Published in final edited form as:

Curr Protoc. 2023 March ; 3(3): e697. doi:10.1002/cpz1.697.

UniProt Tools: BLAST, Align, Peptide search and ID mapping

Rossana Zaru[1,*], **Sandra Orchard**[1], **The UniProt Consortium**[1,2,3,4]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom

[2]Swiss Institute of Bioinformatics, University Medical Center, Geneva, Switzerland

[3]Protein Information Resource, Georgetown University Medical Center, Washington, D.C

[4]Protein Information Resource, University of Delaware, Newark, Delaware

Abstract

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data (UniProt Consortium, 2023). The UniProt Web site receives ~800,000 unique visitors per month and is the primary means to access UniProt. Along with various datasets that you can search, UniProt provides four main tools. These are the 'BLAST' tool for sequence similarity searching, the 'Align' tool for multiple sequence alignment, the 'Peptide search' for retrieving proteins containing a short peptide sequence and the 'Retrieve/ID Mapping' tool for using a list of identifiers to retrieve UniProtKB proteins and to convert database identifiers from UniProt to external databases or vice versa. This unit provides four basic protocols, four alternate protocols, and three support protocols for using UniProt tools. © by John Wiley & Sons, Inc.

BASIC PROTOCOL 1: Basic Local Alignment Search Tool in UniProt

ALTERNATE PROTOCOL 1: Basic Local Alignment Search Tool Through UniProt Text search results pages

SUPPORT PROTOCOL 1: Basic Local Alignment Search Tool Through UniProt Basket

BASIC PROTOCOL 2: Multiple Sequence Alignment in UniProt

ALTERNATE PROTOCOL 2: Align Tool Through UniProt Results Pages and Entry Pages

SUPPORT PROTOCOL 2: Align Through UniProt Basket

BASIC PROTOCOL 3: Peptide Search in UniProt

BASIC PROTOCOL 4: Batch Retrieval and ID Mapping in UniProt

ALTERNATE PROTOCOL 4: Retrieve/ID Mapping Tool Through UniProt TEXT SEARCH Results Pages and, BLAST and ALIGN RESULTS Pages

ALTERNATE PROTOCOL 4: Retrieve/ID Mapping Tool Through UniProt Basket

SUPPORT PROTOCOL 4: Retrieve/ID Mapping Tool Through UniProt SEARCH Box

[*]Corresponding author: rzaru@ebi.ac.uk.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

UniProt; search; navigation; tutorial

INTRODUCTION

UniProt, or the Universal Protein Resource, provides an up-to-date, comprehensive body of protein information at a single site (UniProt Consortium, 2023). The UniProt Knowledgebase (UniProtKB) delivers high-quality functional information for a selected set of protein sequences. To build upon this protein data and to aid analysis, UniProt provides four main tools: 'BLAST' (Basic Local Alignment Search Tool), 'Align' multiple sequence alignment tool, 'Peptide search' and 'Retrieve/ID Mapping' for batch retrievals of UniProt entries and ID mapping between UniProt and external databases. These tools are available on their own dedicated pages on the UniProt Web site and are also accessible directly from other parts of the Web site such as the basket, search/tool results pages, and protein entry pages. Having these tools in the UniProt Web site creates an integrated hub of data and analysis tools, allowing both to leverage each other. For example, if you come across sequences while browsing UniProt databases that you would like to BLAST or align, you can select your sequence and submit it to the relevant tool directly. Consecutively, results from tools provide links directly to all relevant data from UniProt and allow you to filter by attributes like whether you are looking for reviewed (UniProtKB/Swiss-Prot) or unreviewed (UniProtKB/TrEMBL) UniProtKB entries, entries with 3D structures, or entries that are part of a proteome, among others. All data in UniProtKB is freely available to users.

The UniProt Web site can be accessed at http://www.uniprot.org/. The following protocols describe how you can navigate to UniProt tools and use them in your analysis.

BASIC PROTOCOL 1

Basic Local Alignment Search Tool in UniProt

The UniProt Web site provides a Basic Local Alignment Search Tool (BLAST; UniProt Consortium, 2022) which finds regions of local similarity between sequences. This can be used to infer functional and evolutionary relationships between sequences as well as to help identify members of gene families. The BLAST tool page can be reached from a link in the header on all pages of the UniProt Web site.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Click on the 'BLAST' link, which is on the left-hand side of the header bar on all UniProt pages as shown in Figure 1. Alternatively, 'BLAST' can be accessed by clicking on the corresponding tile in the Analysis Tools section on the home page.

   You will see the BLAST input page as shown in Figure 2.

3. To run a BLAST search, enter either a UniProt identifier into the first input box or, a protein or nucleotide sequence into the second input box provided. For example, enter ‘MTMR1_CAEEL’ in the first input box.

   These are the only mandatory input fields.

   You can choose to change parameters from a number of advanced options, as shown in Table 1.

   One of these options is the ability to select the target database to run your search against. Choosing the most appropriate target database allows you to tailor your BLAST search to the type of results you are interested in and to speed up the search. The dropdown menu allows you to restrict your BLAST query to 10 different datasets. The option ‘UniProtKB reference proteomes + UniProtKB/Swiss Prot entries’ is selected by default. In this case, the BLAST will run against proteins that belong to reference proteomes as well as proteins that have been reviewed (UniProtKB/Swiss Prot).

   If you would like to find similar proteins to your query which are well annotated in order to infer a potential function for your protein of interest, you can restrict your target database to ‘UniProtKB Swiss-Prot’ to search only reviewed entries (UniProtKB/Swiss-Prot) from UniProtKB. If you would like to retrieve closely related proteins for which a PDB 3D structure is available, you can select ‘UniProtKB with 3D structure (PDB)’ as the target database to run your BLAST search against.

   If you would like to speed up the BLAST search, you can choose to search UniRef clusters, which reduces redundancy by grouping sequences based on identity (Suzek et al. 2015). There are 3 clusters UniRef100, UniRef90 and UniRef50 which contain protein entries with 100%, 90% and 50% sequence identity, respectively. For example, if you would like to find representative proteins that closely match your query you can choose to run your search against UniRef90. For more information about UniRef clusters see (https://www.uniprot.org/help/uniref). You can also choose to search against the sequence archive UniParc which is a repository of all known protein sequences, including those that have been removed from UniProtKB if highly redundant or of dubious provenance.

   If you are looking for sequence similarity matches from a particular taxonomy group or species, you can use the ‘Restrict by taxonomy’ box. For example, you can find out if a human protein has orthologs in other mammals or specifically in mouse by restricting your target data to UniProtKB entries belonging to the group ‘Mammals’ or the species ‘Mus musculus’, respectively. You can either start typing the species or the group name, which will open a drop-down menu with suggestions where you can select your group or species, or you can enter the NCBI taxonomic identifier for your group or species of interest. You can restrict the search to more than 800’000 different taxa.

Other parameters that you can change are the E-threshold, matrix, filtering, gapped (yes or no), number of hits and HSP per hit. A brief description of each parameters is provided in Table 1. To better understand the effects of E-threshold, matrix, and gapped search changes, refer to UNIT 3.4.

4. Click on the 'Run BLAST' button to execute your query.

   This will take you to the Tool results page as shown in Figure 3 where you can see the progression of your BLAST search. Once the search is finished, click on 'completed' to open the BLAST result page which is shown in Figure 4.

ALTERNATE PROTOCOL 1

Basic Local Alignment Search Tool Through UniProt Text Search Results Pages

Queries can be submitted to the BLAST tool directly through UniProt search result pages as and when you come across a sequence you would like to analyze using a sequence similarity search. This allows for a flexible workflow between browsing data and analyzing it.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Choose the search dataset using the dropdown to the left of the search box. Select either 'UniProtKB', 'UniRef,' or 'UniParc'.
3. Enter a query in the search box, for example 'insulin,' and click on the 'Search' button.

   Before seeing the results, you will be asked how you would like to view them. There are 2 display options, 'Cards' or 'Table' which offer different levels of information. The 'Table' view provides the list of protein entries matching your query with the protein name(s), gene name(s), species and protein length in a table format. The 'Cards' view contains the same information but in addition it provides an overview of the type of annotation in the entry, for example the number of publication or how many 3D structures are associated with the entry. Both views provide checkboxes to select entries.

   Figure 5 shows the search results page in a table format.
4. To run a BLAST search for a protein in your search results, click on the checkbox in the left-hand side column for that protein row.
5. Click on the BLAST button just above the search results table, as shown in Figure 6 which will open the BLAST page from where you can launch the BLAST as described in the previous section.
6. Alternatively, click on a UniProt entry in your results table to be taken to the entry page, which also provides a BLAST button for direct submission, as shown in Figure 7.

SUPPORT PROTOCOL 1

Basic Local Alignment Search Tool Through UniProt Basket

Queries can be submitted to the BLAST tool directly through the UniProt basket feature. The UniProt basket allows you to store entries from UniProtKB, UniRef, or UniParc. You can use the basket to build a set of your proteins across different searches. The basket then allows you to download your dataset to access analysis tools (i.e., BLAST, Align, and Retrieve/ID Mapping). Your basket is saved as long as you do not clear your cookies.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Choose the search dataset through the dropdown to the left of the search box. Select either ‘UniProtKB’, ‘UniRef’, or ‘UniParc’.
3. Enter a query in the search box, for example ‘insulin,’ and click on the ‘Search’ button.

   You will see a search results page as shown in Figure 5.
4. For the entry of interest, click on the checkbox to the left of its accession, and then click on ‘Add’ next to the basket icon button at the top of the results table as shown in Figure 8.
5. When you are ready to analyze the entries in your basket, click on the basket to open it.

   Your entries will be under their dataset tab (UniProtKB, UniRef, or UniParc).
6. Click on the checkbox to the left of your entry of interest, and then click on ‘BLAST,’ as shown in Figure 9.

BASIC PROTOCOL 2

Multiple Sequence Alignment in UniProt

The UniProt Web site provides a multiple sequence alignment tool for proteins called ‘Align.’ This tool runs the Clustal Omega algorithm to find areas of similarity in the entries being aligned. This can be used to find conserved residues and regions that can help infer evolutionary and functional relationships (Simossis et al., 2003).

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Click on the ‘Align’ link, which is available in the header bar on all UniProt pages as shown in Figure 1. Alternatively, ‘Align’ can be accessed by clicking on the corresponding tile in the Analysis Tools section on the home page.

   You will see the Align input page as shown in Figure 10.

3. To execute the multiple sequence alignment, enter either the UniProt identifiers into the first input box or the protein sequences in FASTA format into the second sequence input box provided and click ‘Align sequences.’ For example, paste the following UniProt identifiers ‘MTMR1_HUMAN’, ‘MTMR1_MOUSE’ and ‘MTMR1_CAEEL’ into the first box.

   These are the only mandatory input fields.

   For each UniProt identifiers entered, the corresponding sequence in FASTA format is retrieved and appears in the second box as shown in Figure 11.

   In the advanced parameters you have 2 options that allows you to choose the output sequence order or change the number of iterations. The iteration parameter refers to the number of guide-tree/HMM iterations which can be performed after the initial alignment. It involves re-running the alignment while removing and adding back sequences to see if the alignment score can be improved. It is appropriate to increase the number of iterations when the default number of iterations (here 0) does not produce a satisfactory alignment; however, it is time consuming. While for protein sequences that are similar in length and identity, no iteration is often sufficient to generate an accurate alignment, when aligning a large number of sequences or sequences with low similarity increasing the number of iterations may help to improve accuracy.

4. This will take you to the Tool results page as shown in Figure 12 where you can see the progression of your alignment. Once the alignment is finished, click on ‘completed’ to open the Align result page which is shown in Figure 13.

ALTERNATE PROTOCOL 2

Align Tool Through UniProt Results Pages and Entry Pages

Queries can be submitted to the Align tool directly through UniProt search result pages.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Select ‘UniProtKB’ as the search dataset through the dropdown to the left of the search box.
3. Enter a query in the search box, for example ‘insulin,’ and click on the ‘Search’ button.

   You will see a results page as shown in Figure 5.
4. Click on two or more checkboxes to align these protein entries, as shown in Figure 14.
5. Click on the ‘Align’ button just above the search results table. This will take you to the Tool results page as described previously.

SUPPORT PROTOCOL 2

Align Through UniProt Basket

Queries can be submitted to the Align tool directly through the UniProt basket feature.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Select ‘UniProtKB’ as the search dataset through the dropdown to the left of the search box.
3. Enter a query in the search box, for example ‘insulin,’ and click on the ‘Search’ button.

   You will see a results page as shown in Figure 5.
4. For your entries of interest, click on the checkboxes to the left of their accession numbers in the results table, and then click on ‘Add’ next to the basket icon at the top of the results table.

   You can store UniProt entries in a basket over multiple search sessions and then align them later.
5. When you are ready to analyze the entries in your basket, click on the basket to open it, as shown in Figure 15.
6. Click on the checkboxes to the left of the entries you would like to align and then click on ‘Align.’ This will take you to the Tool results page as described previously.

   You need to select two or more entries to be able to create a multiple sequence alignment.

BASIC PROTOCOL 3

PEPTIDE SEARCH in UniProt

The UniProt website provides a tool that allows you to upload short peptide sequences of at least 3 residues and find all UniProtKB sequences which have an exact match to the query sequence. These peptides sequences can come from proteomics experiments or for the design of peptides for antibody production, for example.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Click on the ‘Peptide search’ link, which is available in the header bar on all UniProt pages, as shown in Figure 1. Alternatively, ‘Peptide search’ can be access by clicking on the corresponding tile in the Analysis Tools section on the home page.

   You will see the Peptide search input page as shown in Figure 16.

3. Enter your peptide(s) into the input box as shown in Figure 17 and click on the 'Run Peptide search' button.

   In the Peptide search input page, you have the option to restrict your search to a specific species using the 'restrict by taxonomy' box. For example, in Figure 17, the search has been restricted to mouse entries.

   In the advanced parameters section, you can also specify whether you would like the tool to treat isoleucine and leucine as equivalent or to restrict the search to Swiss-Prot reviewed entries only.

4. This will take you to the Tool Results page as shown in Figure 18 where you can see the progression of your peptide search. Once the search is finished, click on 'completed' to open the result page which is shown in Figure 19.

BASIC PROTOCOL 4

Batch Retrieval and Id Mapping in UniProt

The UniProt Web site provides a tool that allows you to upload a list of UniProt identifiers and batch retrieve all the corresponding UniProt entries. It allows you to convert or 'map' your identifiers from UniProtKB to over 100 external databases that UniProt is cross-referenced to and *vice versa* (e.g., Ensembl, PDB, Refseq; Huang et al., 2011). This covers a number of databases from different categories including sequence, 3D structure, protein-protein interaction, protein family and groups, chemistry, post-translational modifications, and genome annotations, among others. This tool is called 'Retrieve/ID Mapping.'

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.

2. Click on the 'ID Mapping' link, which is available in the header bar on all UniProt pages, as shown in Figure 1. Alternatively, 'ID Mapping' can be accessed by clicking on the corresponding tile in the Analysis Tools section on the home page.

   You will see the 'Retrieve/ID Mapping' input page, as shown in Figure 20.

3. If you have a list of UniProt IDs (for example, Q9VR99, Q9VEZ5, Q03017, P48607) and would like to retrieve functional information or look at the sequences for all of these entries, paste your list into the input box provided or upload a file.

4. Leave 'From' and 'To' as 'UniProtKB' and click the 'Map IDs' button. This will take you to the Tool results page as shown in Figure 21 where you can see the progression of your mapping. Once the mapping is finished, click on 'completed' to open the Results page as shown in Figure 22.

5. If you have a list of UniProt IDs and would like to map them to IDs from an external database such as Ensembl, RefSeq or one of the model organism databases (or vice versa), upload or paste in your IDs. Then, select the source

database in the ‘From’ dropdown and the target database in the ‘To’ dropdown menus.

For example, if you would like to retrieve the corresponding IDs from the FlyBase database for the 4 *D. melanogaster* proteins mentioned above, paste the UniProt IDs Q9VR99, Q9VEZ5, Q03017, P48607 in the input box and select ‘FlyBase’ in the ‘To database’ dropdown menu as shown in Figure 23.

6. Click the ‘Map IDs’ button as shown in Figure 23.

   You will get a results page with a table showing the mapping between your input IDs and their corresponding IDs from your selected database, as shown in Figure 23.

