HHS Public Access

Author manuscript

Curr Protoc Bioinformatics. Author manuscript; available in PMC 2017 March 24.

Published in final edited form as:

Curr Protoc Bioinformatics. ; 53: 1.29.1–1.29.15. doi:10.1002/0471250953.bi0129s53.

UniProt Tools

Sangya Pundir[1,*], **Maria J. Martin**[1], **Claire O'Donovan**[1], and **The UniProt Consortium**[1,2,3,4]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK [2]SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland [3]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven Street North West, Suite 1200, Washington, DC 20007, USA [4]Protein Information Resource, University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA

Abstract

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data (UniProt C., 2015). The UniProt website receives approximately 400,000 unique visitors per month and is the primary means to access UniProt. Along with various datasets that you can search, UniProt provides three main tools. These are the 'BLAST' tool for sequence similarity searching, the 'Align' tool for multiple sequence alignment 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 three basic protocols, three alternate protocols and two supporting protocols for using UniProt tools.

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 C., 2015). To build upon this protein data and to aid analysis, UniProt provides three main tools; 'BLAST' (Basic Local Alignment Search Tool), 'Align' multiple sequence alignment tool 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 website and are also accessible directly from other parts of the website like the basket, search/ tool results pages and protein entry pages. Having these tools in the UniProt website 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

* spundir@ebi.ac.uk.

filter by attributes like whether you're looking for Reviewed (Swiss-Prot) or Unreviewed (TrEMBL) UniProt Knowledgebase entries or entries with 3D structures or entries that are part of a proteome, etc.

The UniProt website can be accessed at the URL http://www.uniprot.org/. The following three basic 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 website provides a Basic Local Alignment Search Tool (UniProt C., 2015) 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 website.

Necessary Software

An up-to-date web browser and computer

Basic Local Alignment Search Tool in UniProt—

1. Go the UniProt home page at http://www.uniprot.org/.
2. Click on the 'BLAST' link, which is always on the left hand side of the header bar on all UniProt pages as shown in Figure 1.
3. You will see the BLAST input page as shown in Figure 2.
4. To run a BLAST search, enter a protein sequence or nucleotide sequence or UniProt identifier into the sequence input box provided, for example enter 'A4_Human'. This is the only mandatory input field.
5. You can choose to change parameters from a number of optional advanced options, as shown in Table 1.
6. Click on the 'Run BLAST' button to execute your query. You will see a 'Job status: running' page while you wait for results and will then see your BLAST results page as shown in Figure 3.

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 analyze using a sequence similarity search. This allows a flexible workflow between browsing data and analyzing it.

Necessary Software

An up-to-date web browser and computer

Basic Local Alignment Search Tool through UniProt text search results and entry pages—

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 as one of ‘UniProtKB’, ‘UniRef’ and ‘UniParc’ to follow the steps below.
3. Enter a query in the search box, for example ‘insulin’, and click on the search button.
4. You will see a search results page as shown in Figure 4.
5. To run a BLAST search for a protein in your search results, click on the checkbox in the left hand column for that protein row.
6. Now click on the BLAST button just above the search results table and click ‘OK’ in the dialog box that appears, as shown in Figure 5.
7. If you click on a UniProt entry in your results table, you will be taken to the entry page, which also provides a BLAST button for direct submission, as shown in Figure 6.

Supporting 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 UniProt Knowledgebase, 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 data set to access analysis tools, i.e. BLAST, Align and Retrieve/ID Mapping. Your basket is saved as long as you don’t clear your cookies and stores up to 400 entries.

Necessary Software

An up-to-date web browser and computer

Accessing Basic Local Alignment Search Tool through the UniProt basket—

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 as one of ‘UniProtKB’, ‘UniRef’ and ‘UniParc’ to follow the steps below.
3. Enter a query in the search box, for example ‘insulin’, and click on the search button.
4. You will see a search results page as shown in Figure 4. For you entry of interest, click on the checkbox to the left of its accession and then click on the ‘Add to Basket’ button on top of the results table.

5. When you are ready to analyze the entries in your basket, click on the basket to open it.

6. Your entries will be under their dataset tab (UniProtKB/ UniRef/ UniParc). Click on the checkbox to the left of your entry of interest and then click on ‘BLAST’, as shown in Figure 7.

BASIC PROTOCOL 2: Multiple sequence alignment in UniProt

The UniProt website 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 Software

An up-to-date web browser and computer

‘Align’ multiple sequence alignment tool in UniProt—

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.

3. You will see the Align input page as shown in Figure 8.

4. To execute the multiple sequence alignment, enter the protein sequences in FASTA format or UniProt identifiers into the sequence input box provided and click ‘Run align’, for example enter ‘A4_Human’ ‘A4_Mouse and ‘A4_Rat’ separated by a space or each on a new line.

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 Software

An up-to-date web browser and computer

Align Tool through UniProt results pages—

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 as ‘UniProtKB’ and follow the steps below.

3. Enter a query in the search box, for example ‘insulin’, and click on the search button.

4. You will see a results page as shown in Figure 4.

5. Click on two or more checkboxes to align these protein entries, as shown in Figure 9.
6. Now click on the Align button just above the search results table.
7. You will see the ‘Job status: Running’ page while you wait for your results and then see your alignment results page as shown in Figure 10.

Supporting PROTOCOL 2: Align through UniProt basket

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

Necessary Software

An up-to-date web browser and computer

Accessing the Align tool through the UniProt basket—

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 as ‘UniProtKB’ and follow the steps below.
3. Enter a query in the search box, for example ‘insulin’, and click on the search button.
4. You will see a results page as shown in Figure 4.
5. 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 to Basket’ on top of the results table. You can store UniProt entries in a basket over multiple search sessions and then align them later.
6. When you are ready to analyze the entries in your basket, click on the basket to open it, as shown in Figure 11.
7. Click on the checkboxes to the left of the entries you would like to align and then click on ‘Align’. You need to select two or more entries to be able to create a multiple sequence alignment.
8. You will see the ‘Job status: Running’ page while you wait for your results and then see your alignment results page as shown in Figure 10.

BASIC PROTOCOL 3: Batch Retrieval and ID Mapping in UniProt

The UniProt website 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 UniProt to over 100 external databases that UniProt is cross-referenced to and vice versa (e.g. Ensembl, PDB, Refseq, etc.) (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, Genome annotations and others. This tool is called ‘Retrieve/ID Mapping’.

