as the organism. Then go to the 'search terms' filter and select 'insulin' as 'protein name'. The resulting screen with the selected filters is shown in Figure 3.
5. Click on the accession links under the 'Entry column' to view an individual protein entry. For example, clicking on INS_HUMAN will take you to the protein entry for human insulin.
6. The human insulin entry as seen in Figure 4 has a summary at the beginning that displays its protein name, gene name, organism, Reviewed or Unreviewed status, annotation score and evidence level. You can use the buttons along the top to launch tools, view the entry in different formats, add it to your basket or view a history of changes made to this entry. When there is an isoform present, you can

align the sequences to view similarities and differences. You can explore information and the supporting evidence about Function, Subcellular location, Pathology & biotech, etc. by clicking on the buttons in the navigation panel on the left. Further guidance is available at the 'Guidelines for Understanding Results' section.

BASIC PROTOCOL 2: ADVANCED SEARCH AND QUERY BUILDING

You can 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 logic. The advanced search provides a query builder, which helps you expand your query. The advanced search translates search terms that have been entered and filters that have been applied already into the query builder and can be accessed from all pages.

Necessary Software

An up-to-date web browser and computer.

Advanced search within the UniProt dataset

1. Go to the UniProt home page at <http://www.uniprot.org/>.
2. For help on advanced search options available, click on 'text search' under the 'Getting started' area of the home page. You will see a banner saying 'Click here for advanced options'. Clicking on 'options' with the superscript 'i' (for information) will take you to a help page that lists all advanced search fields under UniProtKB.
3. Click on the dropdown to the left of the search box and select your dataset. The default is UniProtKB.
4. Click on 'advanced' to the right of the search box. There are various field types available under advanced search, found by clicking on the dropdown box as seen in Figure 5.
5. Depending on the chosen dataset and field, you can enter some text or choose values from a drop-down list. Where relevant, auto-completion is available. For example, select field type 'Organism' and enter 'Homo sapiens', then select field type 'Protein name' and enter 'insulin'. You can click on the '+' icon to add more rows and further build your query. Hit the search button. The advanced search panel is shown in Figure 6.
6. You will see the search results page with your parameters reflected in the filters, as in Figure 7.
7. Click on the accession links under the 'Entry column' to view an individual protein entry.

BASIC PROTOCOL 3: ADDING PARAMETERS USING ADVANCED SEARCH

The advanced search can be a powerful tool to narrow down to very specific results. Here we look at an example where we use various advanced search fields to find all entries in

UniProtKB with **direct protein sequencing evidence** that are **encoded in the mitochondrial sequence** and have **manual experimental evidence for function**.

Necessary Software

An up-to-date web browser and computer.

Advanced search within the UniProt dataset

1. Go to the UniProt home page at <http://www.uniprot.org/>.
2. Click on the dropdown to the left of the search box and select UniProtKB.
3. Click on 'advanced' to the right of the search box to access advanced search.
4. Click on the dropdown box within advanced search, select the option 'Keywords' and type 'direct protein sequencing' (you can select it from an autocomplete dropdown once you begin typing).

For additional exploration, all keywords can be found under the 'Keywords' supporting dataset <http://www.uniprot.org/keywords/>.

5. Click on the dropdown in the next row and select 'Sequence', then 'Encoded in' in the dropdown to the right and then type 'mitochondrion'.

You can select it from an autocomplete dropdown once you begin typing.

6. Click on the '+' icon at the right hand end of the row to add another parameter. Click on the new dropdown box and select 'function', then click on the 'Evidence' dropdown and select 'Experimental'. You should have a view that looks like Figure 8. You can leave the input field vacant to include all possible values. Hit the search button.
7. You will see the search results page, in this case with just one result that matches all your parameters as shown in Figure 9.
8. You can click on the entry accession to explore the protein entry in detail.

