The HMMER Web Server for Protein Sequence Similarity Search

Ananth Prakash,[1] Matt Jeffryes,[1] Alex Bateman,[1] and Robert D. Finn[1]

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

Protein sequence similarity search is one of the most commonly used bioinformatics methods for identifying evolutionarily related proteins. In general, sequences that are evolutionarily related share some degree of similarity, and sequence-search algorithms use this principle to identify homologs. The requirement for a fast and sensitive sequence search method led to the development of the HMMER software, which in the latest version (v3.1) uses a combination of sophisticated acceleration heuristics and mathematical and computational optimizations to enable the use of profile hidden Markov models (HMMs) for sequence analysis. The HMMER Web server provides a common platform by linking the HMMER algorithms to databases, thereby enabling the search for homologs, as well as providing sequence and functional annotation by linking external databases. This unit describes three basic protocols and two alternate protocols that explain how to use the HMMER Web server using various input formats and user defined parameters. © 2017 by John Wiley & Sons, Inc.

Keywords: bioinformatics • homology • profile hidden Markov model • protein sequence analysis

How to cite this article:
Prakash, A., Jeffryes, M., Bateman, A., & Finn, R. D. (2017). The HMMER web server for protein sequence similarity search. *Current Protocols in Bioinformatics*, *60*, 3.15.1–3.15.23. doi: 10.1002/cpbi.40

INTRODUCTION

The HMMER Web server (*http://www.ebi.ac.uk/Tools/hmmer/*) is an open-access protein sequence similarity search tool that hosts a suite of HMMER algorithms to identify evolutionarily related proteins and/or domains by employing profile hidden Markov models (HMMs; APPENDIX 3A, Schuster-Böckler & Bateman, 2007) for fast and efficient detection of close and remote homologs. The HMMER Web server provides four search interfaces to the corresponding algorithms in the HMMER suite (*http://hmmer.org*): *phmmer*, *hmmscan*, *hmmsearch*, and *jackhmmer*. The functionality of these algorithms are outlined in Table 3.15.1. The HMMER Web server can work with various input formats and user-defined parameters to provide results that are presented to help infer protein sequence conservation, function, and evolution. This article provides detailed protocols for using the Web versions of PHMMER, HMMSCAN, and JACKHMMER algorithms, and ways to navigate and interpret the output.

Basic Protocol 1 and Alternate Protocol 1 describe in detail how to use the basic and advanced search features, respectively, in PHMMER, and interpret the results using a protein sequence as the starting point. The logical organization and interpretation of the output described in Basic Protocol 1 is common to all other protocols and is therefore described in detail; user is referred back to this section in subsequent protocols.

Current Protocols in Bioinformatics 3.15.1–3.15.23, December 2017
Published online December 2017 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpbi.40

Table 3.15.1 Description of Programs in the HMMER Web Server. HMMSEARCH is not discussed in this unit

Program	Description
PHMMER	Searches a single protein sequence against target protein sequence databases
HMMSCAN	Searches a single protein sequence against target profile HMM databases
HMMSEARCH	Searches a protein sequence alignment or a profile HMM against target protein sequence databases
JACKHMMER	Iteratively searches a single protein sequence or a sequence alignment or a profile HMM against target protein sequence databases

Basic Protocol 2 describes how to use HMMSCAN to identify domains on a query sequence. Basic Protocol 3 and Alternate Protocol 3 describes how to use the iterative JACKHMMER search, employing an uncharacterized protein sequence as an example, to identify remote homologs and aid in functional annotation. Useful tips on how to obtain input starting material such as protein sequence and how to generate an input multiple sequence alignment are discussed in Support Protocols 1 and 3, respectively.

The user is advised to take note that the underlying profile HMMs and target sequence databases are regularly updated in the HMMER Web server, and the outputs obtained by users may differ from what is shown in this unit. The HMMER Web server is also under constant development using modern Web technologies, and, as a result, old browsers might lose some functionalities.

BASIC PROTOCOL 1

QUICK SEARCH USING PHMMER

The HMMER Web server is hosted at the European Bioinformatics Institute (EMBL-EBI) and can be accessed at *http://www.ebi.ac.uk/Tools/hmmer/*. This basic protocol demonstrates the simple method to search protein homologs using the 'Quick Search' feature, wherein the query protein sequence is searched against a protein sequence database. This protocol and Alternate Protocol 1 are useful when looking for protein homologs that are closely related in their amino acid sequences.