ALTERNATE PROTOCOL 4

Retrieve/Id Mapping Tool Through UniProt TEXT SEARCH Results Pages and, BLAST and ALIGN RESULTS Pages

Queries can be submitted to the ‘Retrieve/ID Mapping’ tool directly through UniProt search result pages, BLAST results pages or Align results pages.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Select ‘UniProtKB’ as the search dataset through the dropdown to the left of the search box.
3. Enter a query in the search box, for example ‘insulin,’ and click on the ‘Search’ button.

   You will see a results page as shown in Figure 5.
4. Click on two or more checkboxes as shown in Figure 24.

   Similarly, proteins can be selected in the BLAST or Align results pages.
5. Click on the ‘MapIDs’ button just above the search results table. This will take you to the ‘ID mapping’ page as described previously.

   Similarly, the ‘MapIDs’ button is available from the BLAST or Align results pages.

ALTERNATE PROTOCOL 4

Retrieve/Id Mapping Tool Through UniProt Basket

Queries can be submitted to the ‘Retrieve/ID Mapping’ tool directly through UniProt basket. The UniProt basket allows you to store entries from UniProtKB, UniRef, or UniParc. You can use the basket to build a set of your proteins across different searches. The basket then allows you to download your dataset to access analysis tools (i.e., BLAST, Align, and Retrieve/ID Mapping). Your basket is saved as long as you do not clear your cookies.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Choose ‘UniProtKB’ for the search dataset using the dropdown to the left of the search box.
3. Enter a query in the search box, for example ‘insulin,’ and click on the search button.

   You will see a results page as shown in Figure 5.
4. For your entries of interest, click on the checkboxes to the left of their accession numbers in the results table, and then click on ‘Add’ next to the basket icon at the top of the results table as shown in Figure 8.
5. When you are ready to analyze the entries in your basket, click the basket to open it.
6. To map the UniProt IDs to an external database from the basket, click on the checkboxes to the left, and click on the ‘Map IDs’ button in the basket, as shown in Figure 25. This will take you to the ‘ID mapping’ page as described previously.

SUPPORT PROTOCOL 4

Retrieve/Id Mapping Tool Through UniProt SEARCH BOX

Queries can be submitted to the ‘Retrieve/ID Mapping’ tool directly through the UniProt search box.

Necessary Resources—A computer with an up-to-date Web browser

1. Go the UniProt home page at http://www.uniprot.org/.
2. Click on ‘List’ on the left-hand side of the general search box as shown in Figure 26. This will open the ‘Retrieve/ID Mapping’ box where you can paste the IDs that you would like to map.

GUIDELINES FOR UNDERSTANDING RESULTS

UniProt BLAST Results

The UniProt BLAST results appear as shown in Figure 4. The results page provides: (1) a left-hand side panel with filters; (2) a results table which contains the protein entries (hits) that match your query. The left-hand side panel allows you to filter the results based on the % identity, the score or the E-value. You can also filter for reviewed entries (UniProtKB/Swiss-Prot), unreviewed entries (UniProtKB/TrEMBL), entries with specific features such as proteins with a 3D structure, or entries from a specific organism. For example, Figure 27 shows how you can use the filter by taxonomy to display only mouse entries.

In the results table, the protein entries are sorted according to their score in descending order. For each entry, various information is displayed including the name, the organism

and the protein length. The last column on the left provides a basic graphical view of the sequence alignment showing the % of identity, the score and the E-value in small boxes. Clicking on the sequence box opens a window at the bottom of the page where you can see a detailed view of the alignment between the query and the matched entry, as shown in Figure 28. The alignment can be explored as described in the UniProt Align Results section below.

On the top of the page, various tabs allow you to view the results according to their 'Taxonomy', 'Hit Distribution' or as a 'Text Output'.

You can download your alignment in various formats by clicking on the 'Download' button. You can also select one or multiple entries by clicking the checkboxes next to them to run a BLAST, an alignment or store the entries in the basket for further analysis. Using the 'Customize columns' button above the table allows you to change the type of information displayed. For example, in Figure 29, the subcellular location is shown.

The 'API request' tab provides you with the code to run the current job with the same input on the command line using *curl*.

You can save the URL of your sequence alignment to access it at any time for up to 7 days from when you first ran the query. The 'Input Parameters' tab provides detailed information about the alignment parameters including the job identifier which can be used to access your sequence alignment at any time for up to 7 days from when you first ran the query.

UniProt Align Results

Multiple sequence alignments can help understand evolutionary conservation of structurally and functionally important regions of protein sequences (Simossis et al., 2003). To obtain meaningful results and minimize errors in the alignment, it is necessary to align sequences that are likely to be related to each other.

The UniProt Align results appear as shown in Figure 13. The result page displays by default the full sequence alignment ('wrapped' view). You can change how the alignment is displayed by selecting one of the 'view' options above the sequence on the right. The 'overview' mode allows you to zoom in using the grey sequence toggles and explore a specific region of the alignment, as shown in Figure 30.

Above the aligned sequence on the left, there are two dropdown menus that allow you to select sequence annotations (i.e., domains, sites) to view them highlighted across the aligned sequences. For example, in Figure 31, the active site has been selected from the 'Select annotation' and is shown highlighted in the sequence. Clicking on the active site pictogram opens a box where you could find additional information about the site. You can also highlight sequence features by amino acid properties (i.e., hydrophobicity). Both menus are available in the 'overview' and 'wrapped' views.

On the top of the result page there are various tabs that give you the option to change how the alignment is visualized. The alignment can be viewed as 'Trees', as shown in Figure 32, as 'Percent Identity Matrix', or as 'Text Output'.

You can download your alignment in various formats by clicking on the ‘Download’ button. You can also select one or multiple entries by clicking the checkbox next to them to run a BLAST, a new alignment or store the entries in the basket for further analysis.

The ‘API request’ tab provides you with the code to run the current job with the same input on the command line using *curl*.

You can save the URL of your sequence alignment to access it at any time for up to 7 days from when you first ran the query. The ‘Input Parameters’ tab provides detailed information about the alignment parameters including the job identifier which can be used to access your sequence alignment at any time for up to 7 days from when you first ran the query.

UniProt Peptide Search Results

The Peptide Search results provide UniProtKB entries with sequences that contain an exact match for each peptide that you searched for.

The Results page for the peptide search is shown in Figure 19. For each protein entry row, the position of the peptide in the sequence and the sequence of the peptide is provided in the ‘Match’ column. The filters on the left-hand side can be used to restrict the search to specific features. For example, you can display only UniProtKB/Swiss-Prot/reviewed entries.

Protein entries can be selected by clicking the checkboxes and then downloaded or stored in the basket for analysis at a later time point.

The ‘API request’ tab at the top of the results page provides you with the code to run the current job with the same input on the command line using *curl*.

You can also save the URL of your peptide search to access it at any time for up to 7 days from when you first ran the query. The ‘Input Parameters’ tab provides detailed information about the parameters, including the job identifier which can be used to access your peptide result table at any time for up to 7 days from when you first ran the query.

UniProt Retrieve/ID Mapping Results

If you retrieve a batch of UniProt entries for a list of IDs using this tool, you will get a results page as shown in Figure 22. This results page provides filters in the left-hand side panel and a main results table. The results table starts with a column titled ‘From’ that shows your input identifiers. The next columns in the results table show information from the corresponding UniProt entries that were found. You can edit these columns by clicking on the ‘Customize Columns’ button above the results table. You can also run tools such as ‘Blast’ and ‘Align’ and add entries to your basket by selecting the corresponding checkboxes and then clicking on the buttons available at the top of the results page. You can download the full results table or just the list of identifiers using the ‘Download’ button.

If you use the tool to map UniProt IDs to external database IDs (or vice versa; Huang et al., 2011), you will get a results table with two columns showing your input IDs and the corresponding mapped IDs, as shown in Figure 23. You can use the ‘Download’ button to download your results.

The ‘API request’ tab provides you with the code to run the current job with the same input on the command line using *curl*.

You can save the URL of your ID mapping to access it at any time for up to 7 days from when you first ran the query. The ‘Input Parameters’ tab provides detailed information about the parameters including the job identifier which can be used to access your mapping result table at any time for up to 7 days from when you first ran the query.

COMMENTARY

Background Information

UniProt aids scientific discovery by collecting, interpreting, and organizing information so that it is easy to access and use. In addition to providing data through various datasets, UniProt also provides tools to help researchers analyze this data. UniProtKB is the central hub for the collection of functional information and other rich annotations on proteins. It is further divided into the Reviewed (UniProtKB/Swiss-Prot), expertly annotated section and the Unreviewed (UniProtKB/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 tools UniProt provides are BLAST, Align, Peptide Search and Retrieve/ID Mapping. The UniProt Web site was designed following a user-centered design process and is flexible, powerful, and user friendly. It provides many ways of accessing these tools. Using tools within UniProt, you can easily chain activities by (1) searching for data; (2) running a BLAST search for a sequence in your results; (3) running a multiple sequence alignment on sequences in your BLAST results; and then (4) mapping the IDs of these sequences to an external database.

UniProt provides training material through the European Bioinformatics Institute (EMBL-EBI) online training portal, including a quick tour (https://www.ebi.ac.uk/training/online/courses/uniprot-quick-tour/) and a detailed course (https://www.ebi.ac.uk/training/online/courses/uniprot-exploring-protein-sequence-and-functional-info/). UniProt also provides short video tutorials embedded in the Web site, and they are also available on the YouTube channel at https://www.youtube.com/uniprotvideos.

Critical Parameters

The Tools dashboard is where all the results of the tools queries that you have run are stored for 7 days from when you first ran the query. You can access the dashboard by clicking the toolbox icon on the right-hand side of the UniProt head banner as shown in Figure 33. For each job you have the option to resubmit the query with the possibility to modify the

parameters, store the job for more than 7 days (this requires the user to not clear UniProt website cookies) or delete the job.

Acknowledgements

This work was supported by the National Human Genome Research Institute (NHGRI); Office of Director (OD/DPCPSI/ODSS); National Institute of Allergy and Infectious Diseases (NIAID); National Institute on Aging (NIA); National Institute of General Medical Sciences (NIGMS); National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK); National Eye Institute (NEI); National Cancer Institute (NCI); National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health under Award Number [U24HG007822].

Data availability statement

The data in UniProt is freely available and can be accessed at https://www.uniprot.org.

Literature Cited

Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, Chen Y, and Wu CH 2011. A comprehensive protein-centric ID mapping service for molecular data integration. Bioinformatics 27:1190–1191. doi: 10.1093/bioinformatics/btr101. [PubMed: 21478197]

Ladunga I 2009. Finding homologs in amino acid sequences using network blast searches. Curr. Protoc. Bioinformatics 25:3.4:3.4.1–3.4.34. doi: 10.1002/0471250953.bi0304s25.

Simossis V, Kleinjung J, and Heringa J 2003. An overview of multiple sequence alignment. Curr. Protoc. Bioinformatics 3.7:3.7.1–3.7.26. doi: 10.1002/0471250953.bi0307s03.

Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH and UniProt Consortium. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6):926–32. doi: 10.1093/bioinformatics/btu739. [PubMed: 25398609]

UniProt Consortium, 2023. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 10.1093/nar/gkac1052.

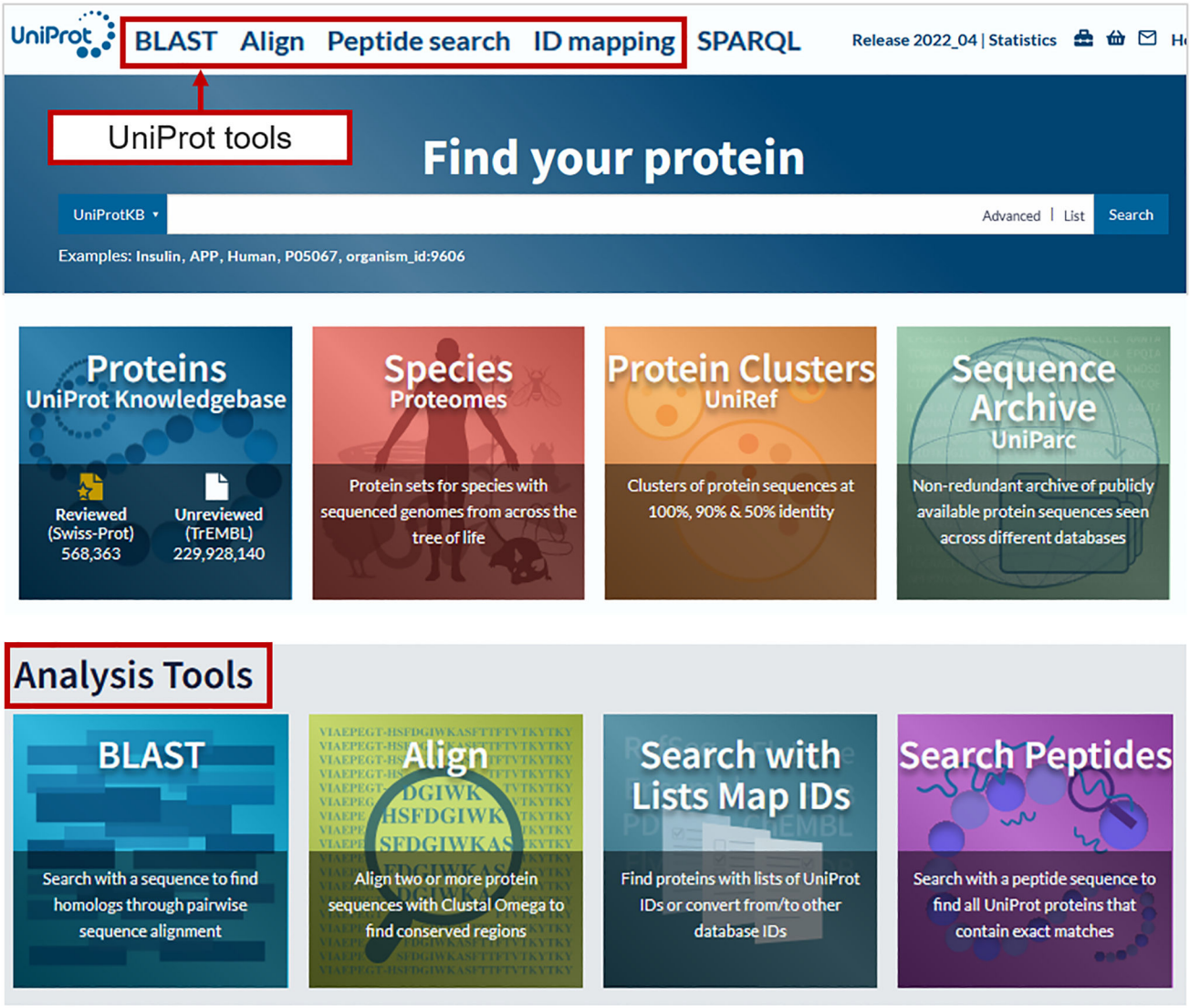

Figure 1.
Tools link in UniProt Web site header.