GUIDELINES FOR UNDERSTANDING RESULTS

UniProtKB Search Results

The UniProtKB results page provides filters on the left and the results table on the right. Filters allow you to select results belonging only to the Reviewed (Swiss-Prot) or Unreviewed (TrEMBL) section or results from just a certain organism. You can also refine your search terms using the 'Search terms' filters. You can view your results classified by their taxonomy, keywords, Gene Ontology, Enzyme class, Pathways, etc. and also follow links to see your results clustered by sequence identity in the UniRef dataset.

The results page also provides a row of buttons along the top to access tools, download your results and add them to your basket. The 'Columns' button can be used to edit the columns you see to add or remove information. You can select entries to launch BLAST searches, run multiple sequence alignments and add selected entries to your basket. You can download the entire results table or just your selected entries.

UniProtKB Protein Entry

The UniProtKB protein entry page, such as the INS_HUMAN example, presents the protein sequence and all the annotation related to the protein. This example belongs to the Reviewed (Swiss-Prot) section, which is expertly annotated by UniProt curators. It has an annotation score of five out of five, indicating a high level of manual annotation. You can view the evidence of annotated information using evidence tags. For example 'Pathology & Biotech' presents a '4 publication' evidence tag, which can be opened to view a summary of the publications cited as seen in Figure 10. Evidence tags are color coded such that gold indicates a manual assertion and blue indicates an automatic assertion.

Aligning Isoforms

The entry page also provides various buttons for tools, formats and to add the entry to your basket along the top of the page. There are also other tools available for sequences in the 'Sequence' section. When there are isoforms present, you can align them using the 'Align' button that is at the top of the page and also within the 'Sequence' section. The results show the sequence alignment including gaps. There is a 'Highlight' panel on the left that allows you to highlight annotations to view them in the alignment. Sequence annotations are mapped to the main canonical sequence so some features may not be available for all isoforms. You can compare highlighted regions across the aligned isoform sequences to view conserved areas and differences. The panel also allows you to highlight amino acid properties. For example, you can view all the hydrophobic amino acids in the aligned sequences by clicking on 'Hydrophobic' as shown in Figure 11.

COMMENTARY

Background Information

UniProt aids scientific discovery by collecting, interpreting and organizing information so that it is easy to access and use. It saves researchers countless hours of work in monitoring and collecting this information themselves. The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information and other rich annotation on proteins. It is further divided into the Reviewed (Swiss-Prot) expertly annotated section and the Unreviewed (TrEMBL) automatically annotated section. The UniProt Archive (UniParc) is a non-redundant archive containing all the publicly available protein sequences in the world. The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from UniProtKB (including isoforms) and selected UniParc entries. UniRef reduces redundancy and provides complete coverage of the sequence space at three levels of sequence identity, i.e. 100%, 90% and 50% identity. The Proteomes dataset provides protein sets for organisms with completely sequenced genomes. Supporting datasets are a collection of meta-information about proteins in UniProtKB entries such as literature citations, taxonomy, subcellular locations, keywords, cross-referenced databases and diseases.

The UniProt website provides powerful searching and filtering features to help users find the exact data they are interested in. This is further enhanced by a flexible and effective advanced search system, which allows users to define their search terms and build their queries. UniProt provides training material through the European Bioinformatics Institute's

train online portal, including a quick tour (<http://www.ebi.ac.uk/training/online/course/uniprot-quick-tourversion-0>) and a detailed course (<http://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional>). UniProt also provides short video tutorials embedded in the website and also available on our YouTube channel at <https://www.youtube.com/uniprotvideos>.

Acknowledgments

This work was supported by the National Institutes of Health [U41HG006104, U41HG007822, U41HG002273, R01GM080646, G08LM010720, P20GM103446]; British Heart Foundation [RG/13/5/30112]; Parkinson's Disease United Kingdom [G-1307]; Swiss Federal Government through the State Secretariat for Education, Research and Innovation; National Science Foundation [DBI-1062520]; and European Molecular Biology Laboratory core funds.