Necessary Resources

An up-to-date Web browser such as Firefox, Safari, or Chrome

Simple sequence similarity search

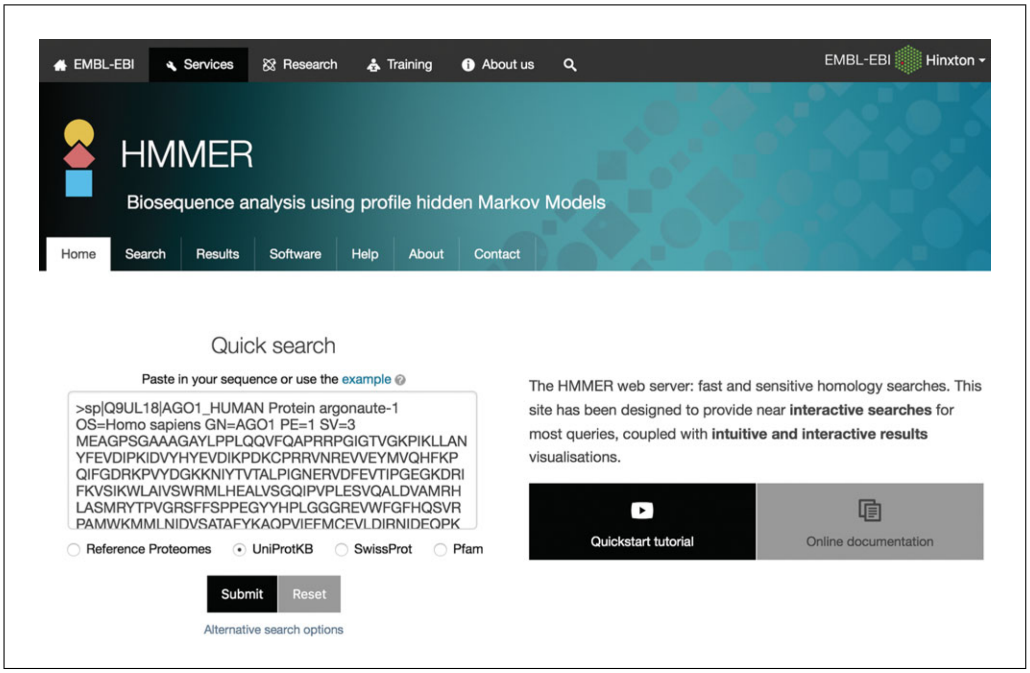
1. Visit the HMMER home page at *http://www.ebi.ac.uk/Tools/hmmer/*.

2. Paste a protein amino acid sequence in the sequence input box.

   If you do not have a protein amino acid sequence on hand, use the example sequence provided or follow Support Protocol 1 to obtain one from a protein sequence database.

   In this protocol the amino acid sequence of the human protein Argonaute-1 (AGO1) (UniProtKB accession: Q9UL18) is used as an example. AGO1 is required for RNA-mediated gene silencing by forming a complex with either miRNA or siRNA and represses the translation of mRNAs that are complementary to them. The amino acid sequence used here is in FASTA format (APPENDIX 1B, Mills, 2014); amino acid sequence without any formatting is also accepted.

3. Select a target database from the options to search and press 'Submit'. Some of the frequently searched against databases include Reference Proteomes, UniProtKB, Swiss-Prot, and Pfam (Fig. 3.15.1).



Quick search

Paste in your sequence or use the example

```
>sp|Q9UL18|AGO1_HUMAN Protein argonaute-1
OS=Homo sapiens GN=AGO1 PE=1 SV=3
MEAGPSGAAAGAYLPPLQQVFQAPRRPGIGTVGKPIKLLAN
YFEVDIPKIDVYHYEVDIKPDKCPRRVNREVVEYMVQHFKP
QIFGDRKPVYDGKKNIYTVTALPIGNERVDFEVTIPGEGKDRI
FKVSIKWLAIVSWRMLHEALVSGQIPVPLESVQALDVAMRH
LASMRYTPVGRSFFSPPEGYYHPLGGGREVWFGFHQSVR
PAMWKMMLNIDVSATAFYKAQPVIEFMCEVLDIRNIDEQPK
```

○ Reference Proteomes ● UniProtKB ○ SwissProt ○ Pfam

The HMMER web server: fast and sensitive homology searches. This site has been designed to provide near **interactive searches** for most queries, coupled with **intuitive and interactive results** visualisations.

Figure 3.15.1 The PHMMER quick sequence search input interface page described in Basic Protocol 1. The input sequence in FASTA format is pasted in the search box and searched against UniProtKB database.

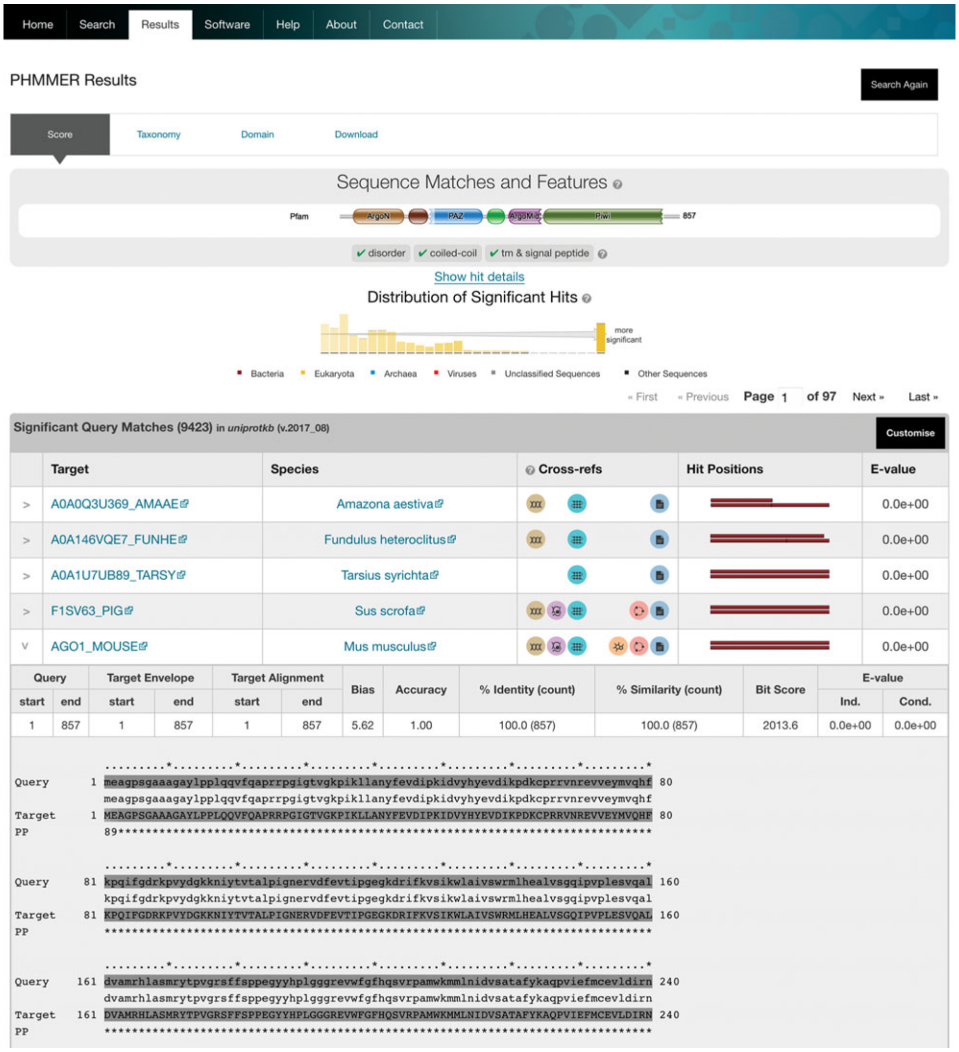
*Each target database is a specialized resource comprising unique annotations and features. Reference Proteomes comprise complete proteomes of well studied model organisms and those that are of interest to biomedical or biotechnological research. UniProtKB (*UNIT 1.27*; Pundir, Magrane, Martin, O'Donovan, & The UniProt Consortium, 2015) comprises a comprehensive collection of protein sequences, classifications, functional annotations, and Gene Ontology information, among other data. Swiss-Prot is a subset of UniProtKB that consists of only high-quality manually curated protein sequences that are usually experimentally characterized; Pfam is a collection of profile HMMs of over 16,700 protein domains (*UNIT 2.5*; Coggill, Finn, & Bateman, 2008). A detailed description of each of these target databases is beyond the scope of this article, but are included in the help pages on the Web site. Please refer to the relevant articles mentioned above for navigating these databases.*