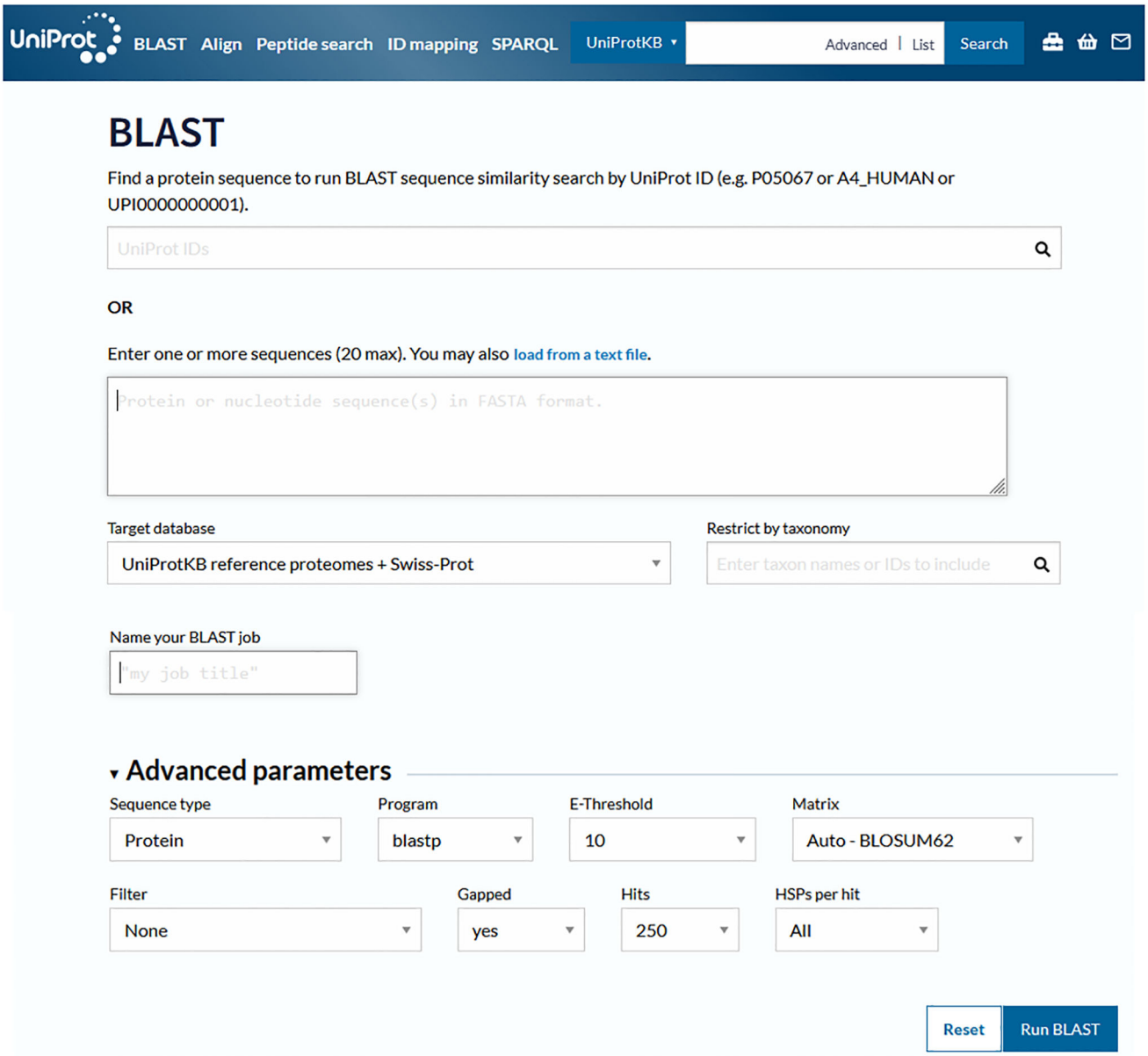

Figure 2.
BLAST query input page.

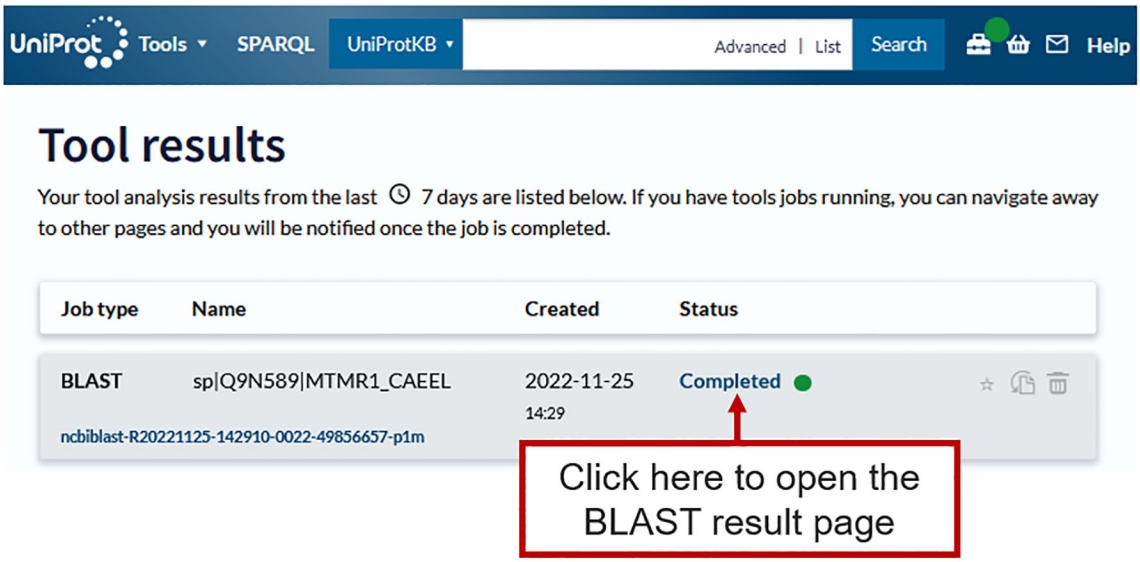

Figure 3.
Tool results dashboard page.

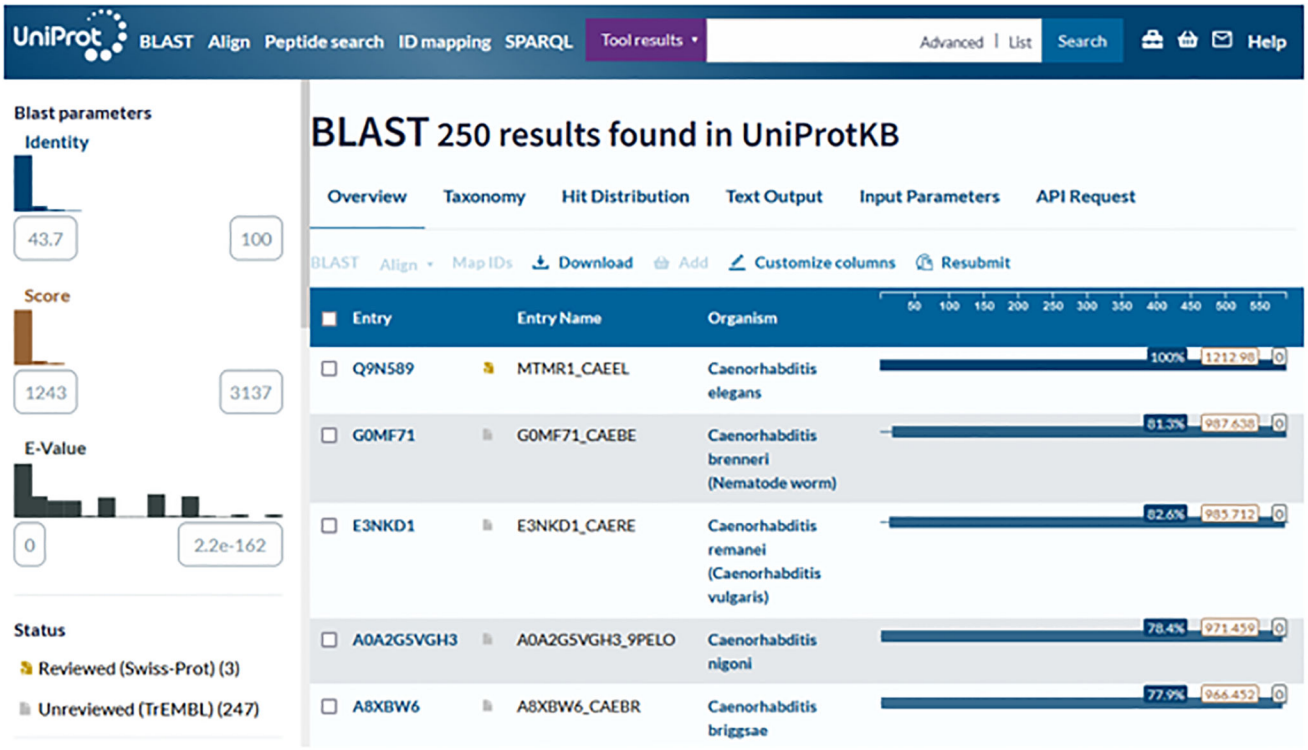

Figure 4.
BLAST results page.

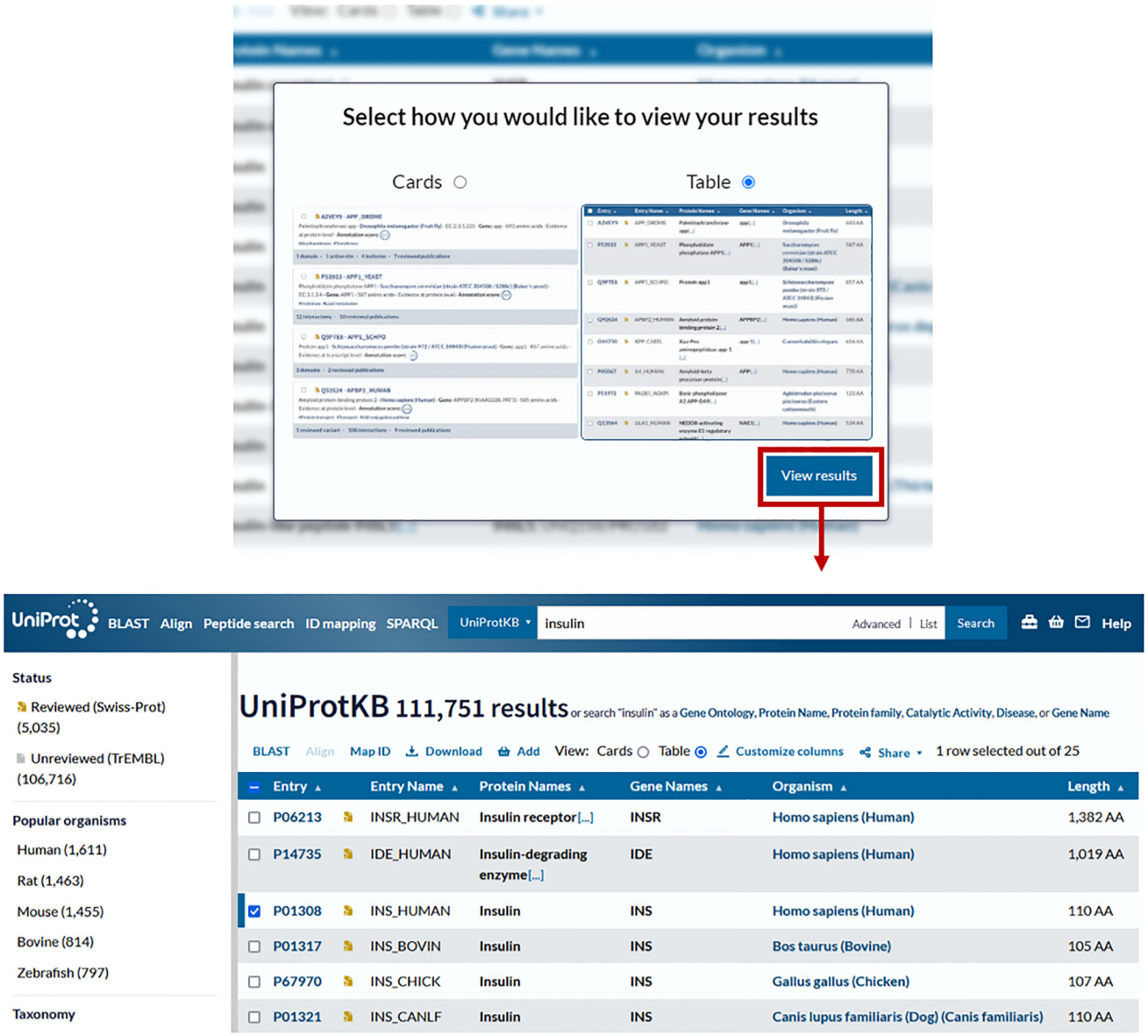

Figure 5.
UniProtKB search results page.

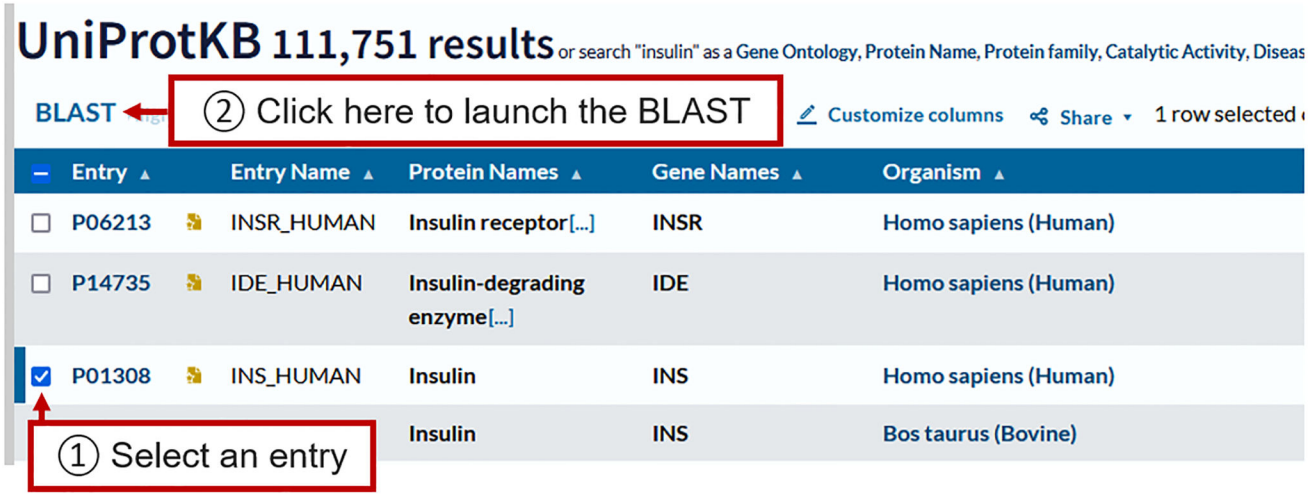

Figure 6.
Selecting a protein to BLAST from the results page.

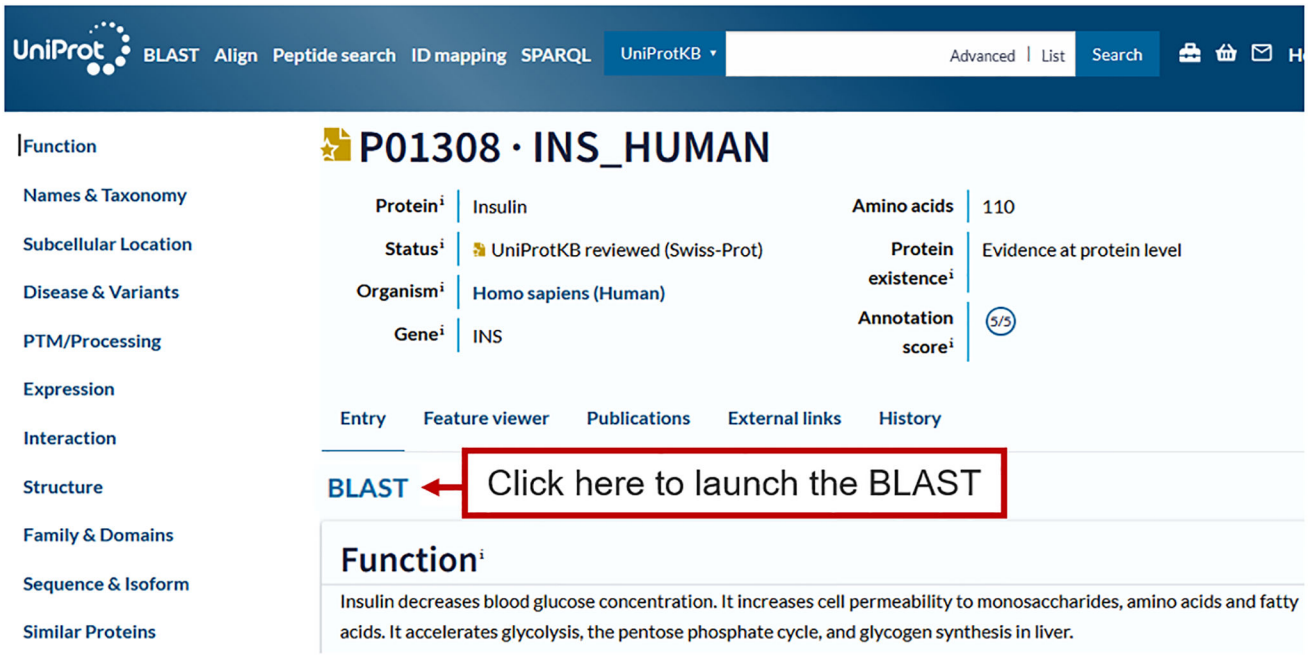

Figure 7.
Running BLAST on a protein from the protein entry page.