Necessary Software

An up-to-date web browser and computer

Retrieve/ID Mapping tool webpage in UniProt—

1. Go the UniProt home page at http://www.uniprot.org/.
2. Click on the ‘Retrieve/ID Mapping’ link, which is available in the header bar on all UniProt pages as shown in Figure 1.
3. You will see the ‘Retrieve/ID Mapping’ input page as shown in Figure 12.
4. If you have a list of UniProt IDs and would like to retrieve results for all of these, paste your list into the input box provided or upload a file.
5. Leave ‘From’ and ‘To’ as ‘UniProtKB’ and click the ‘Go’ button. You will get a UniProtKB batch results page, as shown in Figure 13.
6. If you have a list of UniProt IDs and would like to map them to IDs from an external database (or vice versa), upload or paste in your IDs and then select the source database in the ‘From’ dropdown and the target database in the ‘To’ dropdown and then click the ‘Go’ button. 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 14.

ALTERNATE PROTOCOL 3: ‘Retrieve/ID Mapping’ Tool through UniProt Basket

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

1. Choose the search dataset through the dropdown to the left of the search box as ‘UniProtKB’ and follow the steps below.
2. Enter a query in the search box, for example ‘insulin’, and click on the search button.
3. You will see a results page as shown in Figure 4.
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 to Basket’ on top of the results table.
5. When you are ready to analyze the entries in your basket, click on the basket to open it.
6. You can click directly on the ‘Download’ button in the basket to retrieve the UniProt entries in your basket. To map the UniProt IDs to an external database from the basket, click on the checkboxes to the left and click on ‘Map IDs’ button in the basket, as shown in Figure 11.
7. Your selected IDs will appear in the query box on the ‘Retrieve/ID Mapping’ page.

8. To map them to IDs from an external database, select the target database in the ‘To’ dropdown and then click the ‘Go’ button. 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 14.

GUIDELINES FOR UNDERSTANDING RESULTS

UniProt BLAST Results

The UniProt BLAST results appear as shown in Figure 3. The results page provides (1) a left hand panel with filters and options to view the results arranged by taxonomy or view the raw alignments, (2) an overview graphical list on top of the page and (3) a detailed table of alignments underneath that.

The left hand panel allows you to filter for Reviewed (UniProtKB/Swiss-Prot), Unreviewed (UniProtKB/TrEMBL) entries, entries with 3D structures or entries from a specific organism. It also allows you to view results by taxonomy, in text and XML formats. The ‘view by Taxonomy’ page provides results on a taxonomy tree and also includes an input box (with auto-completion) to allow you to jump to any taxonomy node, as shown in Figure 15. You can click on the taxonomy node itself to go to a page describing this taxonomy node or click on the number of results shown in brackets to view the corresponding BLAST results.

The overview table presents color coded results that are ordered by identity by default. Each row is colored using a heat map scale with red reflecting the closest identity to the query sequence and blue reflecting the least identity. The order can be changed to view the diagram sorted by E Value or Bit Score.

The detailed alignment table shows the alignment graphically with the query shown in black and the subject hit color coded to reflect identity following the same scale as the overview table. You can hover over the diagram to see sequence lengths or click on the diagram to view the sequences aligned in detail. You can select entries in your alignment table by clicking on the checkboxes to their left and then submit them to the Blast tool, the align tool or add them to your basket. You can also download your results in various formats by clicking on the ‘Download’ button.

Your BLAST job is assigned a job identifier that can be seen at the bottom of your BLAST results page in a section titled ‘Job information’. This identifier can be used to access your BLAST page at any time up to seven days from when you first ran your query. A UniProt YouTube video explaining how to use the page is available at https://www.youtube.com/watch?v=UPaConHNP7E.

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 10. The main page shows the sequences aligned and a left hand panel provides highlight options. These options allow you to select sequence annotations like domains, sites, etc. to view them highlighted across the sequences aligned. You can also highlight by amino acid properties like hydrophobicity etc. as shown in Figure 16. You can download your alignment from this page using the 'Download' button.

Scrolling further down the page, you can see a section called 'Results information' which provides your job identifier, number of identical positions, similar positions etc. The job identifier can be used to access your sequence alignment at any time for up to seven days from when you first ran the query.

A UniProt YouTube video explaining how to use the page is available at https://www.youtube.com/watch?v=IAYFLfPQ0Gs.

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 13. This results page provides filters and view options in the left hand panel and a main results table covering the majority of the page. The results table starts with a column titled 'Your list' that shown 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 'Columns' button on top of the results table. You can also run tools like 'Blast', 'Align' and add entries to your basket by selecting the corresponding checkboxes and then clicking on the buttons available on 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 just two columns showing your input IDs and the corresponding mapped IDs, as shown in Figure 14. You can use the 'Download' button to download your results.

Your Retrieve/ID Mapping job is assigned a job identifier that can be seen in the search box in the header on your results page. This identifier can be used to access your job results at any time for up to seven days from when you first ran your query. A UniProt YouTube video explaining how to use the page is available at https://www.youtube.com/watch?v=kLdgjqWoMZc.

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 analyse this data. The tools UniProt provides are BLAST, Align and Retrieve/ID Mapping. The UniProt website 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 like searching for data, then running a BLAST search for a sequence in your results then running a multiple sequence alignment on sequences in your BLAST results and then mapping the IDs of these sequences to an external database.

UniProt provides training material through the European Bioinformatics Institute (EMBL-EBI) 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 they are also available on our YouTube channel at https://www.youtube.com/uniprotvideos.

Critical Parameters

The BLAST tool page on the UniProt website offers advanced options to allow you to change a number of parameters, as shown in Table 1. One of these is the option to change the target database for running your search against, which is UniProtKB by default. If you are looking for sequence similarity matches from a particular taxonomy group, you can choose to restrict your target data to UniProtKB entries only from Bacteria, Eukaryota, mammals, plants, etc. You can also restrict your target database to only those UniProtKB entries from Reference proteomes or from the Reviewed (Swiss-Prot) set if you're looking for well-annotated results. If you'd like to speed up the BLAST search, you can choose to search against UniRef clusters, which reduce redundancy by grouping sequences based on identity. You can also choose to search against UniParc if you would like to search the full archive of sequences (including those that are no longer in the UniProtKB).

Other parameters that you can change are E-threshold, Matrix, Filtering, Gapped (yes or no) and number of hits. To better understand the effects of E-threshold, matrix and gapped search changes, refer to UNIT 3.4 (Ladunga, 2009).

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.

