	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ PRODUCTON
UniProt Database **Module Hours: 3**	**Effective Date: 3/15/2021**	**Revision # 1.0** **M. Guzie** **Checked: A. Siclair**

BACKGROUND

The compilation of biological sequence information into large databases has made massive amounts of intel accessible in common places. These biological databases allow for sequential and functional information to be uploaded, reviewed, and made available to the scientific community to use information in their own research pursuit.

The Universal Protein Resource (UniProt) is a database for protein sequence, function, and annotation data. The UniProt Consortium was established in 2002 as a joining of three separate bioinformatic institutes: The European Bioinformatics Institute, the Swiss Institute of Bioinformatics, and the Protein Information Resource. Each institute has an array of bioinformatic and protein sequence and annotation data. Originally, the European Bioinformatics Institute (EBI) and the Swiss Institute of Bioinformatics together formed the Swiss-Prot and TrEMBL databases, while the Protein Information Resource (PIR) founded the Protein Sequence Database (PIR-PSD). The databases were separate and consisted of different protein sequence availability and annotation, until they merged to form the UniProt database. The launch of UniProt thus allowed for a single site containing detailed protein information consisting of sequence, structure, function, sub-cellular location, modifications, mutational analysis, and other cellular interactions (Apweiler, et al., 2004).

The UniProt website consists of 10 different protein databases, attached search tools, and detailed functional annotation. The main four databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), the UniProt Archive (UniParc), and the Proteasomes database. There are 6 supporting databases which consist of literature citations, taxonomy, subcellular locations, cross-reference database, diseases, and keywords. Associated search tools consist of a BLAST search tool, an Align tool, and a Retrieve/ID Mapping tool. This module will go through the means of searching the available databases, using linked search tools, and accessing the variety of accompanying protein information.

Figure 1

Figure 1: The UniProt Logo found on the database website.

1. PURPOSE

The purpose of this procedure is to learn how to use and navigate the UniProt database to access a variety of protein annotation data.

2. SCOPE

This procedure applies to qualified skills center users.

3. RESPONSIBILITY

3.1. It is the responsibility of the user to understand and perform the procedure described in this document.
3.2. It is the responsibility of the user performing the procedure to fully document any deviations from the written procedure.
3.3. It is the responsibility of the user to become trained in the use of this application.

4. DEFINITIONS

4.1. Annotation Score: A score assigned to a protein in the UniProtKB database based on the available annotative information for that protein.
4.2. Accession Code: The unique 6-character code each UniProtKB protein entry is assigned, used to access and search the protein.
4.3. Taxonomy: The systematic classification of groups of biological organisms based on shared characteristics.
4.4. Proteome: The entire set of proteins produced by a specific cell or organism, as a genome is the entire set of genes and genomic information a cell or organism contains.
4.5. Query: The input information, such as a protein or DNA sequence, which is being input into a database for analysis.

5. MATERIALS/EQUIPMENT

5.1. A computer to perform the procedure.
5.2. The UniProt database.

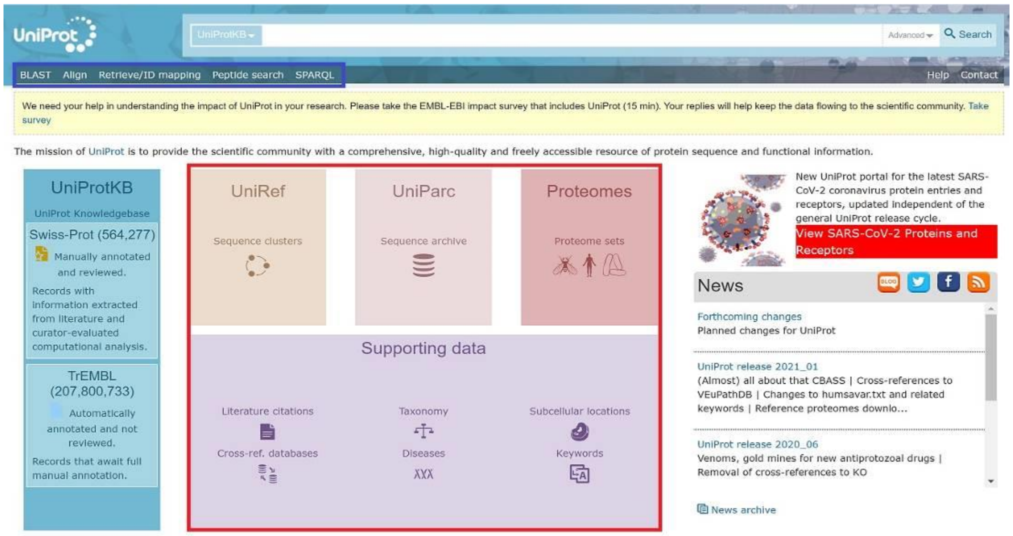
6. PROCEDURE

	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ PRODUCTON
UniProt Database Module Hours: 3	Effective Date: 3/15/2021	Revision # 1.0 M. Guzie Checked: A. Siclair

In this procedure, the search functions of the UniProt website will be explored in order for the user to gain proficiency in navigating the UniProt database along with the ability to access a wide variety of protein annotation data.

6.1. Using the UniProt Knowledgebase (UniProtKB)

6.1.1. Go to the UniProt website: https://www.uniprot.org/. The home page will display the search tools to the upper left of the page and the databases displayed as tiles in the center. (**Figure 2**)



We need your help in understanding the impact of UniProt in your research. Please take the EMBL-EBI impact survey that includes UniProt (15 min). Your replies will help keep the data flowing to the scientific community. Take survey

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

Automatically annotated and not reviewed.