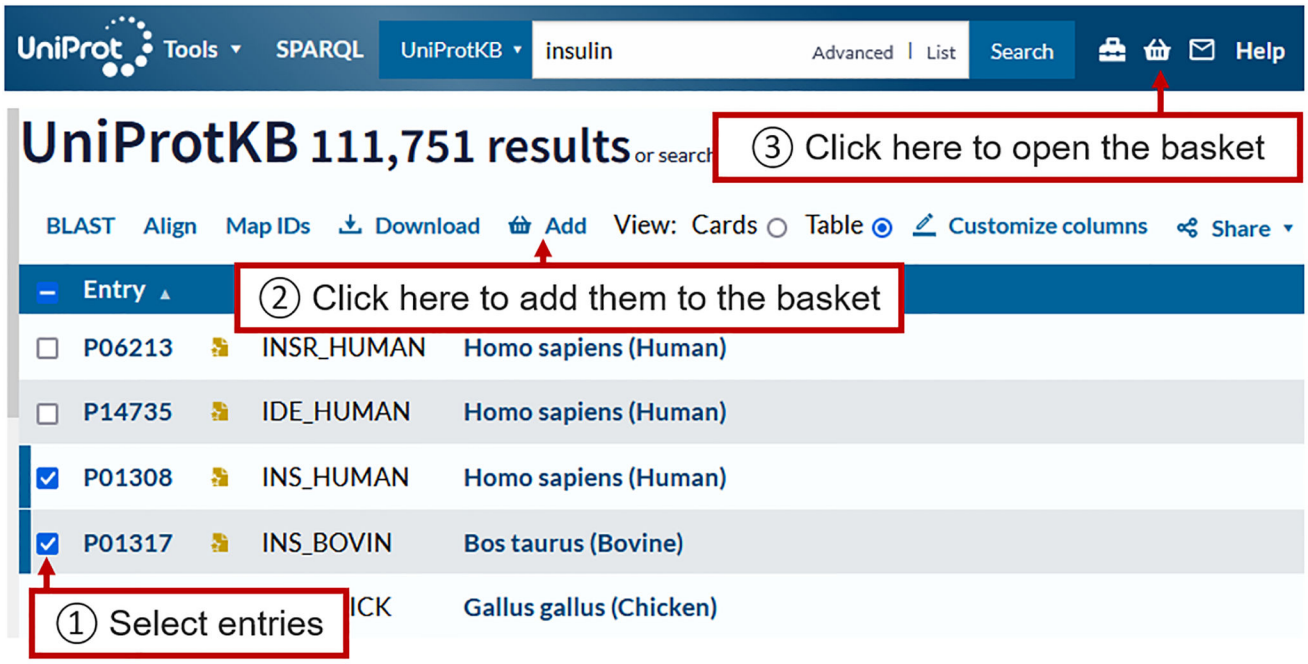

Figure 8.
Adding protein entries to the basket.

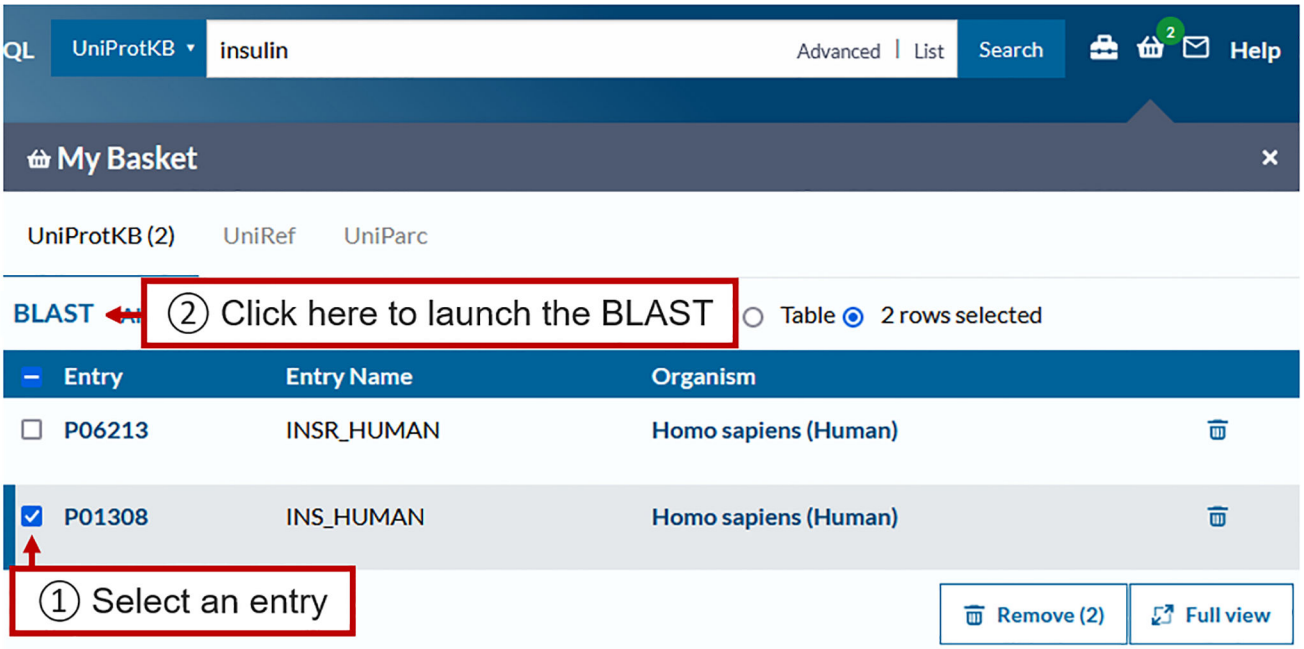

Figure 9.
Running BLAST from the basket.

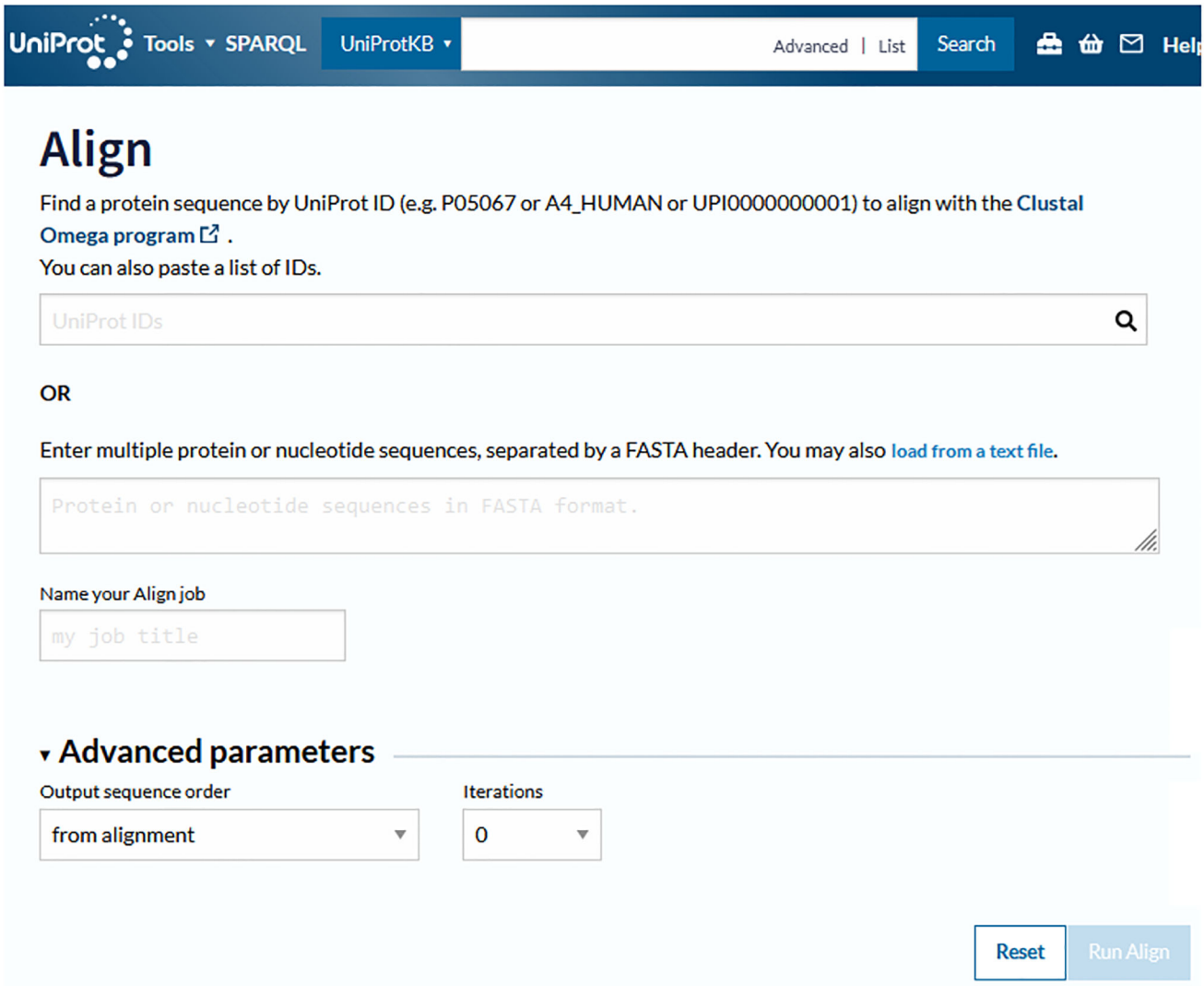

Figure 10.
Align query input page.

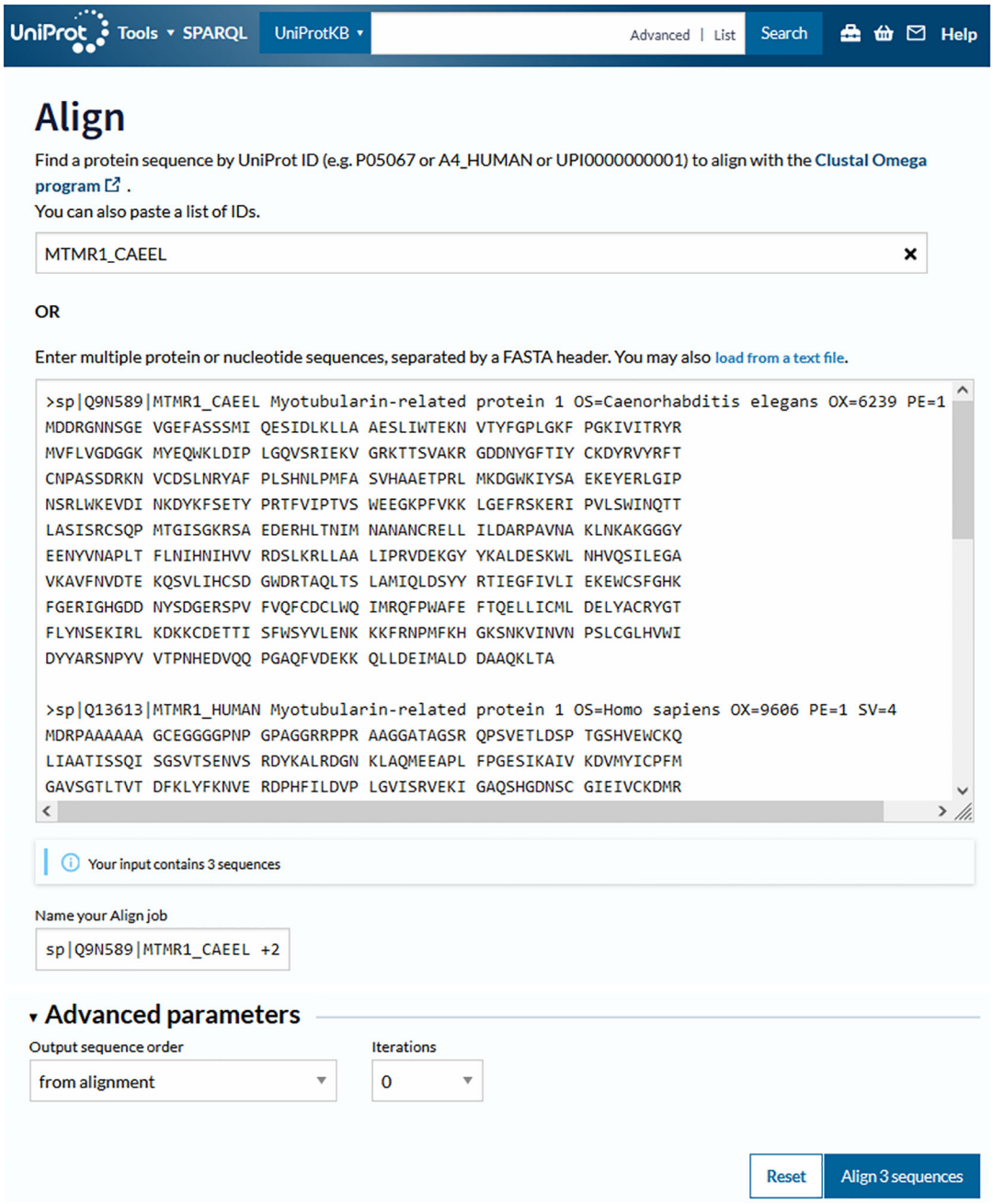

Figure 11.
Align query input page with entries in FASTA format.

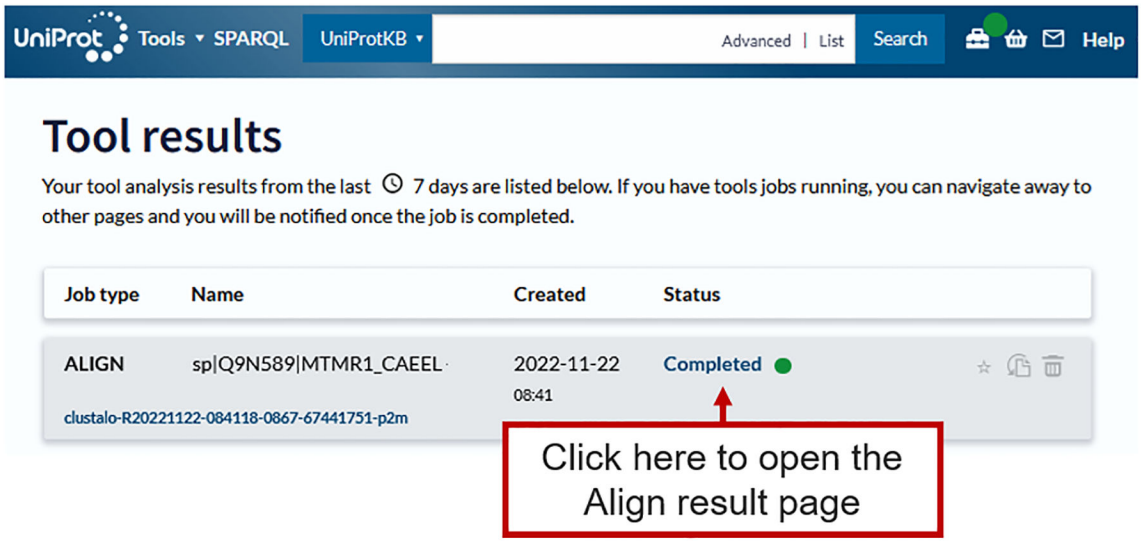

Figure 12.
Tool results dashboard page.