The target database selected in this example is UniProtKB. Clicking the Submit button initiates a similarity search and leads to the Results page.

Interpreting search results

The output in the PHMMER results page is divided into four different tabs—Score, Taxonomy, Domain, and Download. A detailed description of the results from each of the tabs is discussed below.

4. During the process of searching with sequences, the query sequence is converted to a profile HMM and parameterized using a substitution matrix and affine gap open and extend penalties. This profile HMM is what is actually searched against the target sequence databases. The Score tab shows the sequence features of the query together with the matches (homologs) (Fig. 3.15.2). The query sequence is also run against the Pfam database in the background, and its domain architecture is shown. Other sequence features such as disorder regions, coiled-coils, and transmembrane and signal peptides are also searched for and displayed when found as additional graphical tracks.



Significant Query Matches (9423) in *uniprotkb* (v.2017_08)

	Target	Species	Cross-refs	Hit Positions	E-value
>	A0A0Q3U369_AMAAE	Amazona aestiva			0.0e+00
>	A0A146VQE7_FUNHE	Fundulus heteroclitus			0.0e+00
>	A0A1U7UB89_TARSY	Tarsius syrichta			0.0e+00
>	F1SV63_PIG	Sus scrofa			0.0e+00
v	AGO1_MOUSE	Mus musculus			0.0e+00

Query		Target Envelope		Target Alignment		Bias	Accuracy	% Identity (count)	% Similarity (count)	Bit Score	E-value	
start	end	start	end	start	end						Ind.	Cond.
1	857	1	857	1	857	5.62	1.00	100.0 (857)	100.0 (857)	2013.6	0.0e+00	0.0e+00

```
                .........*.........*.........*.........*.........*.........*.........*.........*
Query       1   meagpsgaaagaylpplqqvfqaprrpgigtvgkpikllanyfevdipkidvyhyevdikpdkcprrvnrevveymvqhf 80
                meagpsgaaagaylpplqqvfqaprrpgigtvgkpikllanyfevdipkidvyhyevdikpdkcprrvnrevveymvqhf
Target      1   MEAGPSGAAAGAYLPPLQQVFQAPRRPGIGTVGKPIKLLANYFEVDIPKIDVYHYEVDIKPDKCPRRVNREVVEYMVQHF 80
PP              89******************************************************************************

                .........*.........*.........*.........*.........*.........*.........*.........*
Query      81   kpqifgdrkpvydgkkniytvtalpignervdfevtipgegkdrifkvsikwlaivswrmlhealvsgqipvplesvqal 160
                kpqifgdrkpvydgkkniytvtalpignervdfevtipgegkdrifkvsikwlaivswrmlhealvsgqipvplesvqal
Target     81   KPQIFGDRKPVYDGKKNIYTVTALPIGNERVDFEVTIPGEGKDRIFKVSIKWLAIVSWRMLHEALVSGQIPVPLESVQAL 160
PP              ********************************************************************************

                .........*.........*.........*.........*.........*.........*.........*.........*
Query     161   dvamrhlasmrytpvgrsffsppegyyhplgggrevwfgfhqsvrpamwkmmlnidvsatafykaqpviefmcevldirn 240
                dvamrhlasmrytpvgrsffsppegyyhplgggrevwfgfhqsvrpamwkmmlnidvsatafykaqpviefmcevldirn
Target    161   DVAMRHLASMRYTPVGRSFFSPPEGYYHPLGGGREVWFGFHQSVRPAMWKMMLNIDVSATAFYKAQPVIEFMCEVLDIRN 240
PP              ********************************************************************************
```

Figure 3.15.2 The Score tab from the PHMMER quick search Results page showing domain architecture of query sequence and the resultant homologous sequences. Pairwise alignment for one of the homologous target sequence is shown.

The AGO-1 protein is 857 amino acid residues long and is composed of the N-terminal domain, linker 1, PAZ domain, linker 2, MID, and the PIWI domains. Alignment details and amino acid residue positions for each of these domains can be seen by moving the mouse cursor over them or by clicking the 'Show hit details' option below.