Records that await full manual annotation.

New UniProt portal for the latest SARS-CoV-2 coronavirus protein entries and receptors, updated independent of the general UniProt release cycle.

Forthcoming changes
Planned changes for UniProt

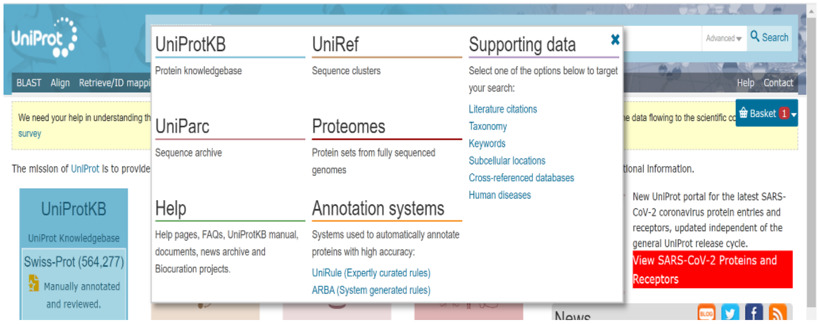
UniProt release 2021_01
(Almost) all about that CBASS | Cross-references to VEuPathDB | Changes to humsavar.txt and related keywords | Reference proteomes downlo...

UniProt release 2020_06
Venoms, gold mines for new antiprotozoal drugs | Removal of cross-references to KO

Figure 2

Figure 2: The UniProt home page.

6.1.2. The database to search from can be chosen by clicking on one of the tiles shown above, or by selecting the drop-down to the left of the search box. It is set to the standard UniProt Knowledgebase (UniProtKB), but by clicking on it, the options for the main searchable databases and supporting databases will be displayed. Begin the first search by selecting the UniProtKB database. (**Figure 3**)



Select one of the options below to target your search:

Protein sets from fully sequenced genomes

Help pages, FAQs, UniProtKB manual, documents, news archive and Biocuration projects.

Systems used to automatically annotate proteins with high accuracy:

We need your help in understanding th … e data flowing to the scientific co … survey

The mission of UniProt is to provide … tional information.

Manually annotated and reviewed.

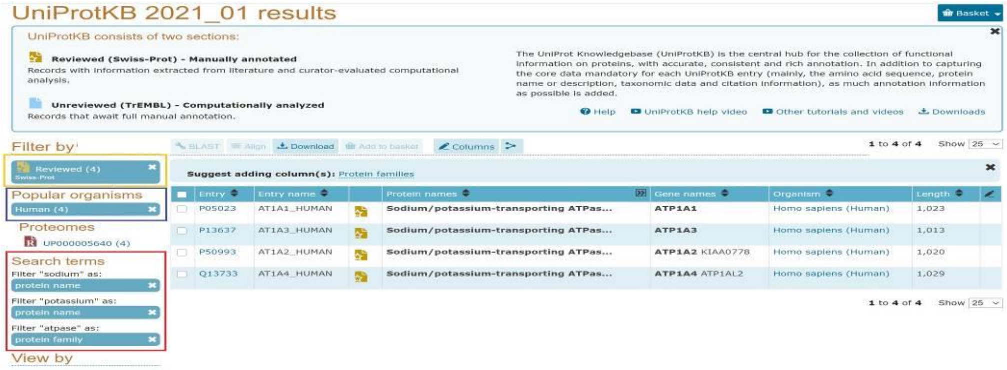
New UniProt portal for the latest SARS-CoV-2 coronavirus protein entries and receptors, updated independent of the general UniProt release cycle.

Figure 3

Figure 3: The 4 main UniProt databases: UniProtKB, UniRef, UniParc, and Proteomes. Supporting datasets are also shown to the right.

6.1.3. The UniProtKB database consists of two sub-databases: SwissProt and TrEMBL. Swiss-Prot contains manually annotated, non-redundant proteins while TrEMBL contains automatically annotated proteins. The collective UniProtKB compiles known information about a specific protein, including associated genes, function, cellular/inter-protein interactions, binding/cofactor sites, expression levels, subcellular location, and variant forms of the same protein. For this reason, UniProtKB is set to the standard search tool and used as a general search tool to find the most information about a given protein. Search for a protein using the organism name followed by the protein name, for example: "Human Sodium Potassium ATPase."

6.1.4. The results page will display a list of proteins matching the search; the search parameters can be narrowed using the "Filter by" box to the left of the results. Shown here: Reviewed was selected, popular organism was set to human, for search terms sodium and potassium were both set to be the protein name, and ATPase to protein family. These parameters narrowed the search to only 4 subunits of the protein. (**Figure 4**)



UniProtKB 2021_01 results

UniProtKB consists of two sections:

Reviewed (Swiss-Prot) - Manually annotated
Records with information extracted from literature and curator-evaluated computational analysis.

Unreviewed (TrEMBL) - Computationally analyzed
Records that await full manual annotation.

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.