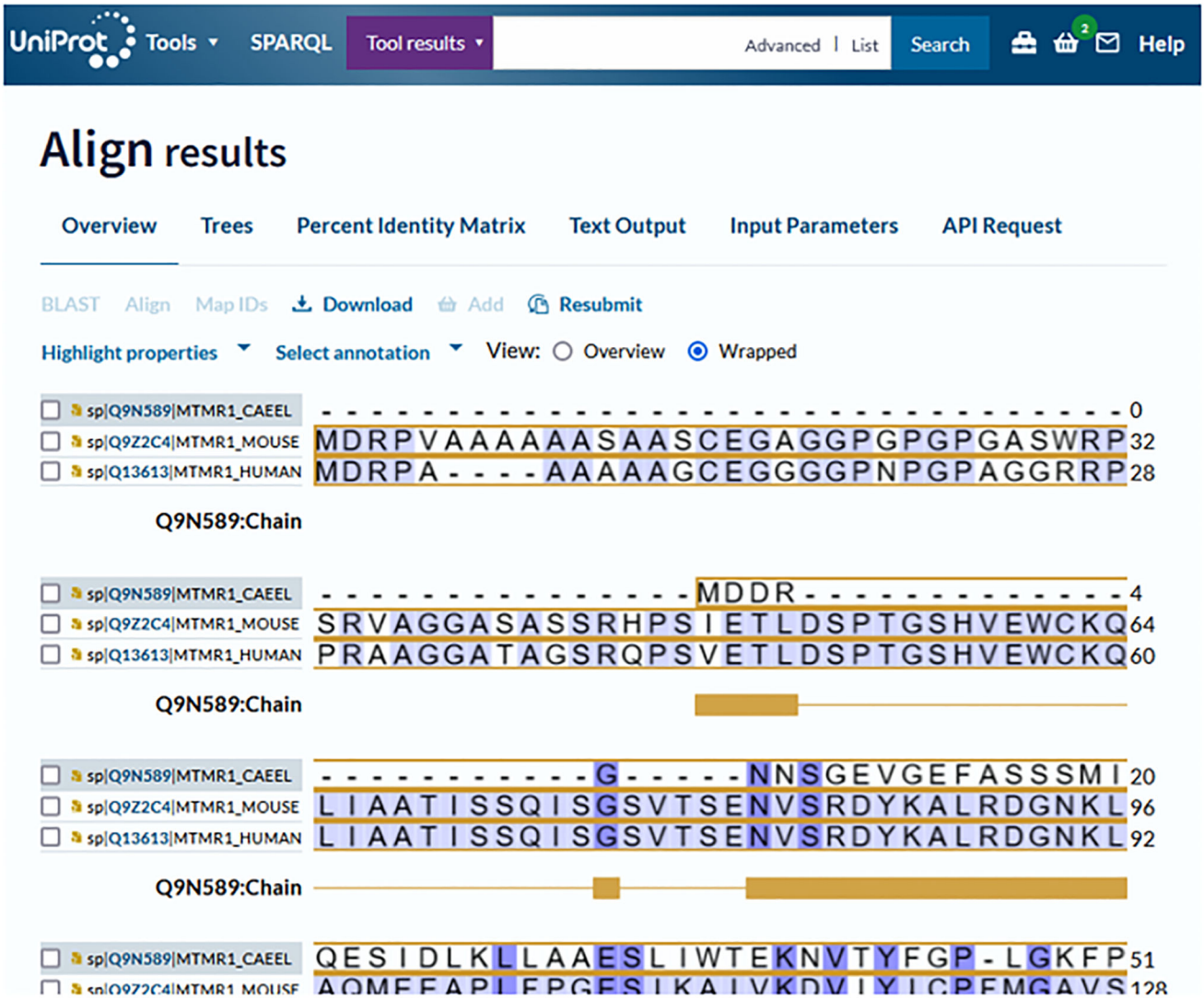

Figure 13.
Align results page.

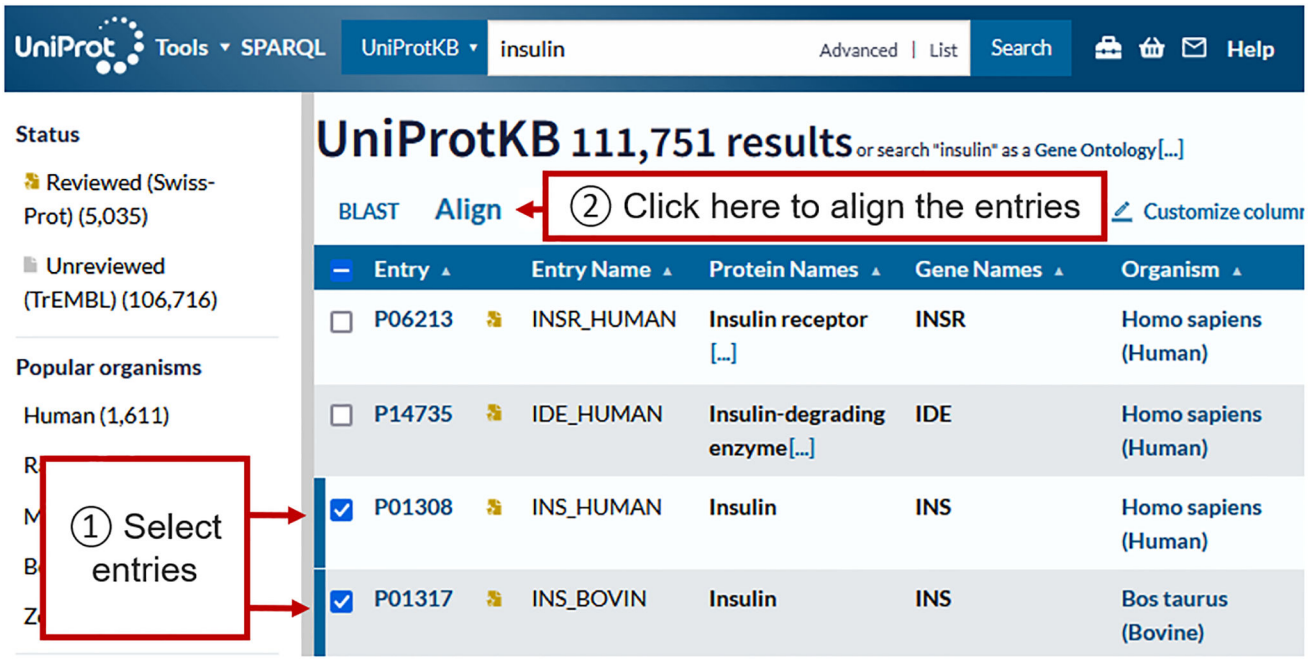

Figure 14.
Align multiple protein sequences from UniProtKB results page.

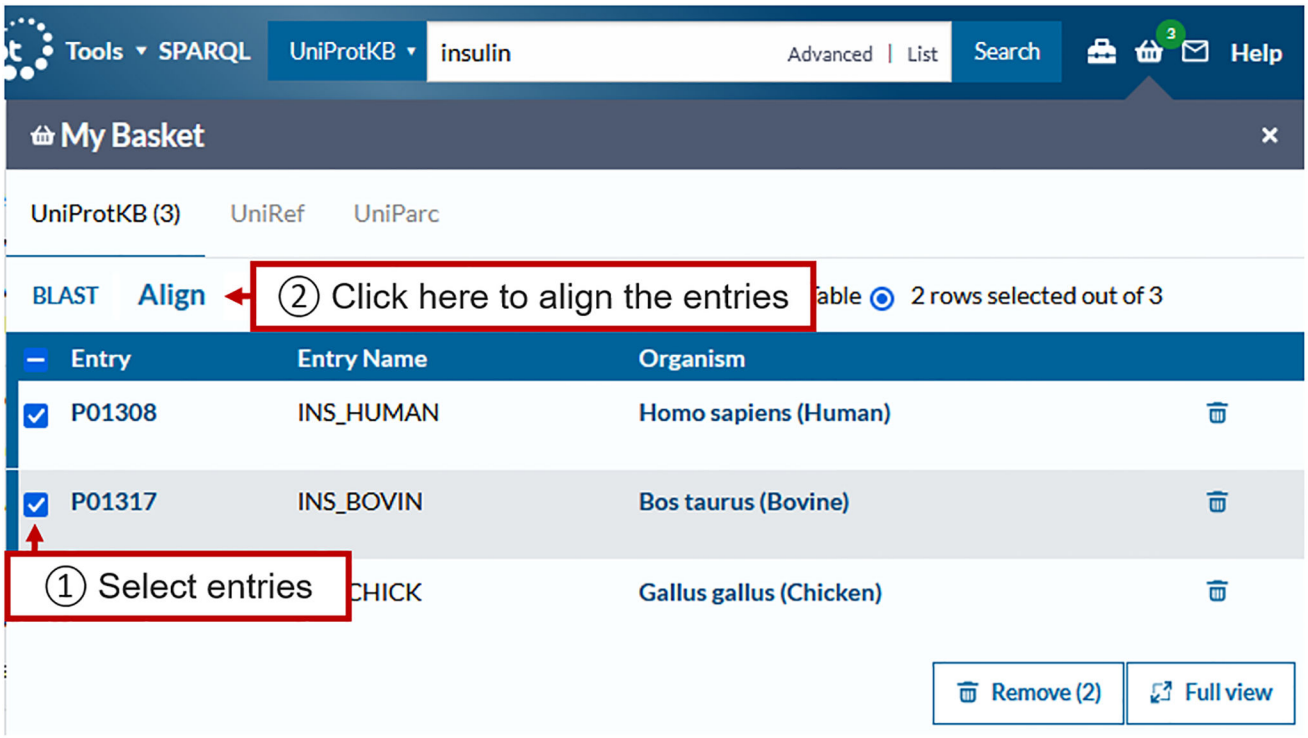

Figure 15.
Aligning proteins from the basket.

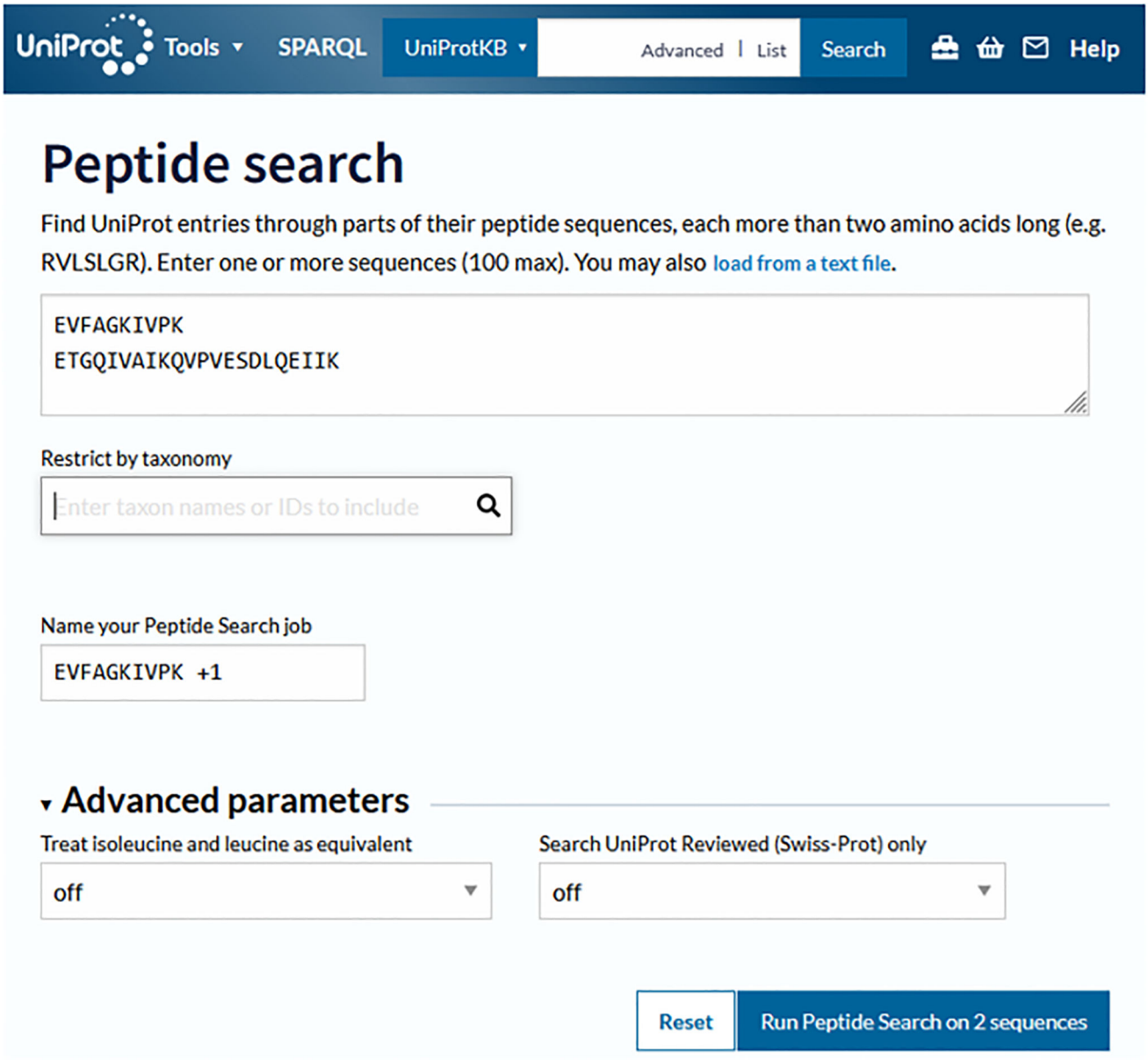

Figure 16.
Peptide search query input page.

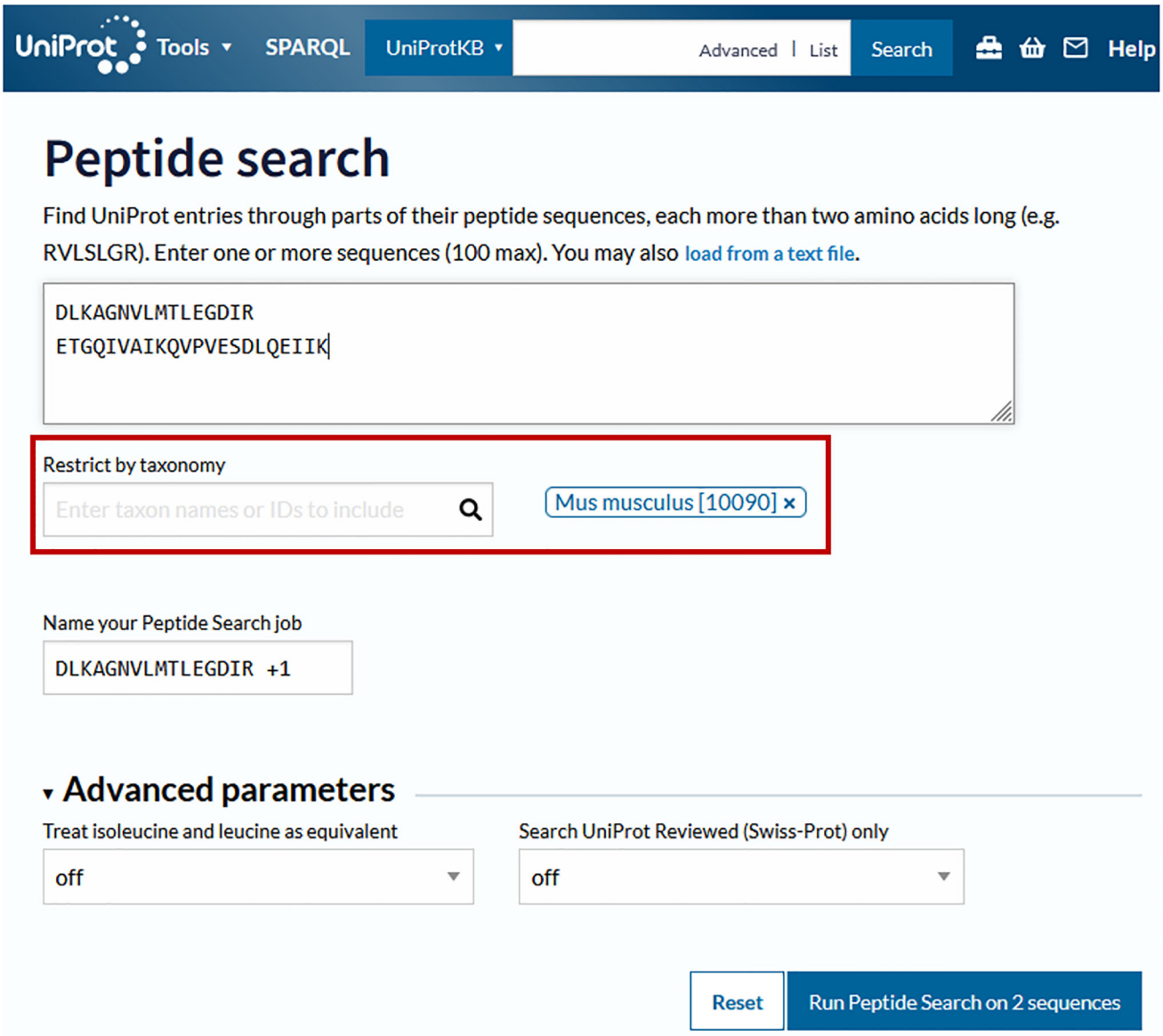

Figure 17.
Taxonomy restriction in the peptide search query input page.