5. The significance of the resulting sequence homologs is measured by their expected-values (E-values; *UNIT 3.1*, Pearson, 2013a). These hits are binned according to their E-values as represented by the histogram from least to most significant, as indicated. The bars are color-coded based on the kingdom.

   Moving the cursor over the bars shows the number of sequences that are representative of them and their E-values. Click on the threshold columns to jump straight to the result table below that highlights sequences that are part of the distribution. There are 987 eukaryotic sequences that have E-values equal to 0 and represent the most significant matches to the query AGO-1.

6. The results table, listing all of the sequence matches (see point 7 below), can be customized by clicking on the Customise button on the far left to either restore to default settings or select among 12 different options to be displayed. The

customize option also allows you to control the number of hits per page to be displayed.

In this example, in addition to the columns shown by default, the Cross-refs and Hit Positions columns are also selected for display.

7. The result table has a total of 9423 protein sequence matches to the query across the UniProtKB database version 2017_08. The first one hundred hits, sorted by their ascending E-values, are shown in the first page. The results are tabulated to show UniProt Identifiers, protein descriptions, species of origin, links to external cross reference databases, schematic representations of matches to the query and target sequences (hit positions), and their E-values.

   Each identified homolog can be viewed by clicking its UniProt identifier, which links to its respective page in the UniProtKB database. As seen from the protein descriptions, the identified homologs are protein argonaute-1 and uncharacterized proteins. The external databases that are cross-referenced are grouped into seven major classes, each shown by a different colored icon, such as (i) genes, genomes, and variations; (ii) gene, protein, and metabolite expression; (iii) protein sequences, families, and motifs; (iv) molecular structures; (v) chemical biology; (vi) systems; and (vii) literature and ontologies. The identified homologs may have annotations for all or most of these features, and they can be viewed individually by selecting the desired group and the cross-reference database. The icon for these groups is absent when no cross-references are available.

   The hit positions depicting the graphical pairwise alignment shows the query sequence on the top and the target sequence at the bottom. This representation provides an easy way to visualize the site of sequence similarity/match between the query and the target sequences and the coverage of their alignments. The first entry in the table, A0A0Q3U369_AMAAE, is argonaute-4 from Amazona aestiva (turquoise-fronted parrot), which is 1676 amino acid residues long. The full-length of the query sequence AGO-1 matches two regions of the homolog with high significance shown by the low E-value.

8. Detailed pairwise alignments of the query with each match can be viewed by clicking on the '>' symbol found to the left of the UniProt identifiers. This reveals a nested view of the matches, with each match and alignment in a table. In the first row of the table, the query start and end values denote the coordinates of the query sequence profile HMM; the target envelope and the target alignment start and end values denote the probable and confident match coordinates of the query with the profile HMM. Other scores displayed include the percentage identity, percentage similarity, bit-score, and independent and conditional E-values. Moving the mouse cursor over the alignment shows a description of the terminologies used in the alignment and its color coding.

9. The Taxonomy tab contains the species distribution of the identified homologs (Fig. 3.15.3). At the top of the page, a taxonomic tree shows distribution of all the identified homologs, with the number of sequences in that taxonomic branch shown within parentheses. The branch can be navigated back and forth by clicking on the arrow buttons. By selecting a particular branch, the species distribution table underneath the tree can be filtered to only show homologs in these organisms. The sequences are grouped based on each organism and their numbers shown. Clicking the Show button for an entry in this table returns to the Score page, but with only the sequences for the species selected shown. This filtering is also applied in the Domain and Download pages, which are discussed below. Filtering can be cancelled by clicking the Cancel button.

   The homologs of human protein AGO-1 that are obtained in this search are largely found in eukaryotic organisms.