Filter by

Reviewed (4) Swiss-Prot

Popular organisms: Human (4)

Proteomes: UP000005640 (4)

Search terms
Filter "sodium" as: protein name
Filter "potassium" as: protein name
Filter "atpase" as: protein family

View by

Suggest adding column(s): Protein families

1 to 4 of 4 Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
P05023	AT1A1_HUMAN	**Sodium/potassium-transporting ATPas...**	**ATP1A1**	Homo sapiens (Human)	1,023
P13637	AT1A3_HUMAN	**Sodium/potassium-transporting ATPas...**	**ATP1A3**	Homo sapiens (Human)	1,013
P50993	AT1A2_HUMAN	**Sodium/potassium-transporting ATPas...**	**ATP1A2** KIAA0778	Homo sapiens (Human)	1,020
Q13733	AT1A4_HUMAN	**Sodium/potassium-transporting ATPas...**	**ATP1A4** ATP1AL2	Homo sapiens (Human)	1,029

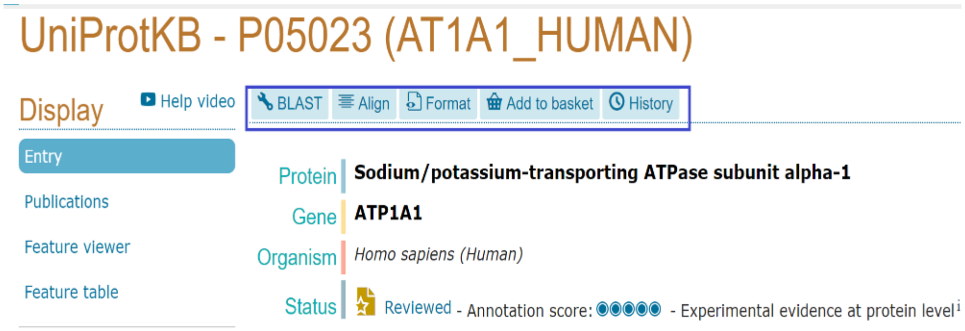
1 to 4 of 4 Show 25

Figure 4

Figure 4: The results for the human sodium potassium ATPase search using filtered results.

6.1.5. Select the protein's accession number (the 6 character code under "Entry") to view the annotations of an individual protein. For example, select P05023 to view the AT1A1_HUMAN sodium/potassium pump alpha subunit 1 entry.

6.1.6. The individual protein entry page will identify the protein by name, gene, organism. It also gives a report on whether the protein is reviewed, the annotation score of the protein, and an evidence level. The annotation score of 5/5 indicates that there is a high amount of annotation related to this protein. Experimental evidence at the protein level indicates that experiments have been done to show the existence of this protein. The buttons lined beneath the entry code can be used to launch the sequence analysis search tools, view the sequential information in a different format, add it to the basket, or view the history of the entry. Explore these functions. (**Figure 5**)


UniProtKB - P05023 (AT1A1_HUMAN)

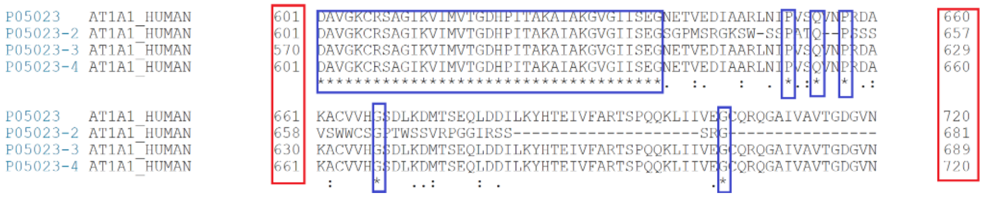
Protein | Sodium/potassium-transporting ATPase subunit alpha-1
Gene | ATP1A1
Organism | *Homo sapiens (Human)*
Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

Figure 5

Figure 5: UniProt entry P05023 for the human sodium potassium ATPase subunit alpha-1.

-**The BLAST tool** can be used to run a BLAST search directly from the UniProt website. BLAST, short for Basic Local Alignment Search Tool, will run the amino acid sequence of the protein through a database and identify sequences which contain aligning segments. BLAST is useful for identifying related proteins or homologs. For more on BLAST, see the BLAST Search Tool Skills Center Standard Operating Procedure. Run a BLAST for the protein.

-**The align tool** can be used to align two or more sequences using the Clustal Omega program. Note that the alignment tool will only be available for a protein which has available isoforms, but different protein sequences can be added into the query box in the FASTA format to search for alignment against. The astericks (*) indicate regions in which all isoforms have aligned. The numbers on the left and right indicate the amino acid position in the chain. Note that these isoforms have different numbers because the isoforms of the sodium potassium pump are different lengths. Run an alignment for the protein if isoforms are available. (**Figure 6**)



```
P05023   AT1A1_HUMAN   601  DAVGKCRSAGIKVIMVTGDHPITAKAIAKGVGIISEGNETVEDIAARLNIPVSQVNPRDA   660
P05023-2 AT1A1_HUMAN   601  DAVGKCRSAGIKVIMVTGDHPITAKAIAKGVGIISEGSGPMSRGKSW-SSPATQ--PSSS   657
P05023-3 AT1A1_HUMAN   570  DAVGKCRSAGIKVIMVTGDHPITAKAIAKGVGIISEGNETVEDIAARLNIPVSQVNPRDA   629
P05023-4 AT1A1_HUMAN   601  DAVGKCRSAGIKVIMVTGDHPITAKAIAKGVGIISEGNETVEDIAARLNIPVSQVNPRDA   660
                            *************************************.  :.   :  . *.:*  * .:

P05023   AT1A1_HUMAN   661  KACVVHGSDLKDMTSEQLDDILKYHTEIVFARTSPQQKLIIVEGCQRQGAIVAVTGDGVN   720
P05023-2 AT1A1_HUMAN   658  VSWWCSGPTWSSVRPGGIRSS-------------------SRG---------------   681
P05023-3 AT1A1_HUMAN   630  KACVVHGSDLKDMTSEQLDDILKYHTEIVFARTSPQQKLIIVEGCQRQGAIVAVTGDGVN   689
P05023-4 AT1A1_HUMAN   661  KACVVHGSDLKDMTSEQLDDILKYHTEIVFARTSPQQKLIIVEGCQRQGAIVAVTGDGVN   720
                             :     *   ..:    : .                    .*
```