Literature cited

Huang, Hongzhan; McGarvey, Peter B.; Suzek, Baris E.; Mazumder, Raja; Zhang, Jian; Chen, Yongxing; Wu, Cathy H. A comprehensive protein-centric ID mapping service for molecular data integration. Bioinformatics. 2011 Apr 15; 27(8):1190–1191. 2011. [PubMed: 21478197]

Ladunga I. Finding Homologs in Amino Acid Sequences Using Network BLAST Searches. Current Protocols in Bioinformatics. 2009; 25:3.4, 3.4.1–3.4.34. [PubMed: 19496060]

Simossis V, Kleinjung J, Heringa J. An Overview of Multiple Sequence Alignment. Current Protocols in Bioinformatics. 2003; 3:3.7, 3.7.1–3.7.26.

UniProt C. UniProt: a hub for protein information. Nucleic acids Res. 2015; 43:D204–D212. [PubMed: 25348405]

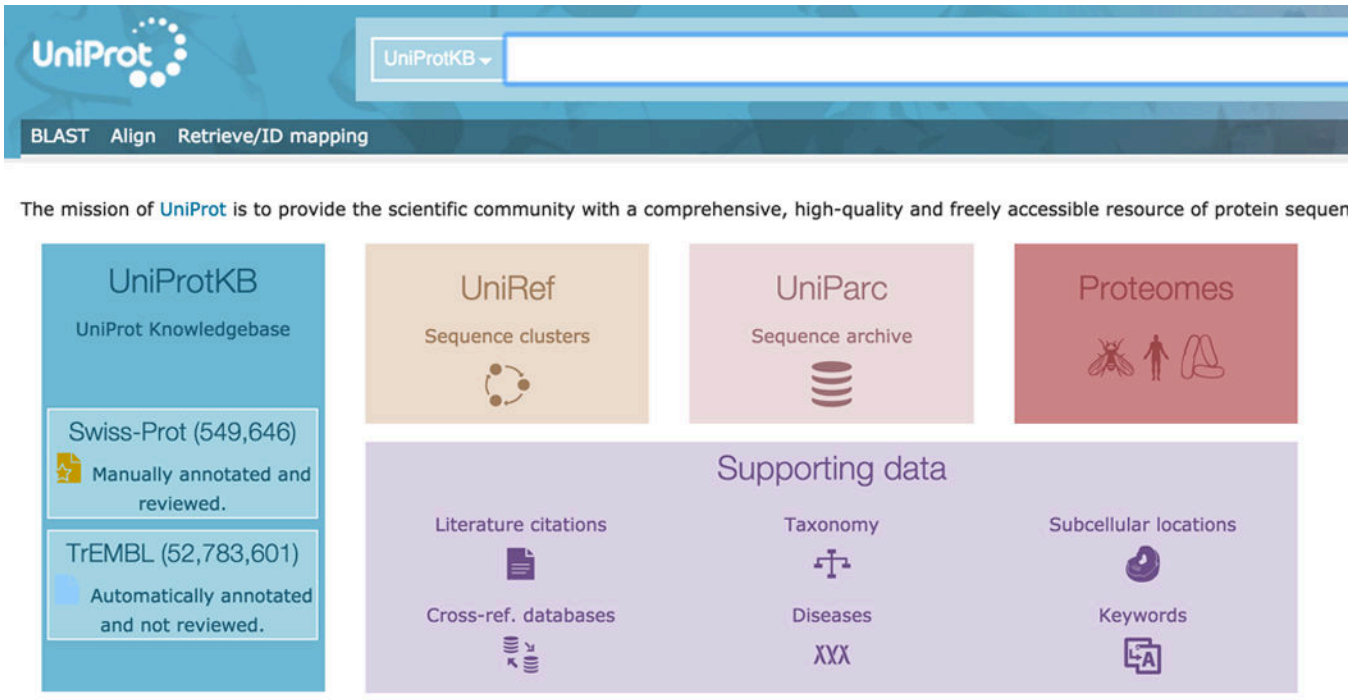

Figure 1.
Blast link in UniProt header

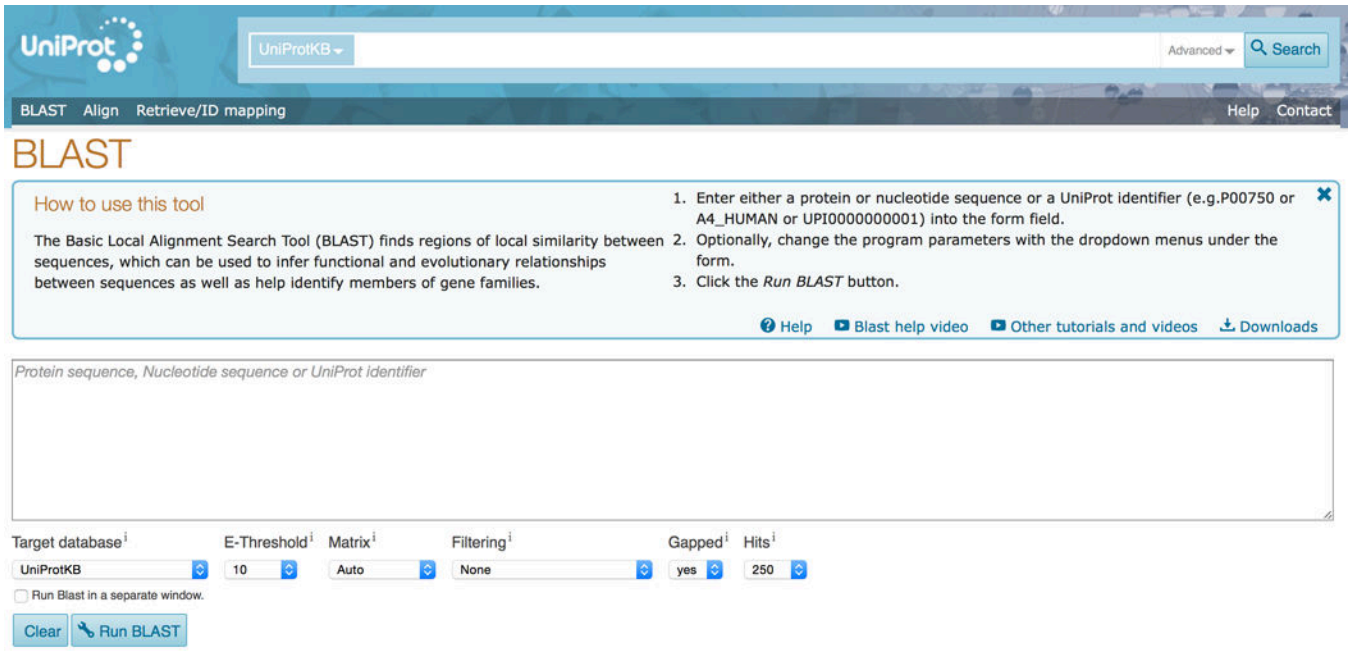

Figure 2.
Blast query input page

Figure 3.
BLAST results page

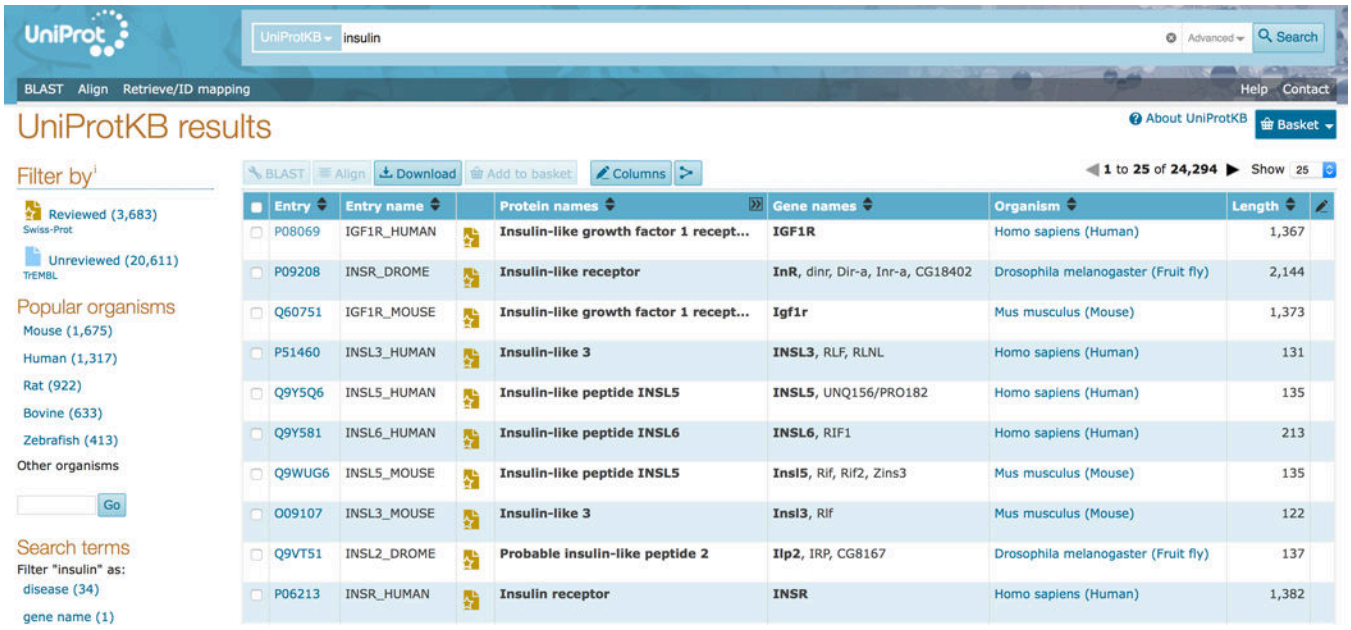

Figure 4.
UniProtKB search results page