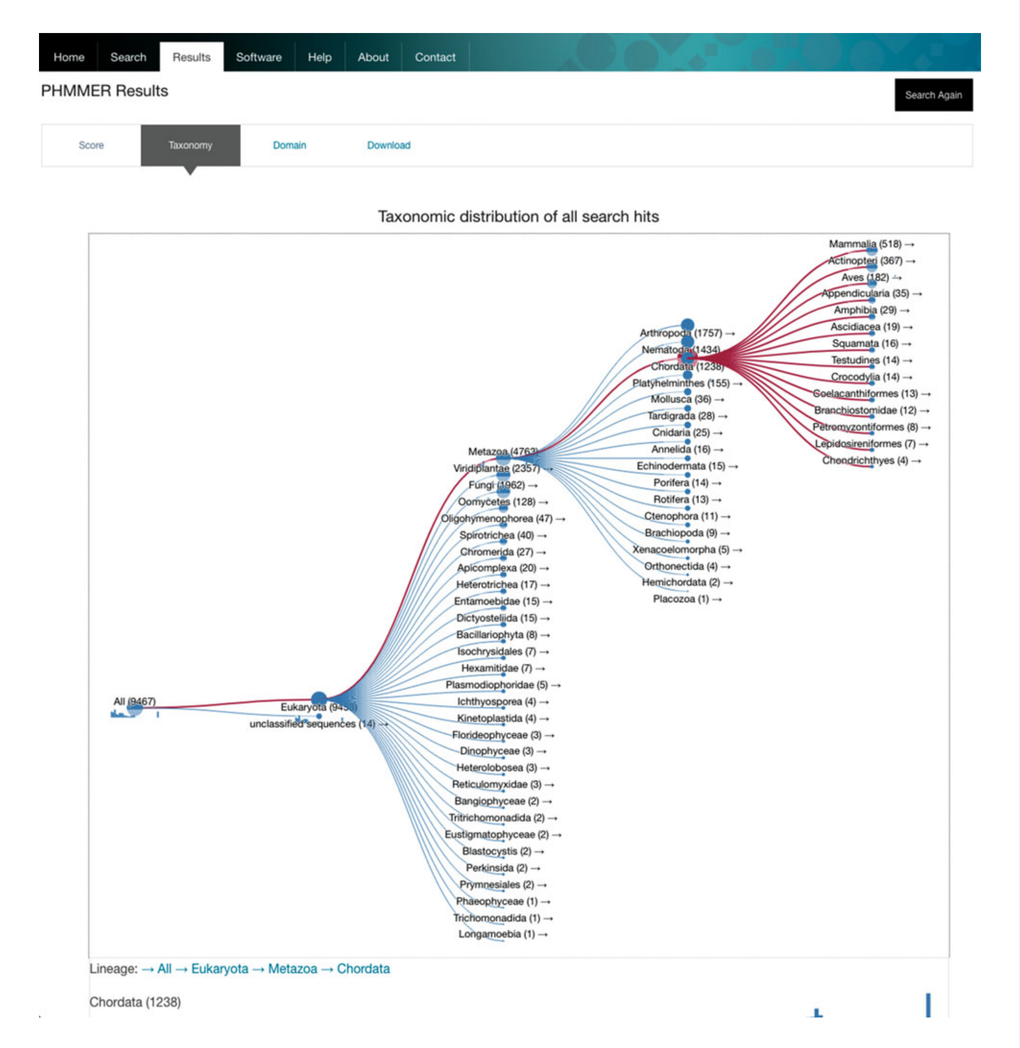


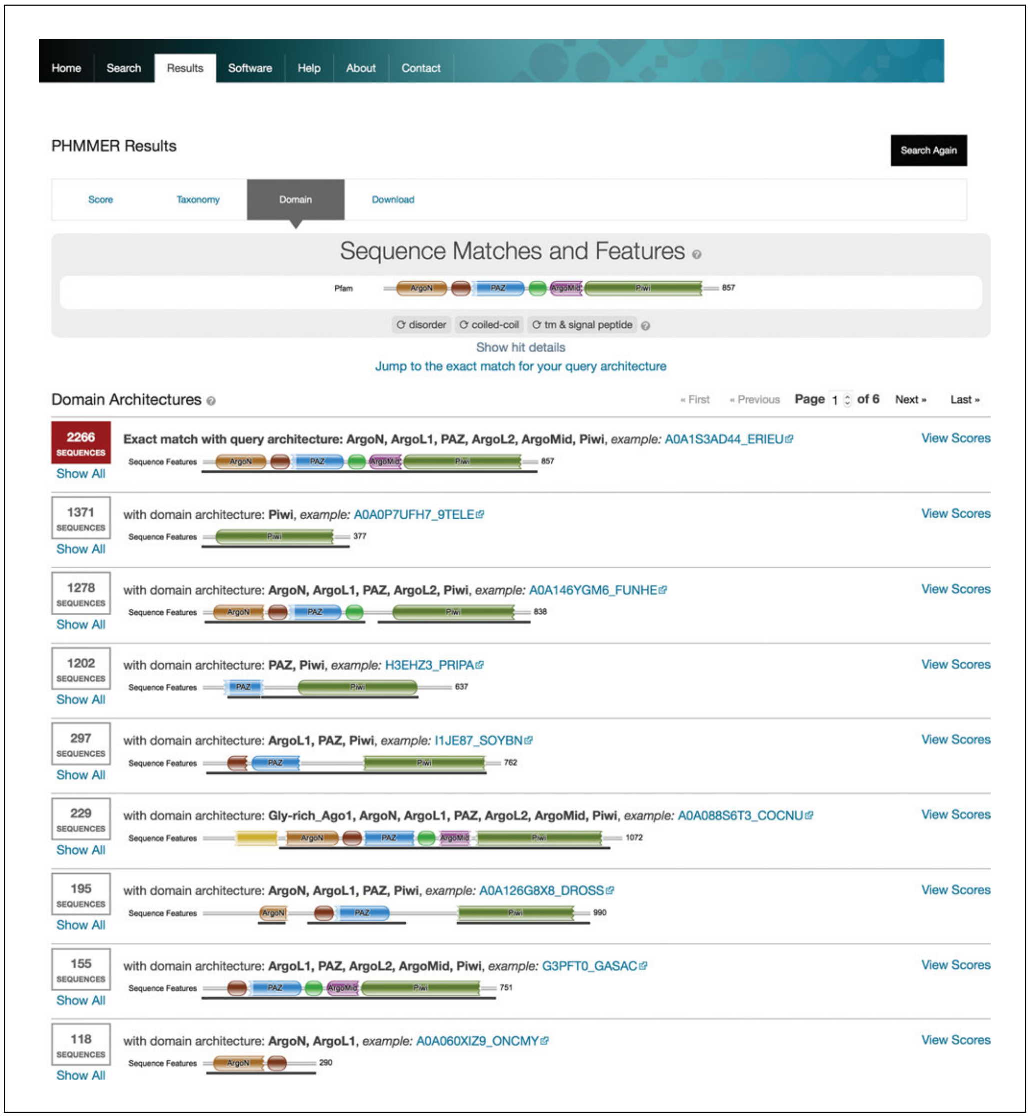
Figure 3.15.3 The Taxonomy tab from the PHMMER quick search Results page showing the distribution of the proteomes in which homologous sequences are found. The branches in the tree can be navigated back and forth by clicking on the arrow button. Numbers on the arrow indicate sequences that are present in the further clades.