Figure 6

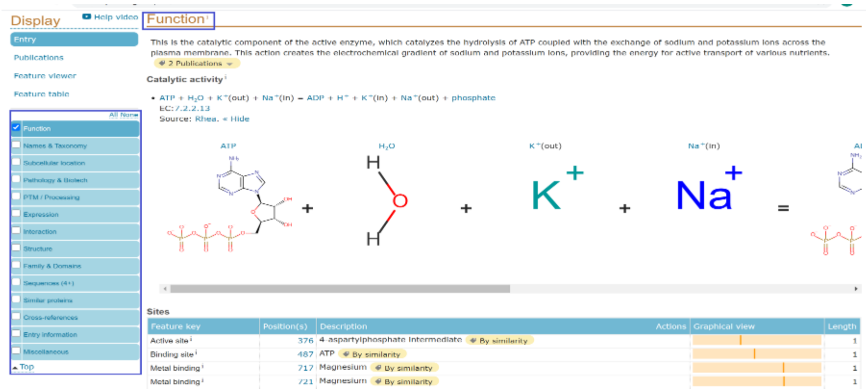
Figure 6: The alignment results for the four different isoforms of UniProtKB entry P05023 (Human sodium potassium ATPase subunit alpha-1).

-**Format** can be used to view the amino acid sequence of the protein in different formats, such as standard text or FASTA format. Different bioinformatic tools may need a protein sequence to be entered in a certain format. The format function is the easiest way to obtain the sequence of a protein. View the protein sequence in all formats.

-Add to basket will keep that protein saved in the basket to return to later.

-History will display what entry version and sequence version the protein is currently at and how recently it was updated to that version. Previous versions can also be selected to view prior entries on the protein.

6.1.7. After viewing the search tools, view and explore the annotative navigation panel lining the left side of the page. Select each feature by clicking the checkmark next to the feature. Here, only function is selected and hence the only data displayed. An overview of the catalytic function of the sodium potassium ATPase is given. (**Figure 7**)



Function

This is the catalytic component of the active enzyme, which catalyzes the hydrolysis of ATP coupled with the exchange of sodium and potassium ions across the plasma membrane. This action creates the electrochemical gradient of sodium and potassium ions, providing the energy for active transport of various nutrients.

2 Publications

Catalytic activity

- ATP + H_2O + K^+(out) + Na^+(in) = ADP + H^+ + K^+(in) + Na^+(out) + phosphate
  EC:7.2.2.13
  Source: Rhea. « Hide

Sites

Feature key	Position(s)	Description	Actions	Graphical view	Length
Active site	376	4-aspartylphosphate intermediate By similarity			1
Binding site	487	ATP By similarity			1
Metal binding	717	Magnesium By similarity			1
Metal binding	721	Magnesium By similarity			1

Figure 7

Figure 7: The functional annotation of UniProtKB entry P05023 (Human sodium potassium ATPase subunit alpha-1).

The function tab also identifies all the known molecular functions and biological functions for the protein of interest, listed beneath the sites. Explore the annotative information that can be found in the rest of the tabs.

-Names & Taxonomy gives an overview on the protein/gene nomenclature. It will also give taxonomic information, such as the classifications of what kingdoms of life and organisms the protein belongs to. It also indicates the proteome if available.

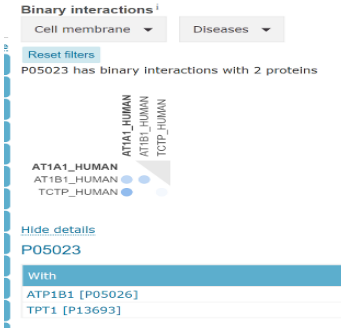
-Subcellular location displays a cellular diagram of where the protein can be found at in the cell. It also includes a written description of each location where the protein can be found along with published research supporting the localization of the protein within that cellular region. This supporting research can be found under both “UniProt annotation” and “GO- Cellular component.” Subcellular location also gives a topology annotation, describing which segments are transmembrane vs. topological (non-membrane region of membrane spanning proteins.) It gives further description of whether the transmembrane segments are helical or beta-sheets, and whether the topological domains are extracellular or cytoplasmic. The length of each segment is also given. Further sequence analysis or a BLAST can be run on each individual segment.

-Pathology & Biotech describes clinical implications of the protein, such as associated diseases with protein mutation/misfunction. It also highlights use of the protein in biotechnology.

-PTM/Processing PTM stands for post-translational modifications; this section shows modifications that occur during protein processing to modify the protein beyond its amino acid sequence and tertiary/quaternary structure. In the “Amino acid modifications” table, each modified residue is identified, and the modification listed (for example, phosphorylation of Tyrosine 10 to phosphotyrosine modulates pumping activity.) There is also a list of available databases compiling information on proteomics and post-translational modifications.

-Expression lists gene expression databases which can be used to identify which regions of the body and cells are expressing the target protein.