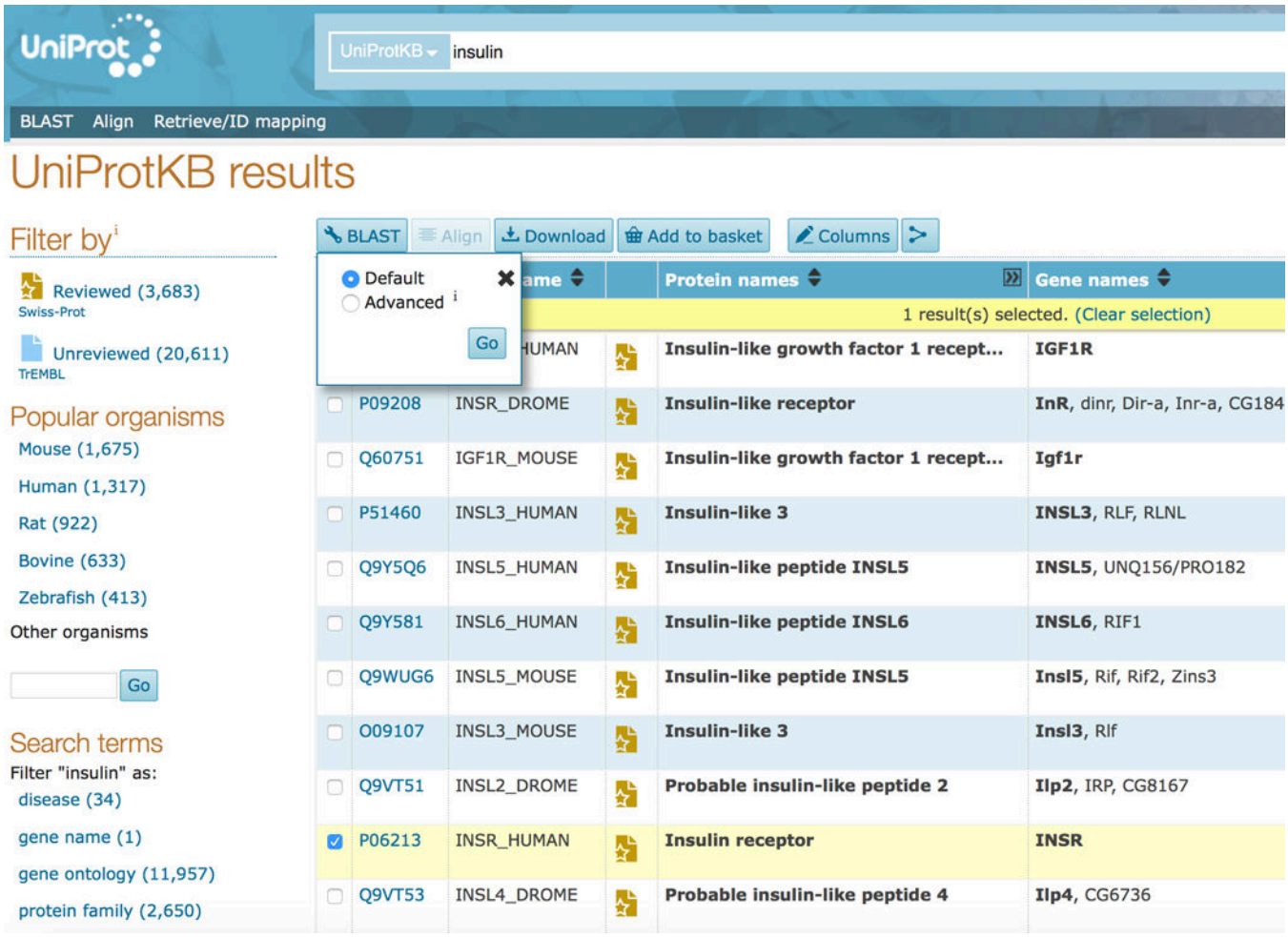

Figure 5.
Selecting a protein to BLAST from the results page

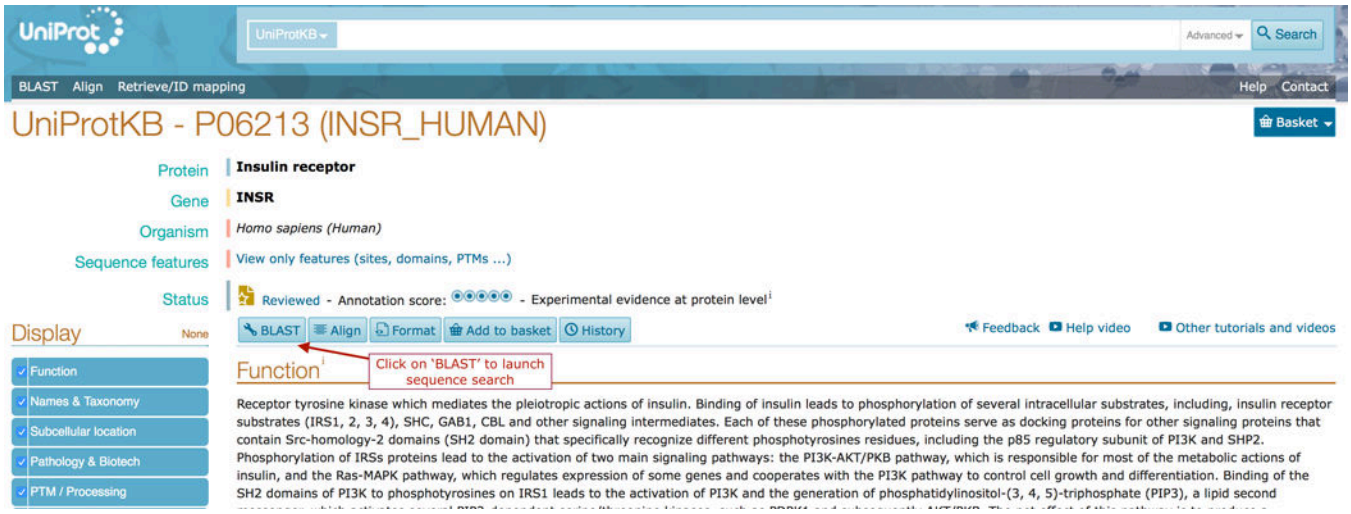

Figure 6.
Running BLAST on a protein from the protein entry page

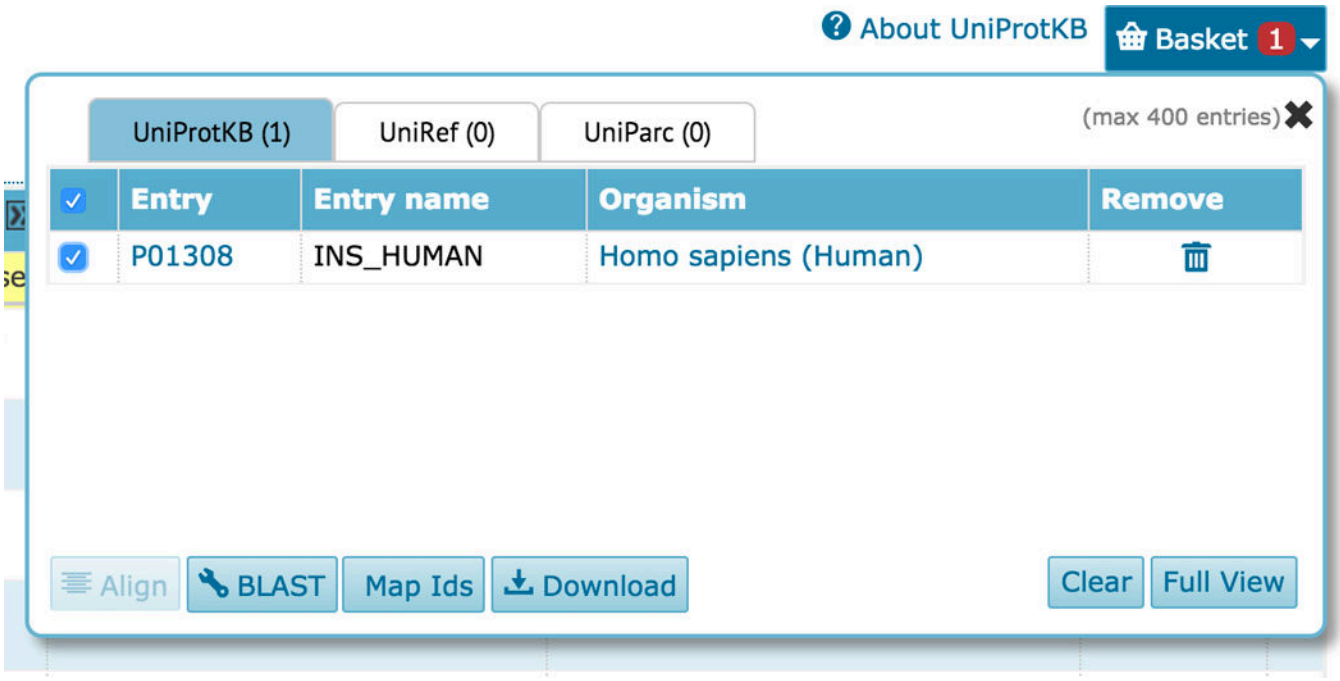

Figure 7.
Running BLAST from the basket

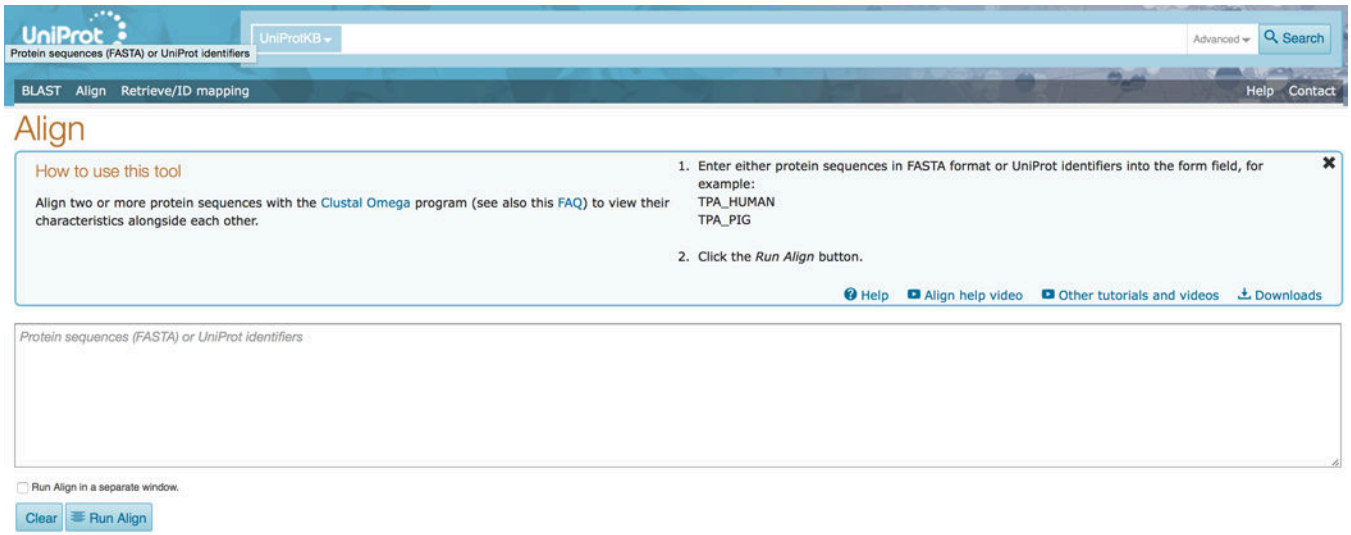

Figure 8.
Align query input page

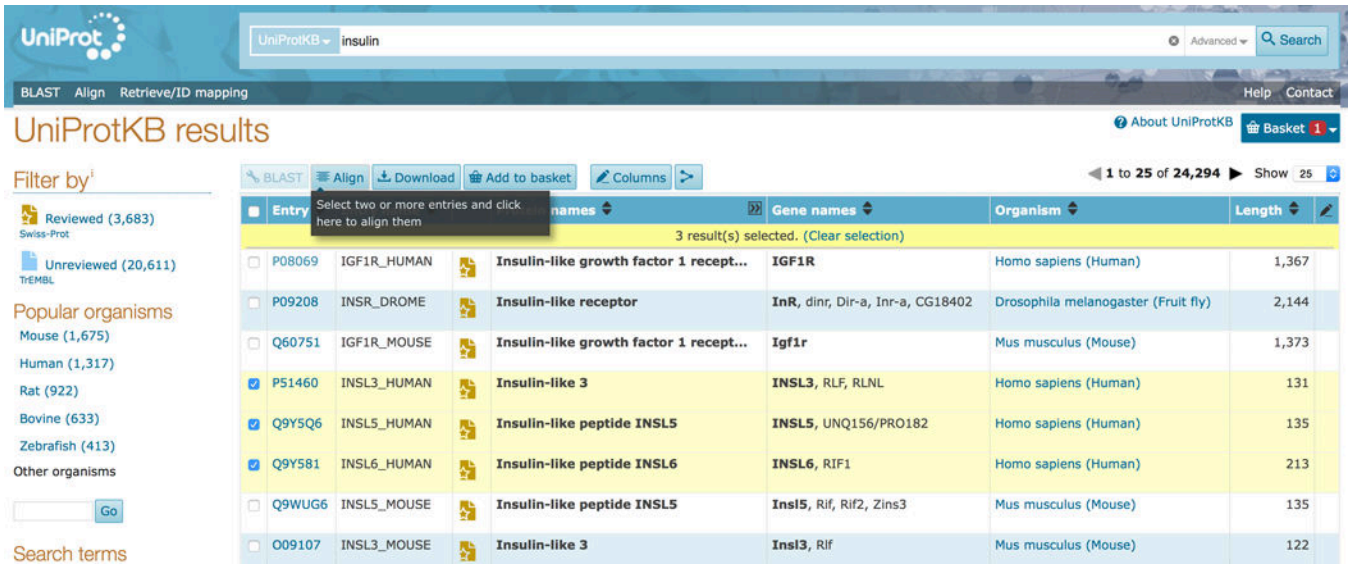

Figure 9.
'Align' multiple protein sequences from UniProtKB results page