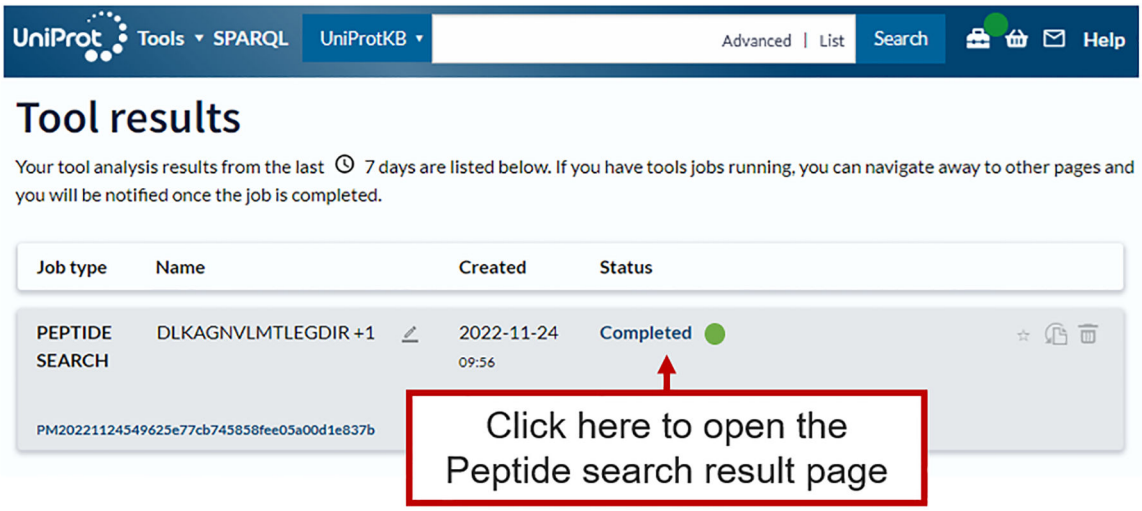

Figure 18.
Tool results dashboard page.

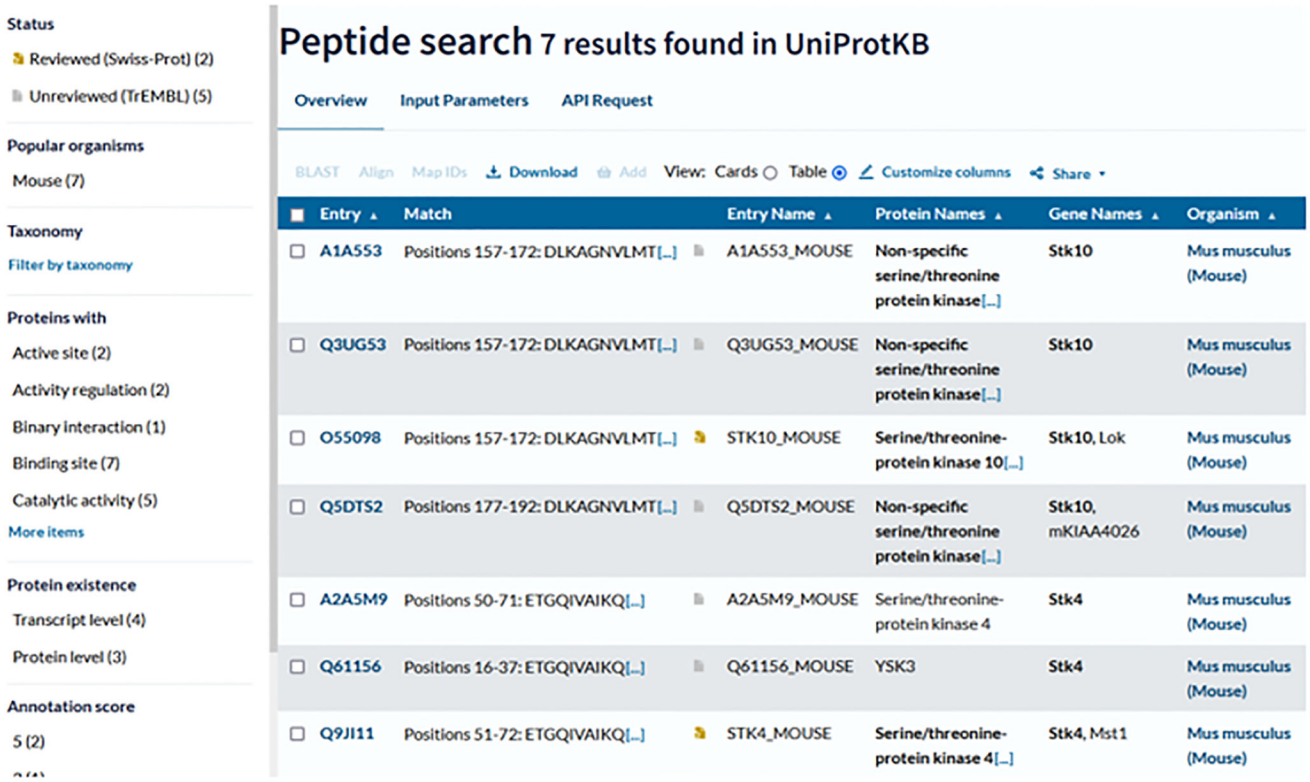

Figure 19.
Peptide search results page.

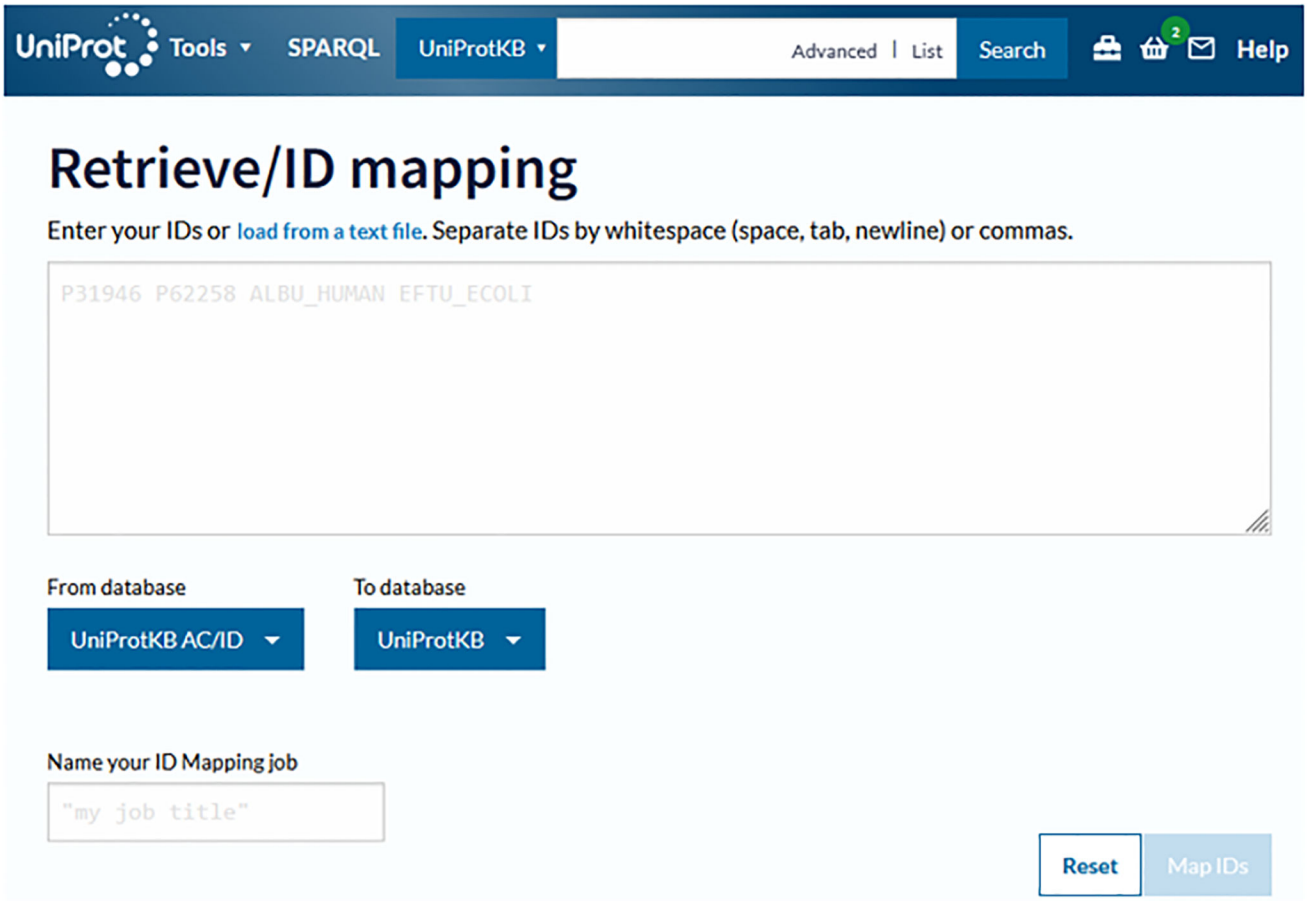

Figure 20.
Retrieve/ID Mapping query input page.

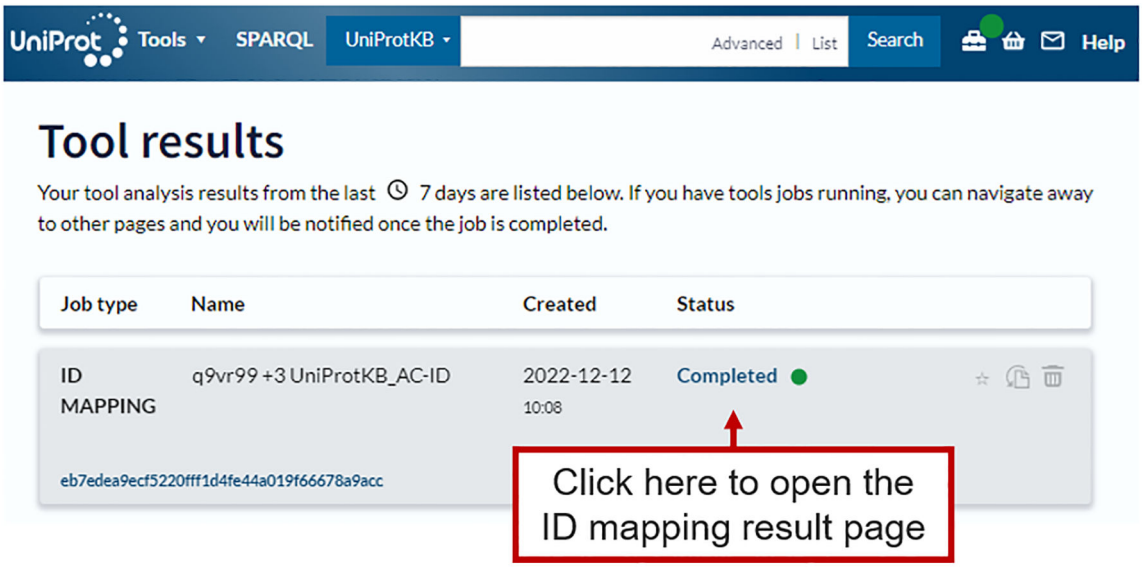

Figure 21.
Tool results dashboard page.

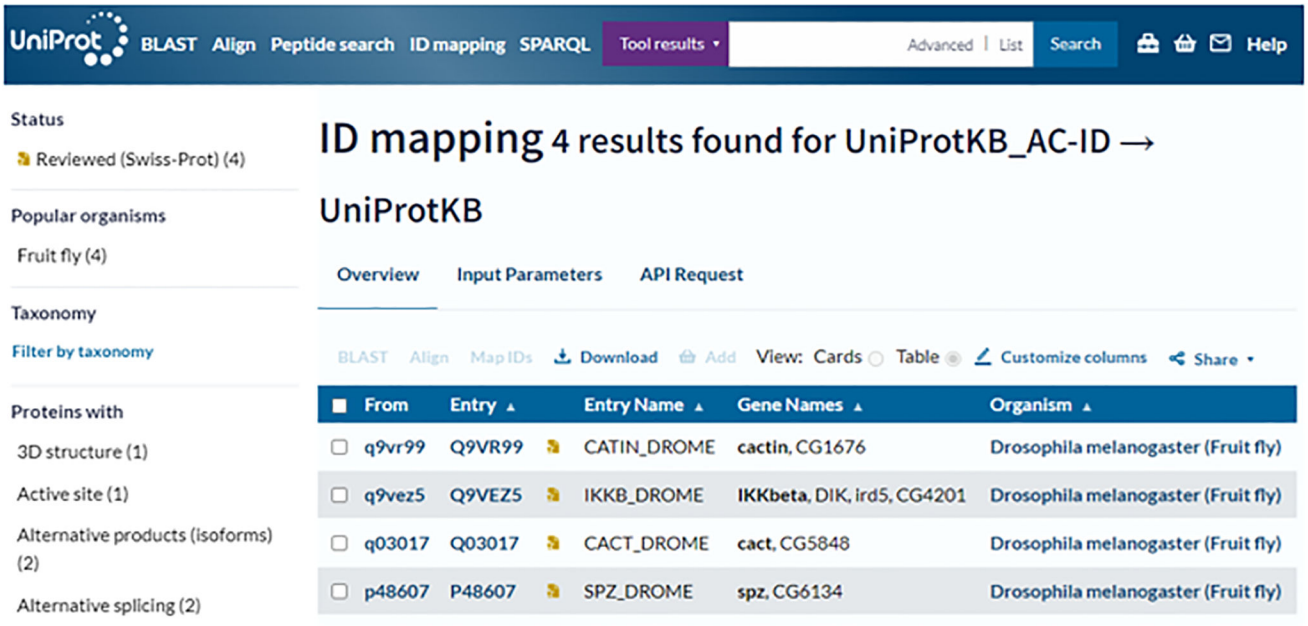

Figure 22.
UniProtKB ID mapping results page.

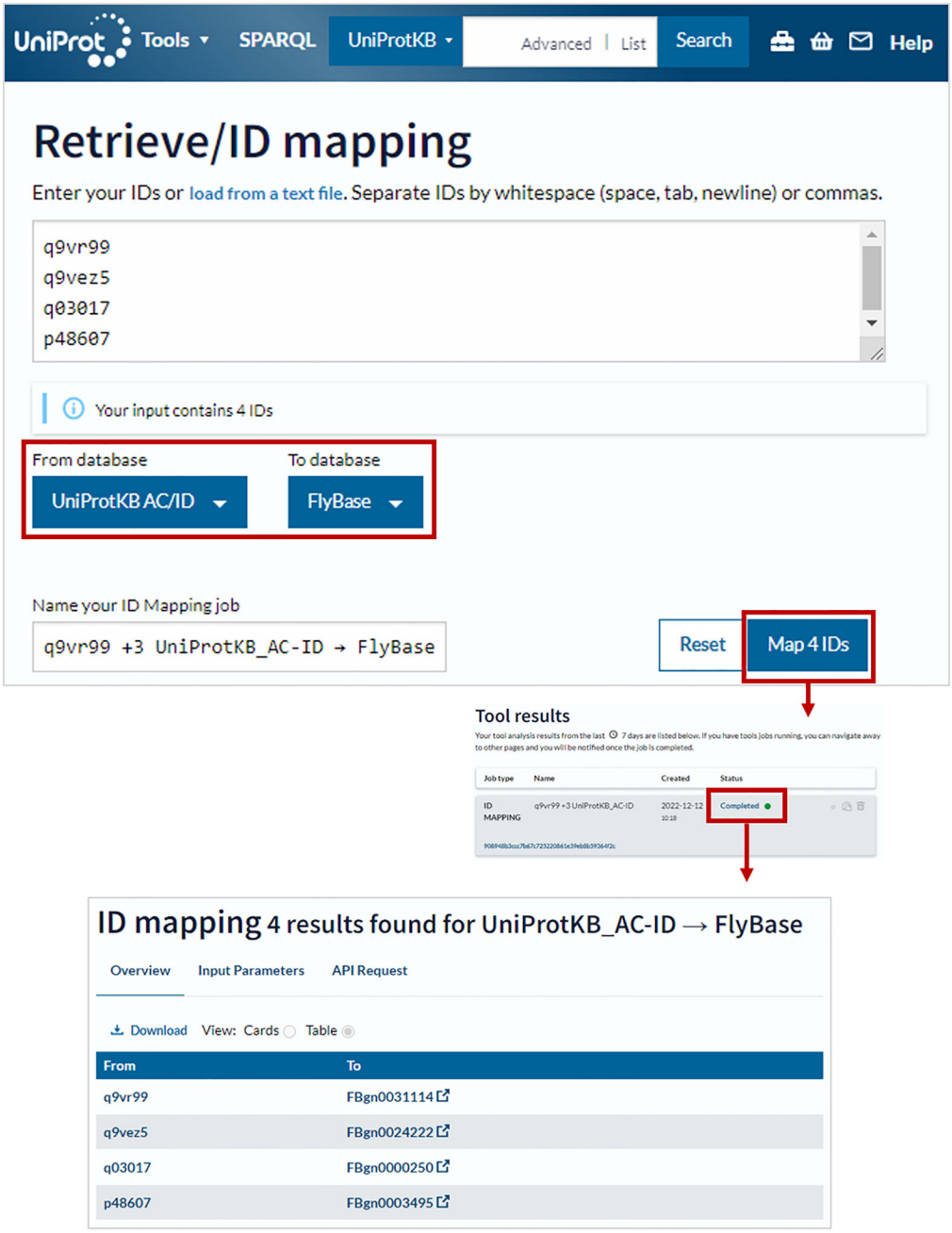

Figure 23.
Mapping UniProtKB IDs to an external database and ID mapping results page.