-Interaction describes the interactions each domain of the protein has with regulatory molecules/antigens/other proteins. The “Binary interactions” section shows how many protein-to-protein interactions the target protein has and with what other proteins in different areas of the cell. It also shows associated diseases. (**Figure 8**)



Binary interactions

P05023 has binary interactions with 2 proteins

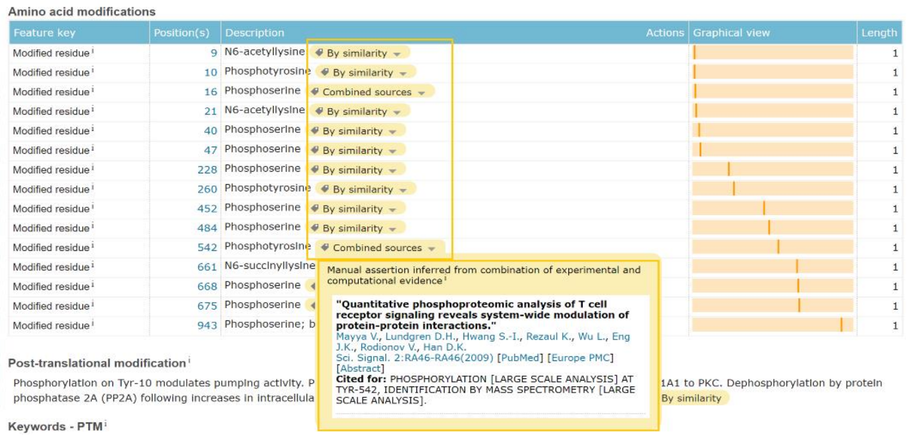
P05023

With
ATP1B1 [P05026]
TPT1 [P13693]

Figure 8

Figure 8: The binary interactions annotations shows the sodium-potassium ATPase (AT1A1) has interactions with AT1B1 and TPT1.

-Structure links 3D structure databases which can be used to find the protein’s tertiary/quaternary structure.
-Family & Domains identifies different domains present in the protein, what family or families the protein belongs to, and links family and domain databases.
-Sequences (4+) gives the one letter amino acid code sequence for the protein and each isoform. The FASTA format for each one is also linked. BLASTs and other analysis tools can be run directly for each isoform. Other potential isoforms mapped using computational analysis are identified, experimental evidence is listed, and other sequence/genome databases are linked.
-Similar Proteins lists proteins with high sequence alignment. 100% identity, 90% identity, or 50% identity can all be used as parameters to show similar proteins.
-Cross-references lists the other databases used in compiling the full annotation by UniProt.
-Entry information is the information regarding the query protein: Entry name, accession code, entry history and status.
-Miscellaneous consists of any relevant information that did not fall under the category of any of the other sections.
Note that in any of these sections, selecting the gold evidence tag will cause the associated evidence to open. Gold tags represent manually annotated evidence (Swiss-Prot), while blue tags represent automatically annotated evidence (TrEMBL). If “Reviewed” was selected when filtering results, only gold tags will be displayed. (**Figure 9**)



Amino acid modifications

Feature key	Position(s)	Description	Actions	Graphical view	Length
Modified residue	9	N6-acetyllysine By similarity			1
Modified residue	10	Phosphotyrosine By similarity			1
Modified residue	16	Phosphoserine Combined sources			1
Modified residue	21	N6-acetyllysine By similarity			1
Modified residue	40	Phosphoserine By similarity			1
Modified residue	47	Phosphoserine By similarity			1
Modified residue	228	Phosphoserine By similarity			1
Modified residue	260	Phosphotyrosine By similarity			1
Modified residue	452	Phosphoserine By similarity			1
Modified residue	484	Phosphoserine By similarity			1
Modified residue	542	Phosphotyrosine Combined sources			1
Modified residue	661	N6-succinyllysine			1
Modified residue	668	Phosphoserine			1
Modified residue	675	Phosphoserine			1
Modified residue	943	Phosphoserine; b			1

Manual assertion inferred from combination of experimental and computational evidence

"Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions."
Mayya V., Lundgren D.H., Hwang S.-I., Rezaul K., Wu L., Eng J.K., Rodionov V., Han D.K.
Sci. Signal. 2:RA46-RA46(2009) [PubMed] [Europe PMC] [Abstract]
Cited for: PHOSPHORYLATION [LARGE SCALE ANALYSIS] AT TYR-542, IDENTIFICATION BY MASS SPECTROMETRY [LARGE SCALE ANALYSIS].

Post-translational modification

Phosphorylation on Tyr-10 modulates pumping activity. P … 1A1 to PKC. Dephosphorylation by protein phosphatase 2A (PP2A) following increases in intracellula … By similarity

Keywords - PTM

Figure 9

Figure 9: Amino acid modifications shown under the PTM/Processing section. By selecting one of the gold tags, published research articles supporting the annotation of the phosphorylation of the tyrosine 542 residue is shown.

6.2. Using Other UniProt Databases

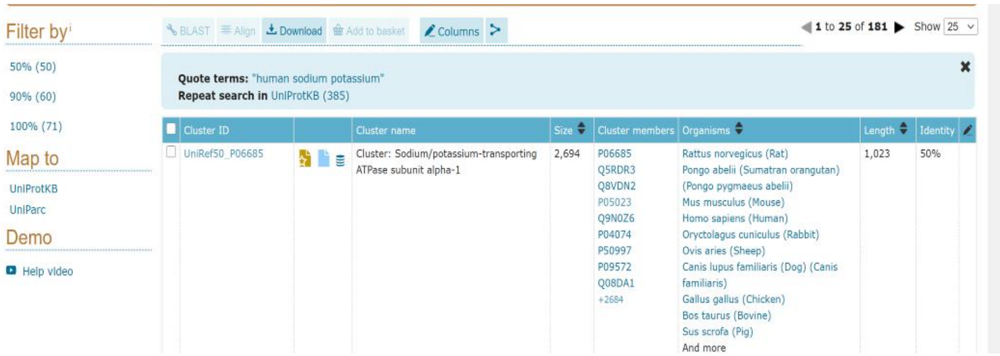
6.2.1. Similar steps as described in 6.1. can be used to run searches using different databases. Return to the UniProt home page and select the UniParc database. UniParc stands for UniProt Archive. The UniParc database is a non-redundant database which contains ONLY PROTEIN SEQUENCES; no annotation is available in the UniParc database. UniParc gives each protein a unique UPI and stores that protein only once to avoid redundancy in the database.