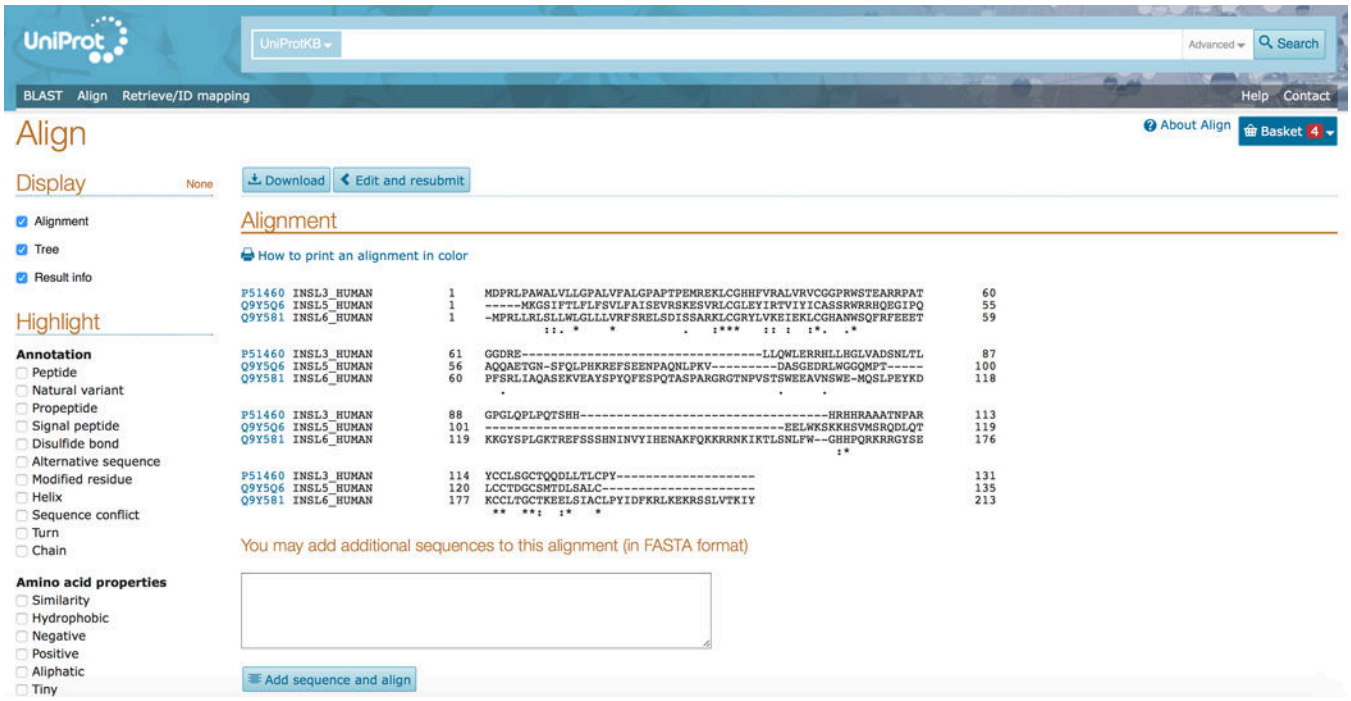

Figure 10.
'Align' results page

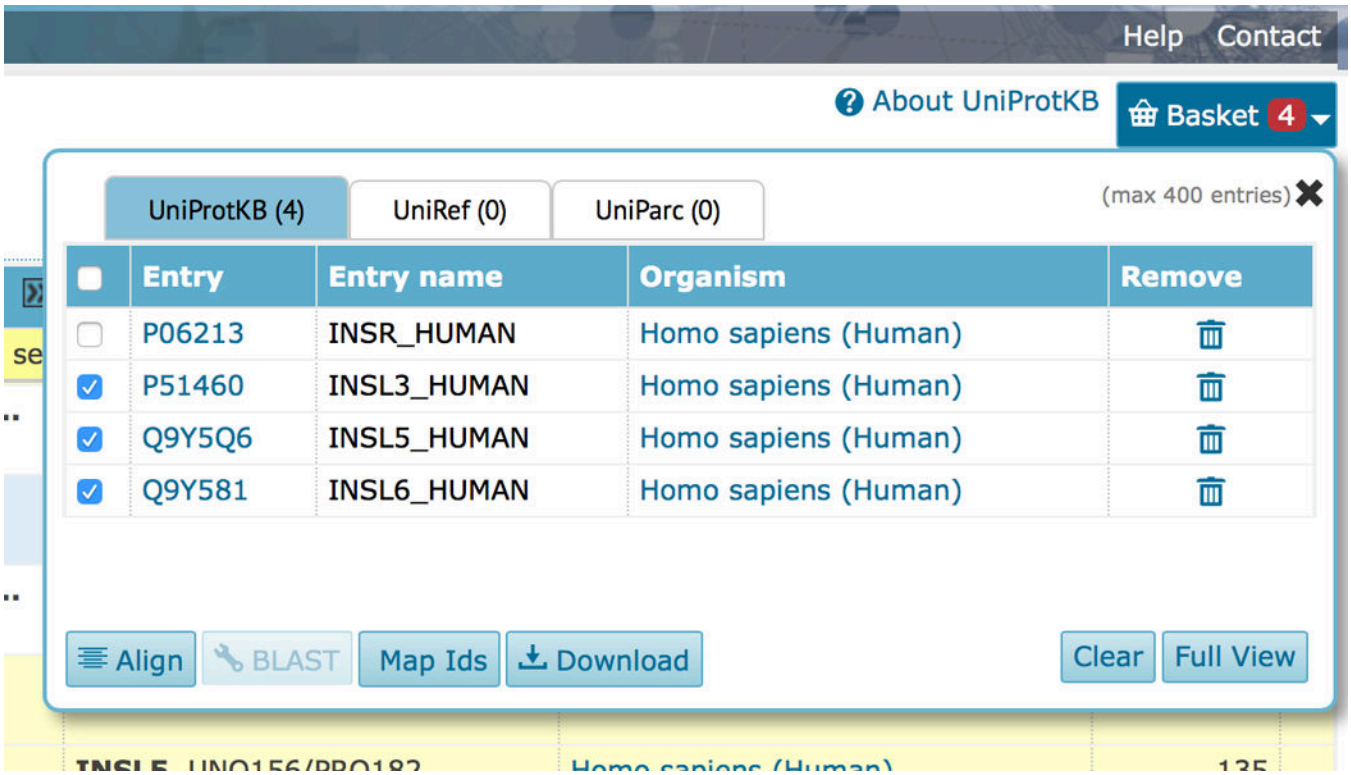

Figure 11.
Aligning proteins from the basket

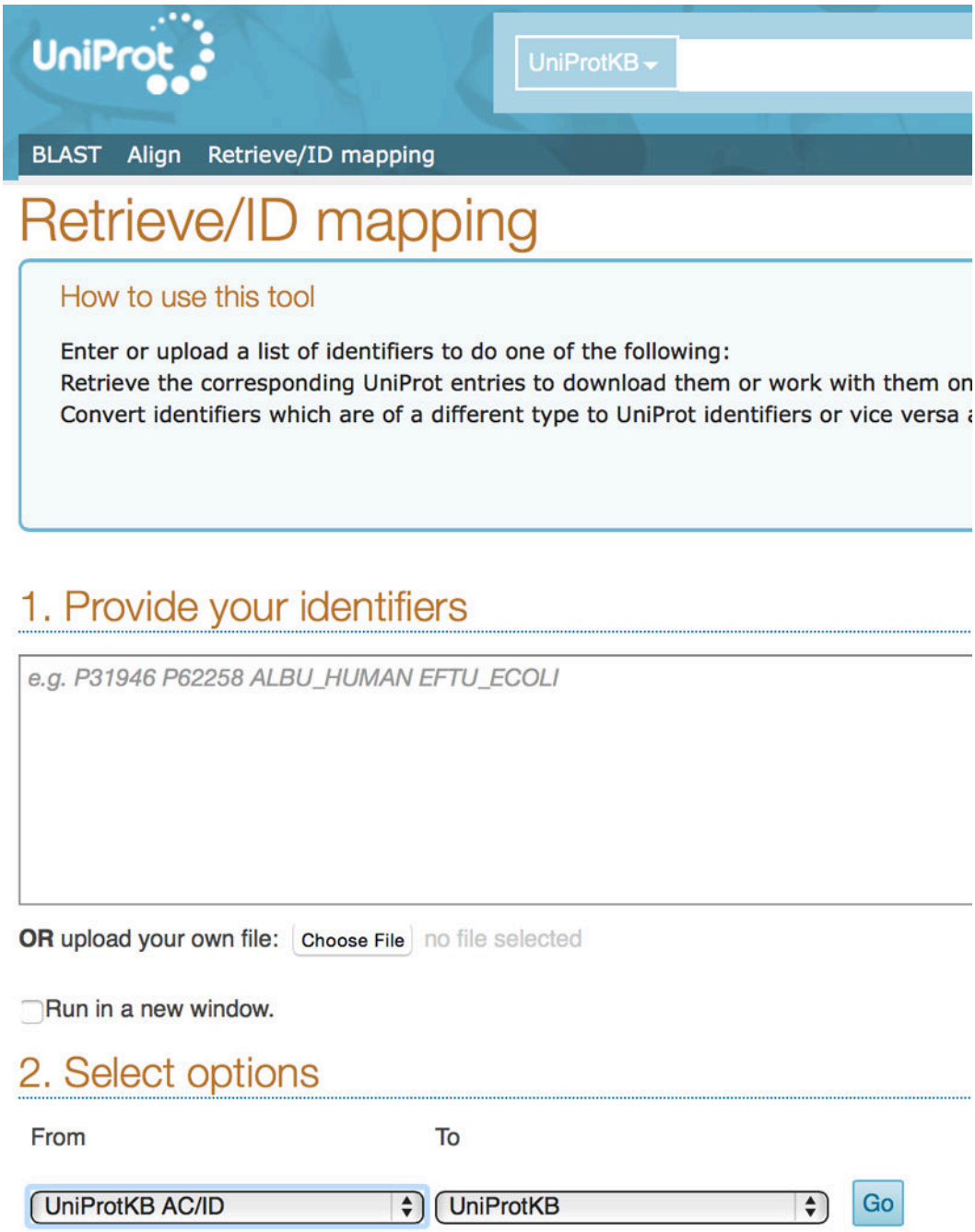

Figure 12.
Retrieve/ID Mapping query input page

Figure 13.
UniProtKB batch results

From	To
P08069	M69229
P08069	AB425196
P08069	BC113610
P08069	AC069029
P08069	AC055807
P08069	AY332722
P08069	X04434
P08069	AC118660
P08069	BC113612
P08069	AC118658
P09208	U28136