10. In the Domain tab, the identified homologs are grouped according to their domain architectures (Fig. 3.15.4). The rows indicate matched sequences that have the same sequential arrangement of domains, and the number of protein sequences in that group is shown to the left of the graphical representation. The Show All option can be clicked to display all proteins within that group. The group of proteins that have a domain architecture identical to that of the query sequence is highlighted with a red background to the sequence count box. This group of sequences can also be found by clicking on the link ‘Jump to the exact match for your query architecture’ on the top. The dark gray line at the bottom denotes the region of target sequence that matches to the query sequence.

    The domain view is an extremely powerful method for interpreting a large result set. Thousands of results can be summarized within a single page.

    In this example, the group with the largest number of hits has a domain architecture that exactly matches with the query.

    To filter the results to show only pairwise alignments of sequences that have a certain domain architecture, click on the View Score option on the left. This works in a



PHMMER Results

Sequence Matches and Features

Show hit details

Jump to the exact match for your query architecture

Domain Architectures

2266 SEQUENCES — **Exact match with query architecture: ArgoN, ArgoL1, PAZ, ArgoL2, ArgoMid, Piwi**, *example:* A0A1S3AD44_ERIEU

1371 SEQUENCES — with domain architecture: **Piwi**, *example:* A0A0P7UFH7_9TELE

1278 SEQUENCES — with domain architecture: **ArgoN, ArgoL1, PAZ, ArgoL2, Piwi**, *example:* A0A146YGM6_FUNHE

1202 SEQUENCES — with domain architecture: **PAZ, Piwi**, *example:* H3EHZ3_PRIPA

297 SEQUENCES — with domain architecture: **ArgoL1, PAZ, Piwi**, *example:* I1JE87_SOYBN

229 SEQUENCES — with domain architecture: **Gly-rich_Ago1, ArgoN, ArgoL1, PAZ, ArgoL2, ArgoMid, Piwi**, *example:* A0A088S6T3_COCNU

195 SEQUENCES — with domain architecture: **ArgoN, ArgoL1, PAZ, Piwi**, *example:* A0A126G8X8_DROSS

155 SEQUENCES — with domain architecture: **ArgoL1, PAZ, ArgoL2, ArgoMid, Piwi**, *example:* G3PFT0_GASAC

118 SEQUENCES — with domain architecture: **ArgoN, ArgoL1**, *example:* A0A060XIZ9_ONCMY

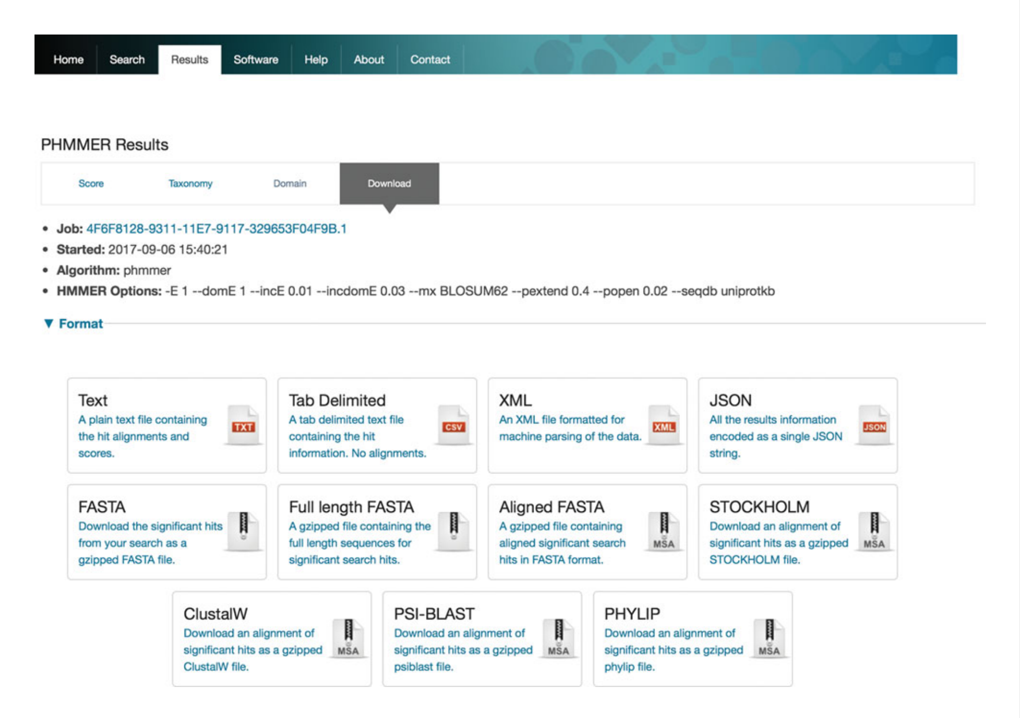
Figure 3.15.4 The Domain tab from the PHMMER quick search Results page showing the resultant homologous sequences grouped by their domain architectures. The group containing homologs with identical domain architectures as that of the query sequence is highlighted by a red box.

similar way to the taxonomy filtering described in step 9, and is applied across all tabs.

Download results

11. Once the desired set of matches, which could be all matches or those of a certain taxonomic lineage and/or certain domain architecture, have been selected or filtered, proceed to the Download tab at the top to download the results (Fig. 3.15.5). The download page displays the provenance of the job: the unique job identifier assigned to this search, which is a link that can be used to retrieve the results later and that can also be used to share the result; the start date and time of the search; the algorithm used, which in this example is PHMMER; and the default parameters that were used by the algorithm to search homologs. If filters have been applied at this stage, but you want the entire search result to be downloaded, click ‘Cancel’.

    There are 11 possible file formats that can be downloaded, and the description of each file format is explained in the download page.