6.2.2. Run a search for the same protein using the accession code. For UniParc, the accession number must be used, rather than the organism followed by protein name format.

6.2.3. Once the target protein has shown up, click the UPI (found under "Entry"). This will redirect the page to the sequence page. The same tools are available at the top of the page; BLAST, alignment (only can be ran if multiple UniParc entries are selected), format, and add to basket. View the protein sequence in each format. Note that the RDF/XML format is a downloaded file.

6.2.4. Next, return to the homepage to use the UniRef database. The UniRef database stands for UniProt Reference Clusters. This database consists of clustered sets of sequences from the UniProtKB and select sequences from UniParc. This clustering of similar (and often nearly identical) sequences together limits redundancy in the database.

6.2.5. Run a search for the same protein using its accession number or name-UniRef can recognize both.



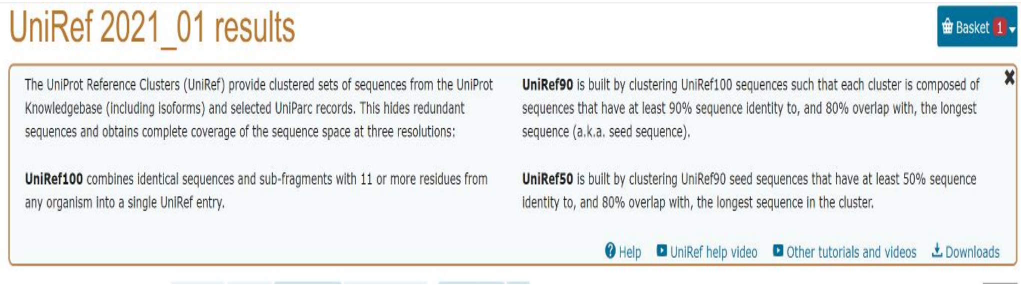
Quote terms: "human sodium potassium"
Repeat search in UniProtKB (385)

□	Cluster ID		Cluster name	Size	Cluster members	Organisms	Length	Identity
□	UniRef50_P06685		Cluster: Sodium/potassium-transporting ATPase subunit alpha-1	2,694	P06685 Q5RDR3 Q8VDN2 P05023 Q9N0Z6 P04074 P50997 P09572 Q08DA1 +2684	Rattus norvegicus (Rat) Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii) Mus musculus (Mouse) Homo sapiens (Human) Oryctolagus cuniculus (Rabbit) Ovis aries (Sheep) Canis lupus familiaris (Dog) (Canis familiaris) Gallus gallus (Chicken) Bos taurus (Bovine) Sus scrofa (Pig) And more	1,023	50%

Figure 10

Figure 10: The UniRef search results page with filtering options on the left.

On the left, the search can be filtered using UniRef100 (100%), UniRef90 (90%), or UniRef50 (50%). (**Figure 10**) As described at the top of the UniRef search page, the UniRef100 combines identical sequences and subfragments from a given organism into a single entry (**Figure 11**). UniRef90 and UniRef50 are then built by combining UniRef100 sequences of 90% or 50% identity, respectively. Thus, each consecutive cluster (100→90→50) yields less results as sequences are clustered based on a smaller percent of required sequence similarity. The user can also choose UniProtKB or UniParc to use as a map.



UniRef 2021_01 results

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records. This hides redundant sequences and obtains complete coverage of the sequence space at three resolutions:

UniRef100 combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry.

UniRef90 is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to, and 80% overlap with, the longest sequence (a.k.a. seed sequence).

UniRef50 is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to, and 80% overlap with, the longest sequence in the cluster.

Figure 11

Figure 11: The algorithm used for each level of UniRef.

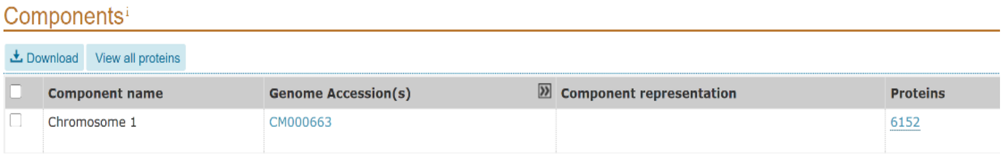
6.2.6. Select the Cluster ID. This will bring the screen to members of all the component cluster and identify the sequence which is common among the members of the cluster. From here, the same tools are available for use; BLAST, format, add to basket, and the multiple different members of the cluster can be selected and the align tool can be used.

6.2.7. Return to the home page and select the Proteome database. Proteomes are the set of proteins expressed entirely by an organism based off their sequenced genome.

6.2.8. Choose an organism and type its name in to the search box to see if their proteome is available (for example, humans.) Select search.