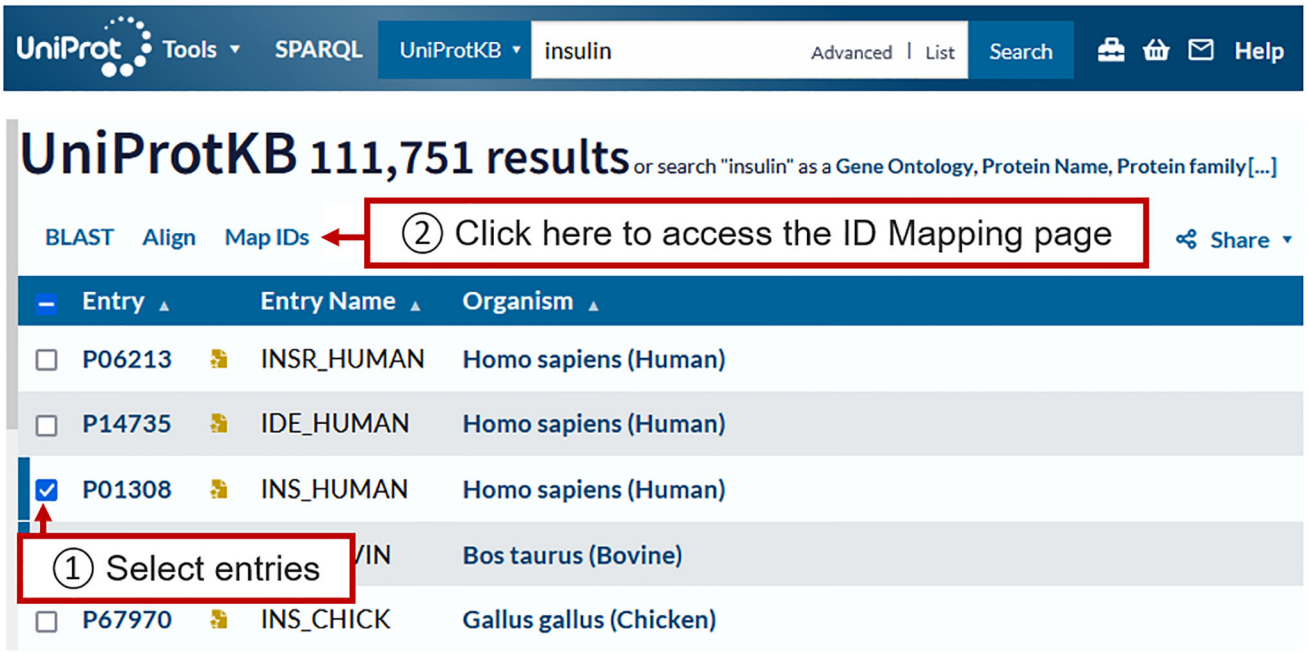

Figure 24.
Mapping UniProtKB IDs from the results page.

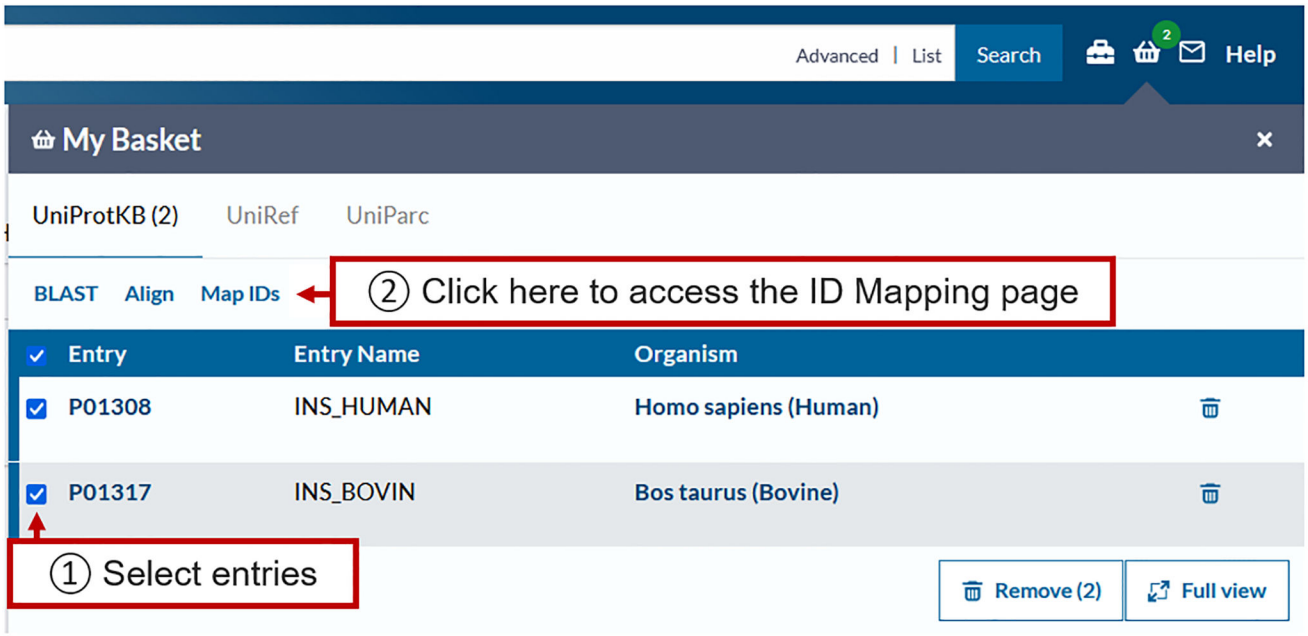

Figure 25.
Mapping UniProtKB IDs from the basket.

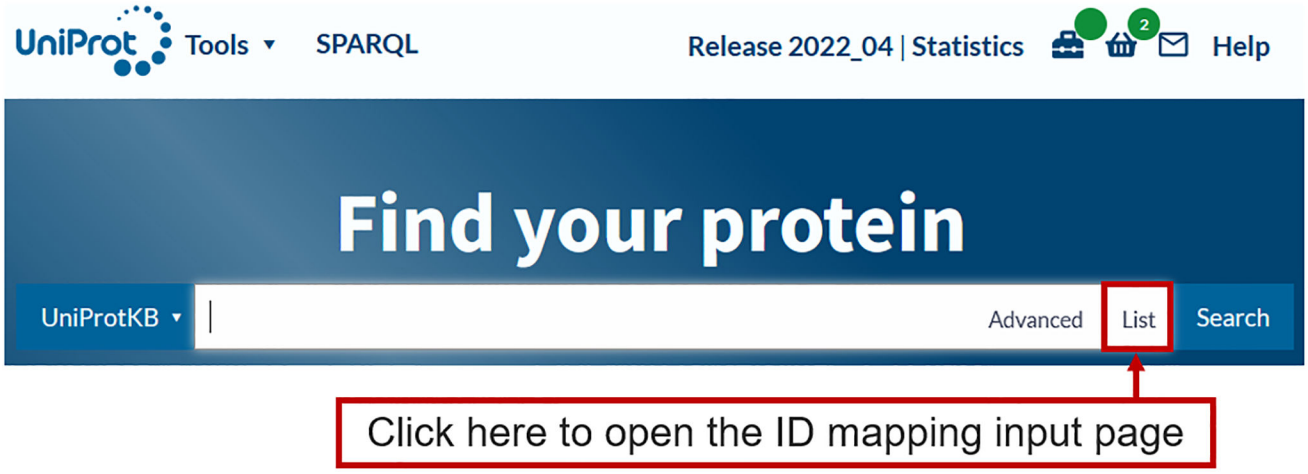

Figure 26.
Mapping UniProtKB IDs from the text search box.

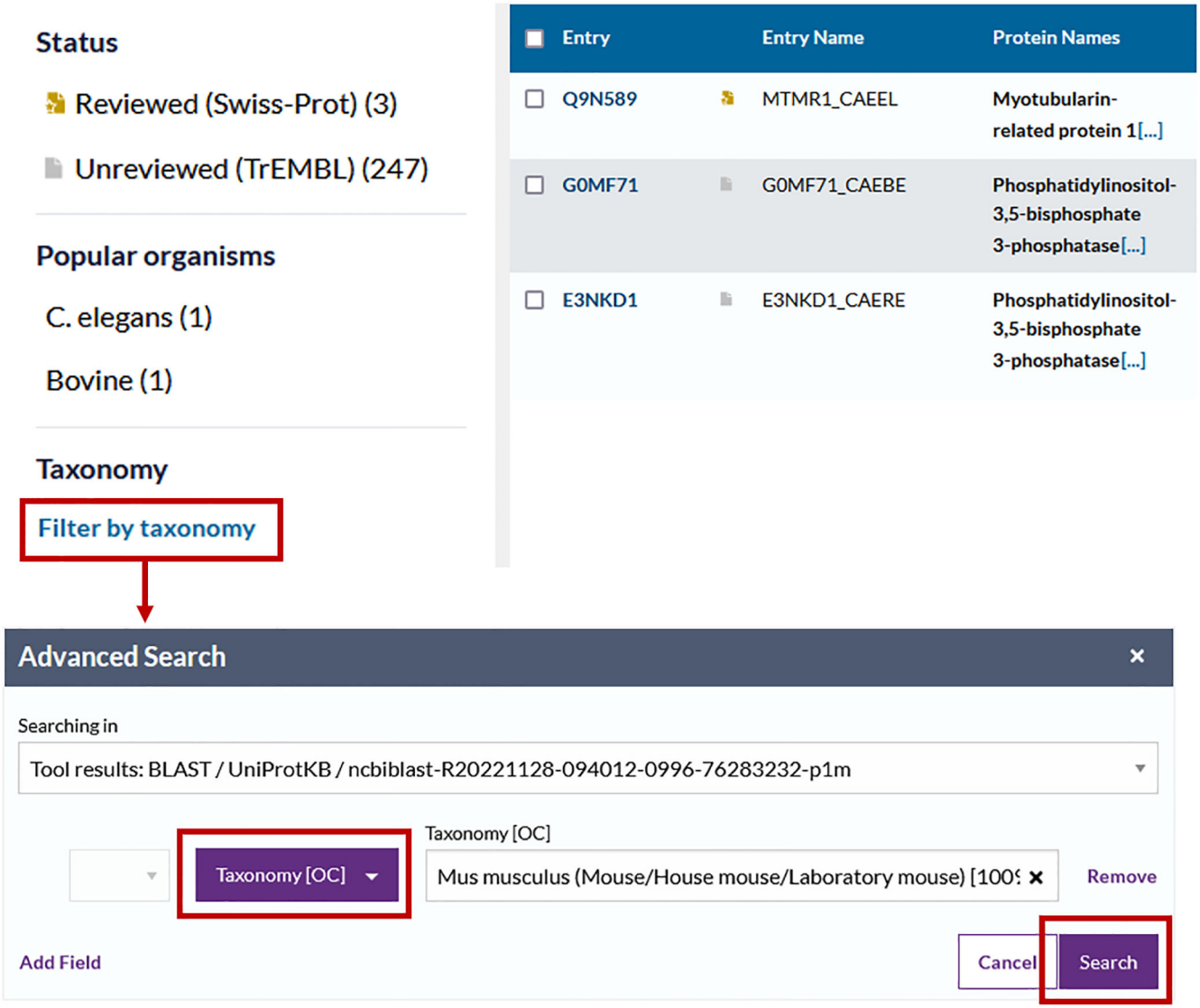

Figure 27.
Taxonomy filter in the BLAST results page.

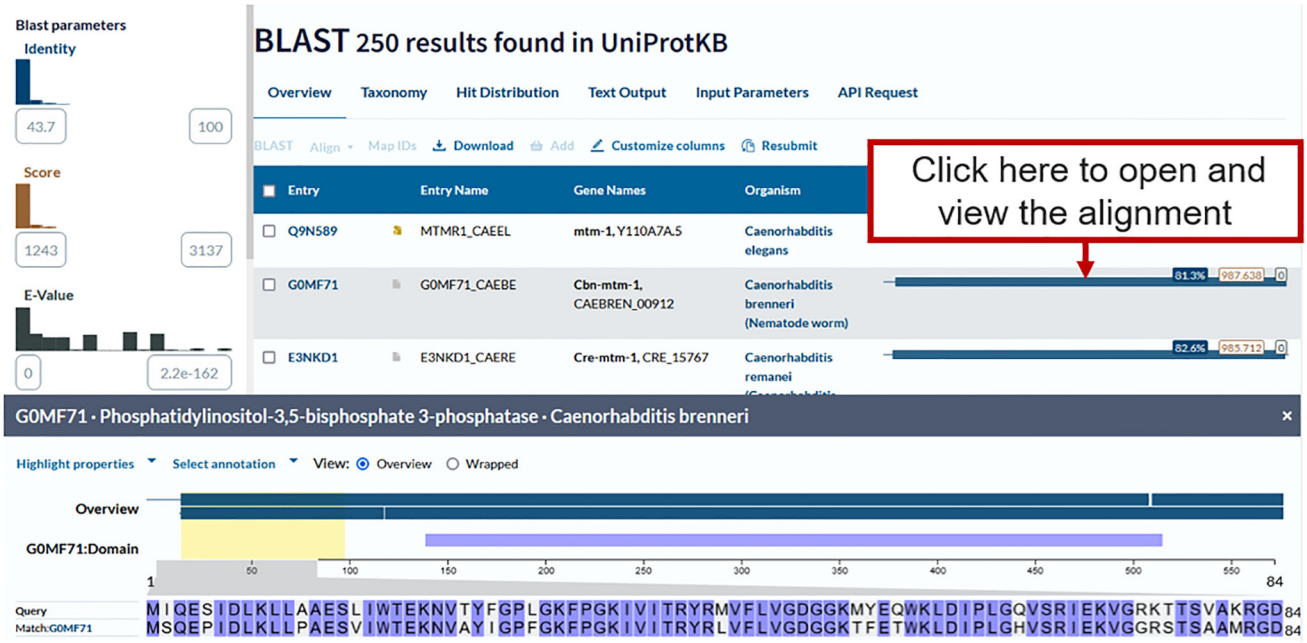

Figure 28.
Selecting and viewing the query/hit alignment in the BLAST results page.