PHMMER Results

- **Job:** 4F6F8128-9311-11E7-9117-329653F04F9B.1
- **Started:** 2017-09-06 15:40:21
- **Algorithm:** phmmer
- **HMMER Options:** -E 1 --domE 1 --incE 0.01 --incdomE 0.03 --mx BLOSUM62 --pextend 0.4 --popen 0.02 --seqdb uniprotkb

▼ Format

Text
A plain text file containing the hit alignments and scores.

Tab Delimited
A tab delimited text file containing the hit information. No alignments.

XML
An XML file formatted for machine parsing of the data.

JSON
All the results information encoded as a single JSON string.

FASTA
Download the significant hits from your search as a gzipped FASTA file.

Full length FASTA
A gzipped file containing the full length sequences for significant search hits.

Aligned FASTA
A gzipped file containing aligned significant search hits in FASTA format.

STOCKHOLM
Download an alignment of significant hits as a gzipped STOCKHOLM file.

ClustalW
Download an alignment of significant hits as a gzipped ClustalW file.

PSI-BLAST
Download an alignment of significant hits as a gzipped psiblast file.

PHYLIP
Download an alignment of significant hits as a gzipped phylip file.

Figure 3.15.5 The Download tab from the PHMMER quick search Results page showing the various file formats the results can be downloaded. The search parameters are indicated at the top.

ALTERNATE PROTOCOL 1

PHMMER ADVANCED SEARCH

The PHMMER sequence similarity search protocol described above can also be searched using the advanced input options available. This alternate protocol demonstrates how to use the advanced search options in order to obtain customized search results.

Necessary Resources

An up-to-date Web browser such as Firefox, Safari, or Chrome

Input amino acid sequence

For simplicity and ease of comparison, the human protein AGO-1, which was used in Basic Protocol 1, is chosen as input here. If you do not have a protein sequence in hand, refer to Support Protocol 1 for help.

1. Visit the HMMER home page: *http://www.ebi.ac.uk/Tools/hmmer/*.

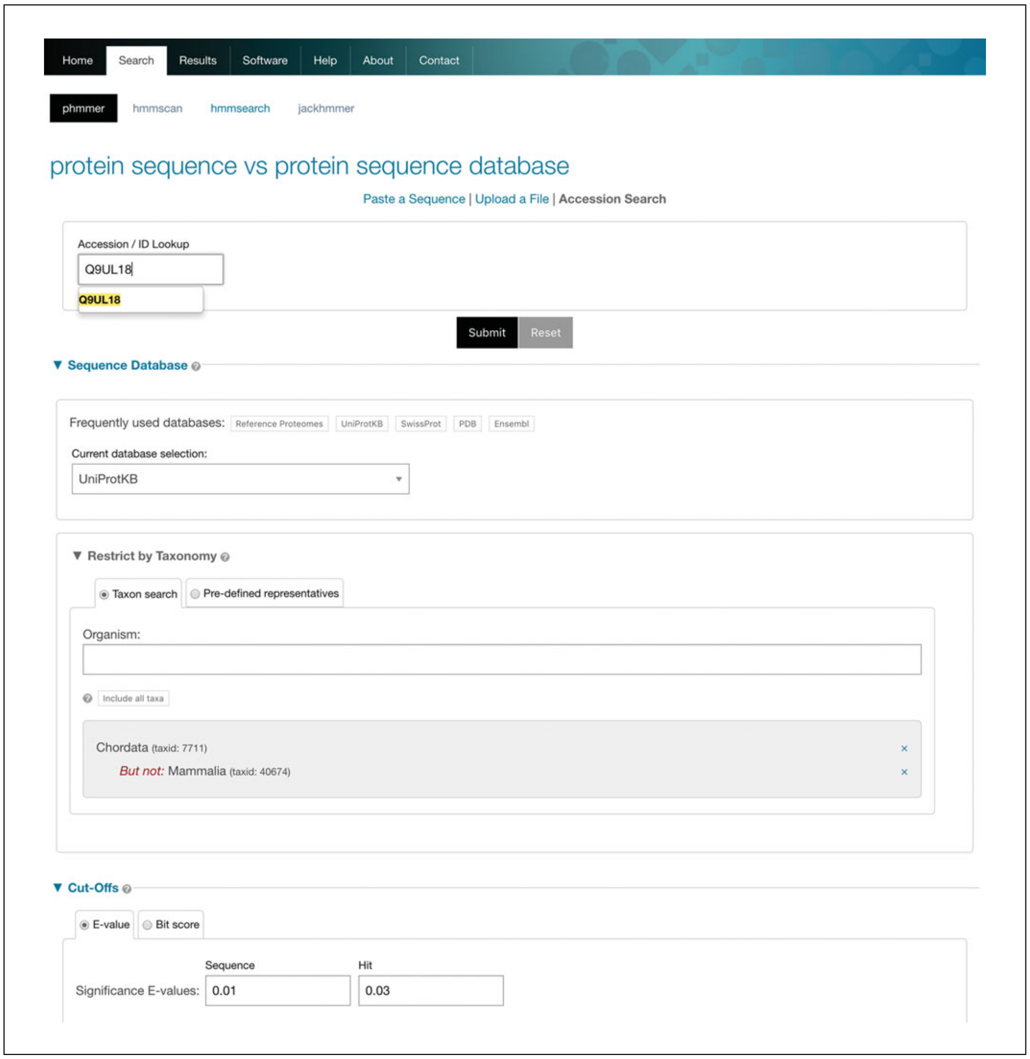
2. Click on the Search tab at the top or click on the ‘Alternative search options’ beneath the Submit and Reset buttons.

3. Instead of pasting the amino acid sequence, there are other input options that allow you to either upload a plain text file containing the input sequence or use the protein accession or identifier. The Accession Search option is case insensitive and can accept UniProt, PDB, and Ensembl accessions or identifiers (Fig. 3.15.6). Typing the first 3 to 5 characters in the accession or identifier in the lookup box displays an autocompletion list of potential entries to select from.