6.2.9. If the organism is available, select the proteome ID; if not, search until one is found.

6.2.10. The results will yield an overview of the organism's proteome and the proteome divided into chromosomes. Each chromosome will have a genome accession number followed by a number of the proteins represented by each chromosome. By selecting this number of proteins, the page will redirect to the annotated UnitProtKB database of all the proteins belonging to that genome. Select a chromosome and then select the protein number; identify how many proteins belong to that chromosome. (**Figure 12**)



Components

□	Component name	Genome Accession(s)	Component representation	Proteins
□	Chromosome 1	CM000663		6152

Figure 12

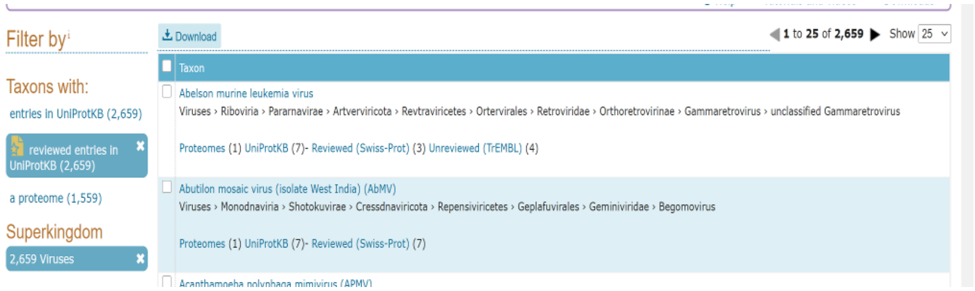
Figure 12: The chromosome number selected for a paticular organism using the proteome database search, indicating the genome accession number and how many proteins belong to that chromosome.

6.2.11. Next, explore the usage of the remaining supporting databases:

-Literature Citations will generate results of publications which are used/cited in the UniProtKB annotations.

-Cross-Ref databases consists of all the databases which are used to compile information in the UniProt annotations which are highlighted in the cross-reference seciton of each annotation.

-Taxonomy consists of organisms organized in their taxonomic heirarchial tree structure. Results can be filtered by taxons with: UniProtKB entries, reviewed UniProtKB entries, or complete available proteomes. Results can also be filtered by Superkingdom: Bacteria, Viruses, Eukarya, or Archea. (**Figure 13**)



□	Taxon
□	Abelson murine leukemia virus Viruses › Riboviria › Pararnavirae › Artverviricota › Revtraviricetes › Ortervirales › Retroviridae › Orthoretrovirinae › Gammaretrovirus › unclassified Gammaretrovirus Proteomes (1) UniProtKB (7)- Reviewed (Swiss-Prot) (3) Unreviewed (TrEMBL) (4)
□	Abutilon mosaic virus (isolate West India) (AbMV) Viruses › Monodnaviria › Shotokuvirae › Cressdnaviricota › Repensiviricetes › Geplafuvirales › Geminiviridae › Begomovirus Proteomes (1) UniProtKB (7)- Reviewed (Swiss-Prot) (7)
□	Acanthamoeba polyphaga mimivirus (APMV)

Figure 13

Figure 13: Search results using the taxonomy dataset with filtered results for reviewed UnitProbKB entries of viruses.

-Diseases compiles information on human diseases involving protein disfunction. The disease can be searched by name using the format “name:disorder” or “name:disease” (for example, type “name:alzheimers” into the search bar. Run a search for a disease.

-Subcellular location can be used to identify and find information about cellular locations, organelles, and structures. A specific location should be searched using the format “name:location” (For example, “type name:plasma membrane” into the search bar.) Run a search for a cellular location or organelle.

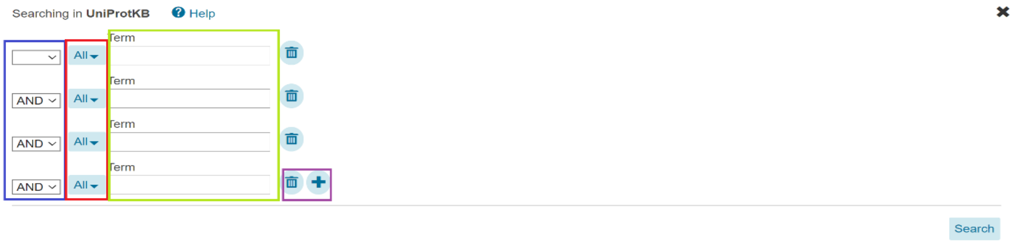
-Keywords UniProtKB entries are associated with keywords that can be used to retrieve subsets of entries. By typing the keyword in the search bar, such as “acetylation”, it will give a match to the keyword and a description of its meaning, then the option to run the search of the keyword through UniProtKB and find all matching proteins associated with the keyword. Run a search for a biological keyword.

6.3. Using Advanced Search Features

6.3.1. The user can set specific parameters in order to narrow the search field when using the UniProt databases. To use the advanced search feature, return to the UniProt home page.

6.3.2. Advanced search can be used on any of the databases, but the means by which the search can be narrowed will differ for each database due to the nature of what each database holds. Use the UniProtKB database as it has the most annotation and most ways in which the search field can be narrowed.

6.3.3. Click on the “Advanced” drop down button next to the “Search” button. The advanced search input page will be displayed. Notice there are four separate input bars which can be specified. This allows for the user to run a search taking multiple inputs into account in relation with each other (note that additional slots for input can be added using the + button; slots can be removed using the trash button.) The first box is a dropdown which can be changed to AND or NOT. This allows the user to simultaneously search for a specific result while filtering out results which they do not want. The second box is a dropdown which lists all the possible ways in which the search can be specified, such as organism, function, sequence, etc. The “Term” box is where the user specifies the parameters for that specific search function. (**Figure 14**)

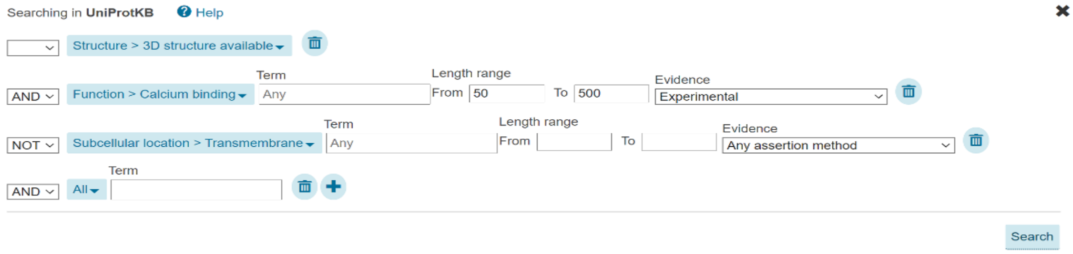

Searching in UniProtKB Help

	All	Term
AND	All	Term
AND	All	Term
AND	All	Term

Figure 14

Figure 14: The advanced search function on UniProtKB.