P09208	M13568
P09208	M14778
P09208	U18351
P09208	AE014297
Q60751	AF056187
Q60751	M33422
Q60751	U00182
P51460	X73637
P51460	BC053345
P51460	S72482
P51460	AY082014
P51460	BC106721
P51460	AC005952
P51460	AK302780
P51460	BC071706

Figure 14.
Mapping UniProtKB IDs to an external database

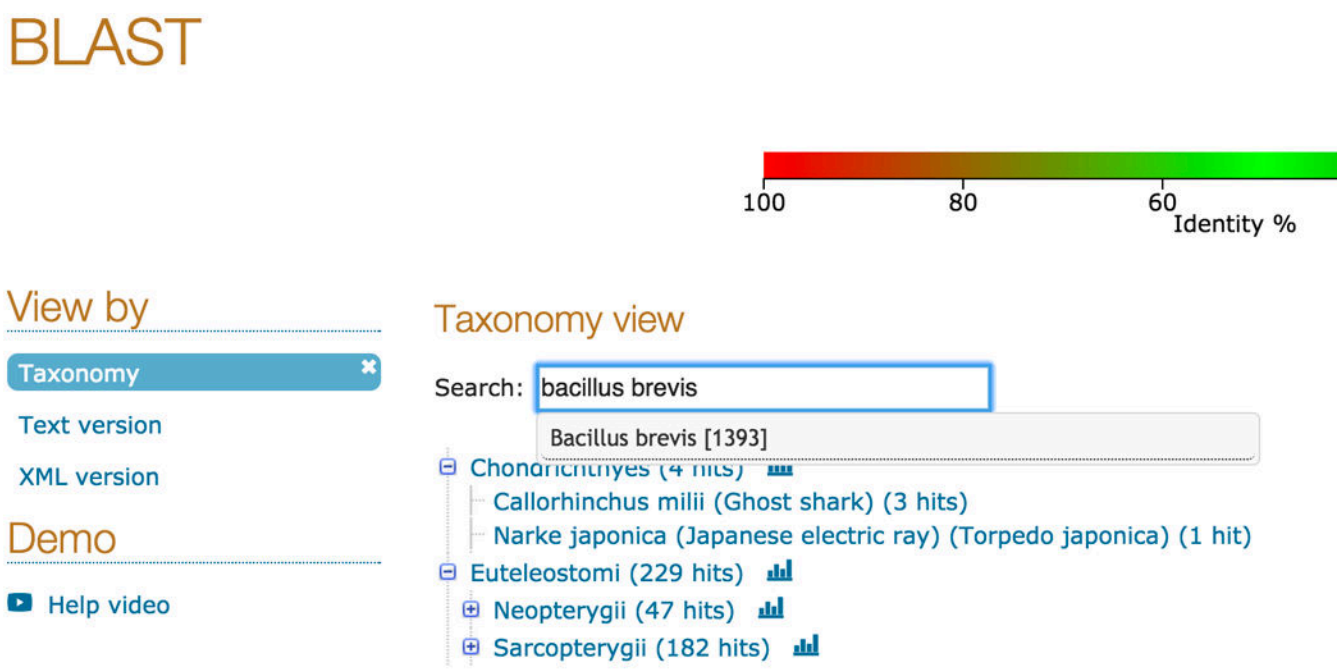

Figure 15.
BLAST results viewed by ‘Taxonomy’

Figure 16.

'Align' results page options to highlight annotations and amino acid properties

Table 1

BLAST advanced options

Database	Database against which the search is performed: UniProtKB or clusters of sequences with 100%, 90% or 50% identity.
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 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 for each position in an alignment that is 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 Xs.
Gapped	This will allow gaps to be introduced in the sequences when the comparison is done.
Hits	Limits the number of returned alignments.