Figure 1.
UniProt datasets dropdown

Figure 2. UniProtKB search results page

Figure 3. UniProtKB search results with filter applied

Figure 4.
 UniProtKB proteoin entry INS_HUMAN for human insulin

Figure 5.
UniProtKB advanced search field types available

Figure 6.
Advanced search example

Figure 7.
UniProtKB search results through advanced search

Figure 8.
Adding parameters to UniProtKB advanced search

Figure 9.
UniProtKB search result for additional advanced search parameters

Figure 10.
Evidence tag in INS_HUMAN protein entry

Align

Display All None

- ☒ ALIGNMENT
- ☐ TREE
- ☐ RESULT INFO

Highlight

- Annotation**
- ☐ Natural variant
 - ☐ Propeptide
 - ☐ Sequence conflict
 - ☐ Helix
 - ☐ Turn
 - ☐ Peptide
 - ☐ Chain
 - ☐ Beta strand
 - ☐ Disulfide bond
 - ☐ Signal peptide
- Amino acid properties**
- ☐ Similarity
 - ☒ **Hydrophobic**
 - ☐ Negative
 - ☐ Positive
 - ☐ Aliphatic
 - ☐ Tiny
 - ☐ Aromatic
 - ☐ Charged
 - ☐ Small
 - ☐ Polar
 - ☐ Big
 - ☐ Serine Threonine

[Download](#) [Edit and resubmit](#)

Alignment

How to print an alignment in color

P01308	INS_HUMAN	1	MALWMRLPLLLALLALWGPDPAAAFVNQHL	CGSHLVEALYLVCGERGFFYTPKTRREAED	60
F8WCM5	INSR2_HUMAN	1	MALWMRLPLLLALLALWGPDPAAAFVNQHL	CGSHLVEALYLVCGERGFFYTPKTRREAED	60

P01308	INS_HUMAN	61	LQVGQVELCGGPGAG-----SLOPLA--	LEGS LQKRG-----IV-----E	93
F8WCM5	INSR2_HUMAN	61	LQASALSLSSSTSTWPEGLDARAPPALVVTAN	IGAGGSSSRQFRQRALGTSDSPLVF	120
***. :.*... : : * * : . : : *					
P01308	INS_HUMAN	94	QCCTSI	CSLYQLENYCN-----	110
F8WCM5	INSR2_HUMAN	121	IHCPGAAGTAOGLEYRGRRVTTEL	VWEEVDSSFPQPGSESLPAQPPAOPAFQPEPQQA	180
* . . . * : *					
P01308	INS_HUMAN	111	-----	-----	110
F8WCM5	INSR2_HUMAN	181	PSPEVSCCGLWPRRP	QRSON	200

You may add additional sequences to this alignment (in FASTA format)

Add sequence and align

Figure 11.
Highlighted alignment for INS_HUMAN sequences

Table 1

UniProt query syntax

human antigen	All entries containing both terms.
human AND antigen	
human && antigen	
“human antigen”	All entries containing both terms in the exact order.
human-antigen	All entries containing the term human but not antigen.
human NOT antigen	
human ! antigen	
human OR mouse	All entries containing either term.
human mouse	
antigen AND (human OR mouse)	Using parentheses to override Boolean precedence rules.
anti*	All entries containing terms starting with anti. Asterisks can also be used at the beginning and within terms. Note: Terms starting with an asterisk or a single letter followed by an asterisk can slow down queries considerably.
author:Tiger*	Citations that have an author whose name start with Tiger. To search in a specific field of a dataset, you must prefix your search term with the field name and a colon. To discover what fields can be queried explicitly, observe the query hints that are shown after submitting a query or use the query builder (see below).
length:[100 TO *]	All entries with a sequence of at least 100 amino acids.
citation:(author:Arai author:Chung)	All entries with a publication that was coauthored by two specific authors.