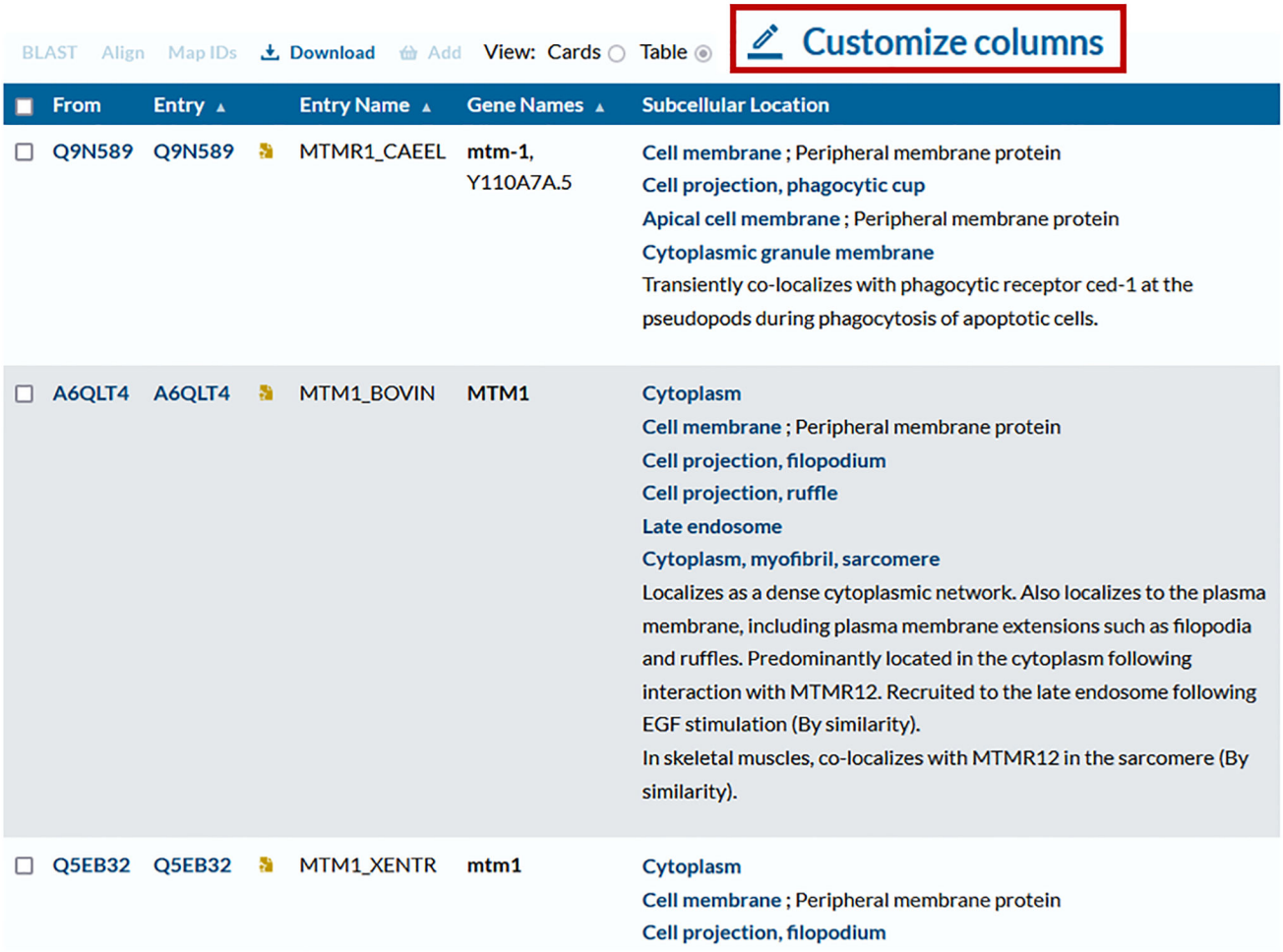

Figure 29.
Customizing columns of the BLAST results table.

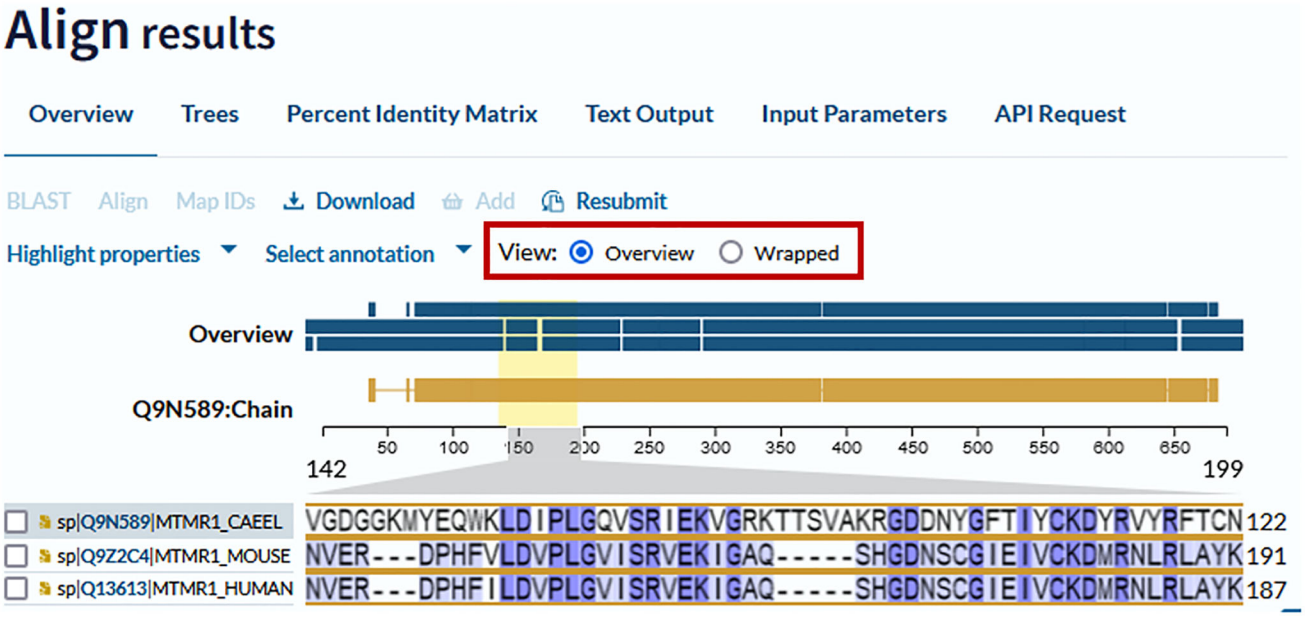

Figure 30.
Overview display of the Align results page.

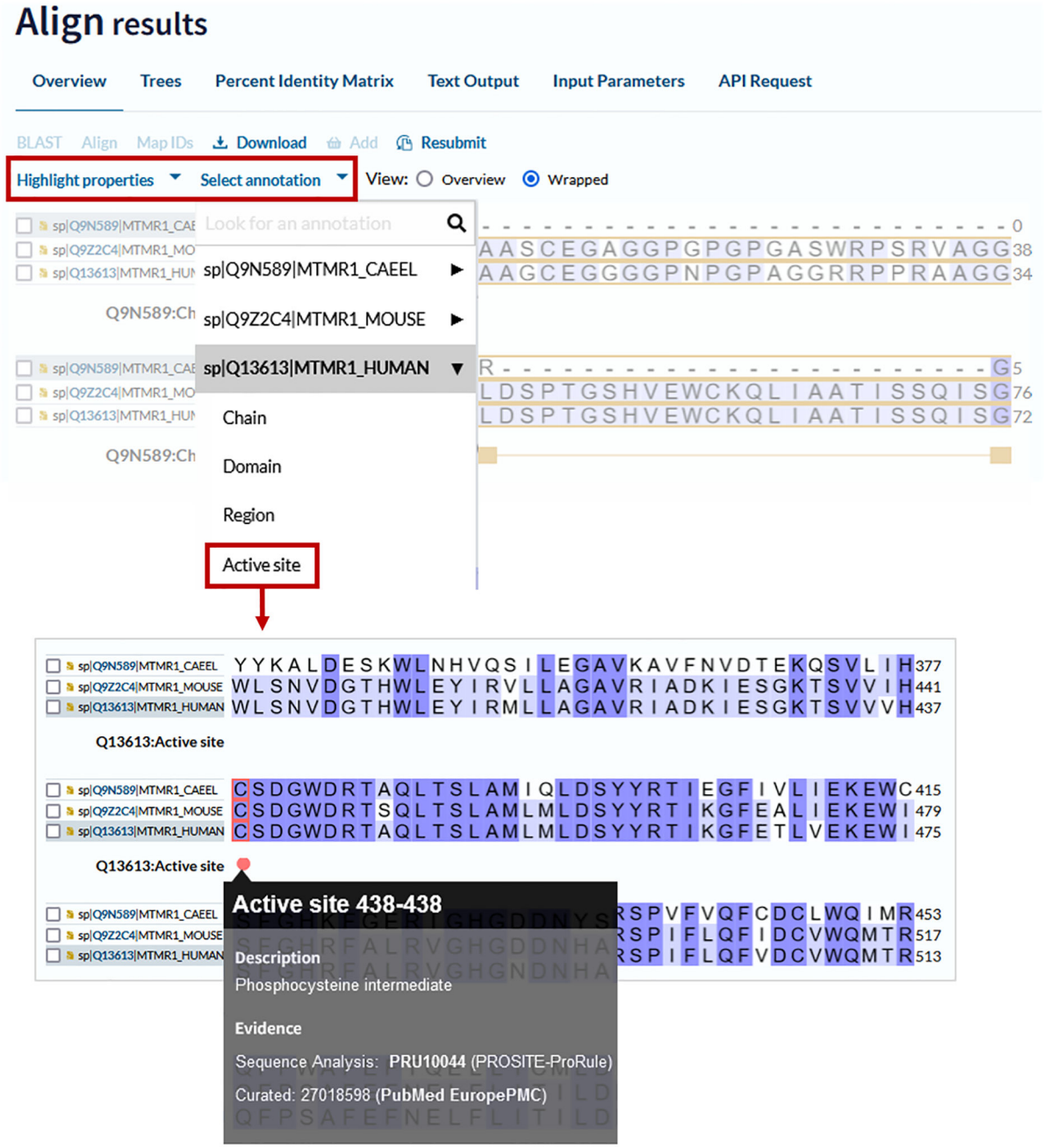

Figure 31.
Highlighting sequence features in the alignment.

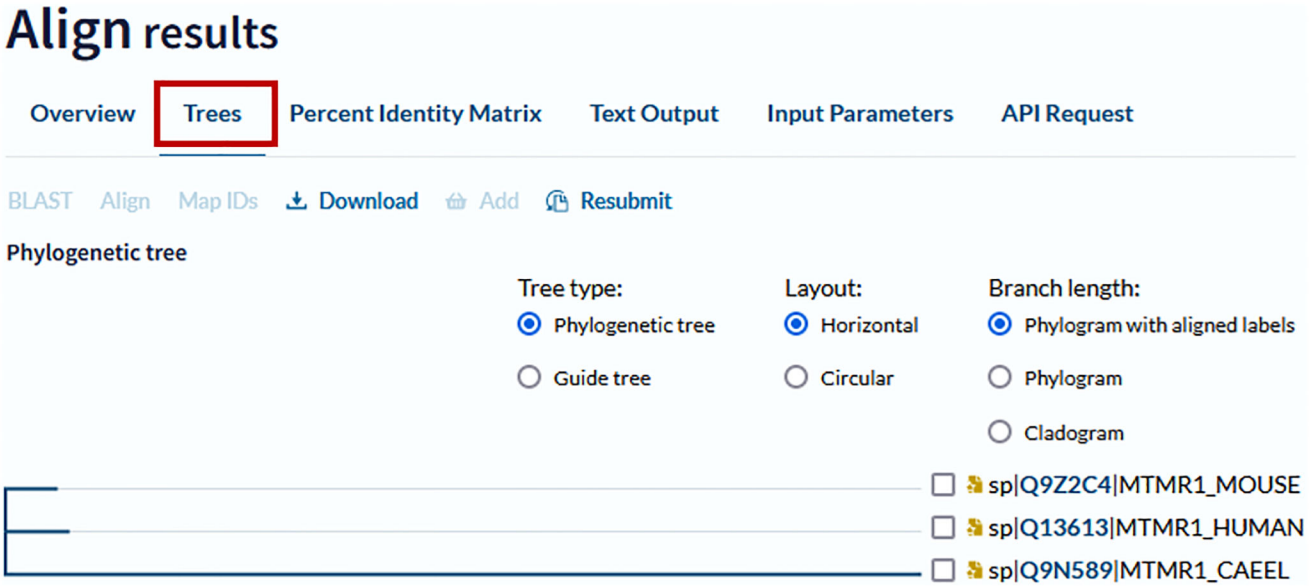

Figure 32.
Tree view of the alignment in the Align results page.

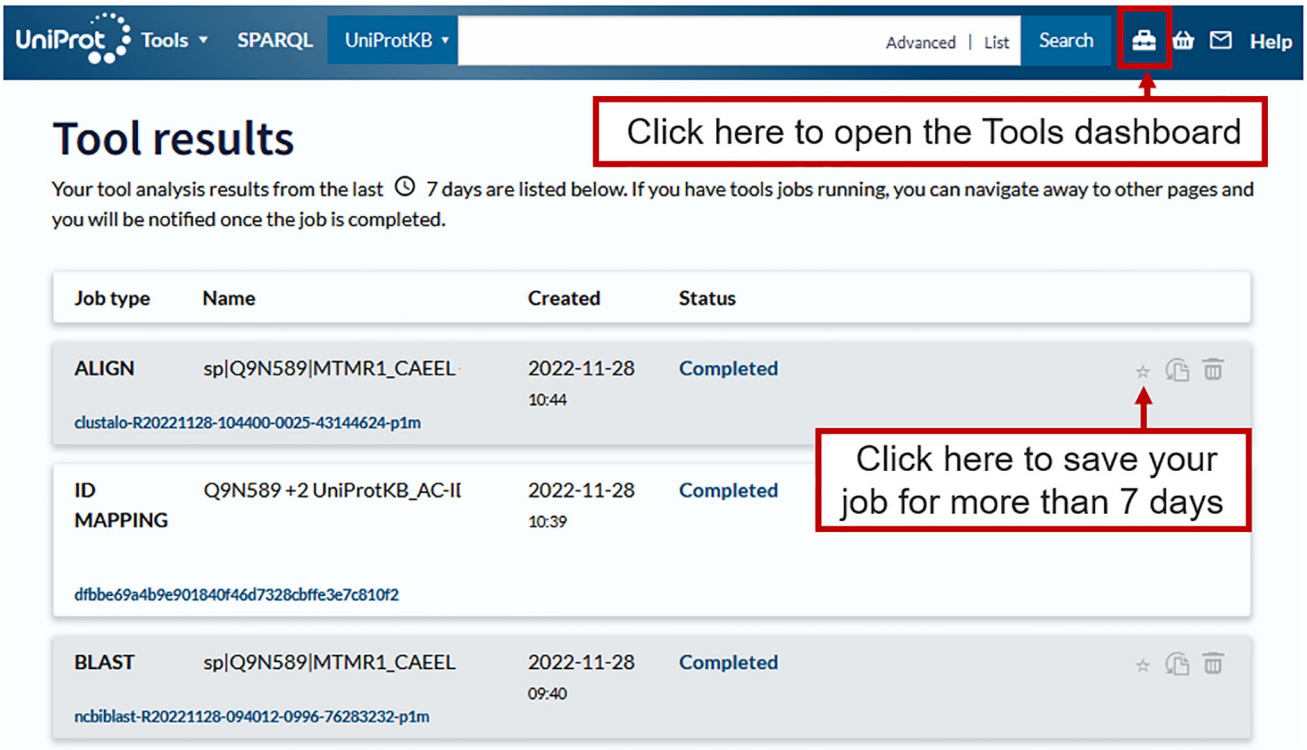

Figure 33.
Tools dashboard page.

Table 1

BLAST Advanced Options

ADVANCED OPTION	SIGNIFICANCE
Database	Database against which the search is performed; UniProtKB or clusters of sequences with 100%, 90%, or 50% identity.
Restrict by taxonomy	Select a specific taxon against which the search is performed
E-Threshold	The expectation value (E) threshold is a statistical measure of the number of expected matches in a random database. The lower the E-value, the more likely the match is to be significant. E-values between 0.1 and 10 are generally dubious and over 10 are unlikely to have biological significance. In all cases, those matches need to be verified manually. You may need to increase the E-value threshold if you have a very short query sequence, to detect very weak similarities or similarities in a short region, or if your sequence has a low complexity region and you use the 'filter' option.
Matrix	The matrix assigns a score for each position in an alignment. The BLOSUM matrix assigns a score based on the frequency with which that substitution is known to occur among consensus blocks within related proteins. BLOSUM62 is among the best of the available matrices for detecting weak protein similarities. The PAM set of matrices is also available. If 'Auto' is set the matrix will be selected depending on the query sequence length.
Filtering	Low-complexity regions (e.g., stretches of cysteine in Q03751 or, in some cases, hydrophobic regions in membrane proteins) tend to produce spurious, insignificant matches with sequences in the database which have the same kind of low-complexity regions but are unrelated biologically. If 'Filter low complexity regions' is selected, the query sequence will be run through the program SEG, and all amino acids in low-complexity regions will be replaced by an X.
Gapped	This will allow gaps to be introduced in the sequences when the comparison is done.
Hits	Limits the number of returned alignments. The user can limit the number of returned alignments to 50, 100, 250, 500, 750 or 1000. By default, the limit is set to 250.
HSPs per hit	Limits the number of high-scoring Segment Pair (HSP) which are local alignment with no gaps that achieves one of the highest alignment scores in a given search.