6.3.4. Fill out the advanced search specifying three different sets of parameters. Use at least one “NOT” option to exclude a certain set of results. (**Figure 15**)

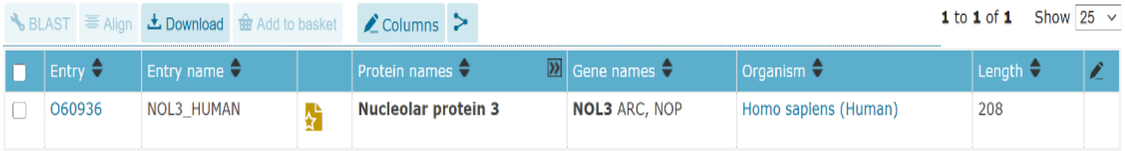

Searching in UniProtKB Help

	Structure > 3D structure available				
AND	Function > Calcium binding	Term: Any	Length range: From 50 To 500	Evidence: Experimental	
NOT	Subcellular location > Transmembrane	Term: Any	Length range: From To	Evidence: Any assertion method	
AND	All	Term			

Figure 15

Figure 15: An advanced search specifying proteins with an available 3D structure, AND that bind calcium with a range of 50-500 (and evidence specified
to be experimentally determined), that are NOT found in the plasma membrane.

6.3.5. Select “Search.” The resulting proteins which fit the specifications will appear. (**Figure 16**)



1 to 1 of 1 Show 25

	Entry	Entry name		Protein names	Gene names	Organism	Length
□	O60936	NOL3_HUMAN		Nucleolar protein 3	NOL3 ARC, NOP	Homo sapiens (Human)	208

Figure 16

Figure 16: The resulting protein from the advanced search conducted from Figure 14.

6.3.6. The protein of interest can be further examined by selecting its entry code to view all the available annotations previously discussed in this module.

7. TROUBLE SHOOTING

7.1. When running an advanced search, if the search keeps leading to no results, make the parameters broader (such as a longer sequence length). Removing a requirement for experimental evidence or for 3D structure will also lead to a greater number of results, if these are not a critical part of your search.

8. REFERENCES

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., . . . Yeh, L. L. (2004, January 1*). Uniprot: The universal protein knowledgebase.* National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308865/.

Pundir, S., Magrane, M., Martin, M. J., O'Donovan, C., & The UniProt Consortium. (2015, June 19). *Searching and navigating Uniprot databases.* National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4522465/.

UniProt Consortium: European Bioinformatics Institute, Protein Information Resource, SIB Swiss Institute of Bioinformatics. (n.d.). UniProt. Retrieved from https://www.uniprot.org/.

9. MODULE MASTERY TASK

Use the UniProt databases to compile as much research as you can about a protein of interest.

9.1. UniProt Knowledgebase: Protein Annotation

9.1.1. Choose a protein of interest and search it using UniProtKB. Compile all the following information about your protein on a document:

-Protein accession number, entry number, name, organism name, and sequence length

-A descriptive overview of the proteins general function and any catalytic activity (include mechanisms/net reactions) can use screenshot or drawing to include mechanism

-What kingdoms of life/organisms the protein belongs to

-Where the protein is found in the cell, including organelles

-The qualitative topology/structure of the protein (transmembrane domains, secondary structures)

-What diseases the protein is involved in

-Uses of the protein in biotechnology

-What post-translational modifications the protein has (Glycosylation, phosphorylation, etc.)

-What other proteins the protein interacts with

-Find the 3D structure of the protein linked to other databases if available (provide a link)

-What domains are present in the protein, and what family/superfamily(s) the protein belongs to

-Copy the sequence in the normal text format and the FASTA format

-What similar proteins are identified

-Any additional interesting miscellaneous information

-Include a link to at least one publication supplying experimental evidence

-Run a BLAST for the protein and make note of several proteins which show alignment.

-Run an alignment if isoforms of the protein are available. Take a screenshot and mark the aligning segments.

9.2. Other UniProt Databases

9.2.1. Choose an organism and search the Proteome database. Search until you find an organism with an available proteome. Make note of how many proteins are in the organism’s proteome and select one, making a note of its name and general function.

9.2.2. Use the accession code to run a UniParc and copy the FASTA sequence of the protein and save it.

9.2.3. Run a UniRef on the protein using its accession number. The UniRef100 cluster should have only 1 member, but select the UniRef90 and UniRef50 clusters and make note of how many members are in each cluster. What does this tell you about this protein? Does it have many “relatives” based on 90%/50% sequence similarity or not?

9.3. Advanced Search

9.3.1. You are working in a lab studying gene expression. You are interested in the transcription factor of a certain gene as you have been trying to regulate the expression of the gene product. You believe the transcription factor to be glycosylated because in mild glucose starvation conditions, the gene product is down regulated. As the transcription factor has not been identified, you know there is no 3D structure available. Use the advanced search to input 3 parameters which may help you narrow down your search of this transcription factor.