   In this example, the UniProt accession number of human AGO1 (Q9UL18) is selected from the drop-down list.

Advanced input options

4. More target database options are available in addition to the frequently used databases against which the query sequence was searched using Basic Protocol 1.



protein sequence vs protein sequence database

Accession / ID Lookup: Q9UL18

Frequently used databases: Reference Proteomes, UniProtKB, SwissProt, PDB, Ensembl

Current database selection: UniProtKB

Restrict by Taxonomy: ● Taxon search ○ Pre-defined representatives

Organism:

Chordata (taxid: 7711)
But not: Mammalia (taxid: 40674)

Cut-Offs: ● E-value ○ Bit score

	Sequence	Hit
Significance E-values:	0.01	0.03

Figure 3.15.6 The PHMMER advanced search input page as discussed in Alternate Protocol 1. Input using Accession or ID lookup is shown here.

The target search can also be limited to search sequences from a certain taxonomic group, such as genus, species, or a sub-species, by selecting the option Restrict by Taxonomy. The 'Taxon search' tab can be used to look up organisms using either their scientific nomenclature or their generic name; for example, type either `Danio rerio` or `zebra fish`.

More advanced options allow excluding searches against sequences of one or more genus or organisms. Once the phylum is 'included', as described above, the desired genera or specific organisms within this phylum can be 'excluded' by naming them in the organism search box and selecting the option Exclude.

In this example the search for AGO-1 homologs is restricted to non-mammalian vertebrates. To achieve this, first type `Chordata` *in the organism search box and select 'Include: Chordata (taxid: 7711)' from the lookup list. This should select the phylum Chordata, and should be indicated below. Now type* `Mammalia` *in the organism search box and select 'Exclude: Mammalia (taxid: 40674)' from the lookup list. These custom filters should now be indicated in the box below as 'Chordata (taxid: 7711) But not: Mammalia (taxid: 40674)'. Multiple genera or organisms can be excluded from searches in a similar way.*

To browse through the list of proteomes taxonomically and select the target organism sequence database of choice, click on the 'Predefined representatives' tab. This option is helpful to easily select multiple target genera or organisms of choice. For example, if the names of target bacterial species are not known, then select the genus.

5. The ‘Cut-offs’ option allows the use of a significance threshold to refine the search and report significant hits. These thresholds can be set as either E-values (default option) or as Bit-scores. There are two cut-off options available, the ‘Significance E-values’ and the ‘Report E-values’.

   The Significance E-values are set to determine whether a sequence can be considered significant or not. The Sequence cut-off value applies for the entire sequence, while the Hit cut-off applies for each match within the sequence and is useful when searching for homologs of a query that has sequence repeats.

   The number of insignificant sequence hits to be displayed can be limited by setting the Report E-values. Sequence hits with E-values higher than the Significance E-values but lower than the Report E-value are displayed with a yellow background separated from the significant hits by a red line. Refer to the HMMER User’s Guide for a description of Bit-Scores (*http://hmmer.org/documentation.html*).

   The significance of the resultant homologs can be gauged by their E-values. E-values offer a means to check if a resultant sequence found by the search is a true homolog of the query or if it is found within the database by the algorithm just by chance. E-values closer to 1 indicate that the sequence hit is just found by chance, and an E-value close to 0 indicates a true homolog. In this example, the default E-values are used, which are 0.01 and 0.03 for the Sequence and Hit options, respectively, of Significance E-values and 1 for both the Sequence and Hit options of Report E-values.

More advanced options

6. Click on the Advanced button to show more advanced input options such as Customise Results, Gap Penalties, and Filters.

7. The Customise Results option is similar to that described in Basic Protocol 1 (step 6). The options allow up to 12 different fields to be displayed along with the number of entries to be displayed per page.

   In addition to the default options, select ‘Cross-refs’ and ‘Hit positions’.

8. The Gap Penalties options allows setting the gap opening and gap extension penalties in an alignment between the query sequence (or the model) and the target sequence.

   In this example the default values are used, which are 0.02 and 0.4 for gap opening and gap extension, respectively.

9. The substitution scoring matrix is used to alter the stringency of alignment between the query and the target sequence (*UNIT 3.5*, Pearson, 2013a). There are five scoring matrices available—BLOSUM45, BLOSUM62, BLOSUM90, PAM30, and PAM70. BLOSUM90 and BLOSUM62 can be used to search more closely related sequences compared to BLOSUM45. PAM70 can be used to find more distantly related sequences compared to PAM30.

   The default matrix BLOSUM62, is used in this example.

10. Composition bias filter is turned on by default to prevent too many sequences passing the filter and slowing down search.

11. Scroll up and press Submit.

Interpreting search results

For the most part, the search results are explained as in Basic Protocol 1 (steps 4 to 10), since the same query was used as input. The differences, however, lie in the number of target hits obtained and their taxonomic distribution.

Compared to 9423 protein sequences that were identified initially, after restricting to non-mammalian vertebrates, the search has resulted in 715 hits, which indicates that the vast majority of human AGO-1 homologs are found in mammals. The taxonomic distribution of the resulting homologs indicates that most non-mammalian homologs belong to classes Aves (birds) and Actinopteri (ray-finned fishes). The search restriction is displayed at the beginning of the result table. The search can be repeated again without any restrictions using the Resubmit button.

Another round of search

The following steps are optional.

12. The Search Again box at the top right hand side of the result page can be clicked if another round of searching needs to be performed. There are two available options: (a) 'Perform a new search with new input', which redirects to the blank sequence submit page, or (b) 'Perform a new search with these results', which takes into account the result from the current search and performs a second iteration.

    The second option to perform a new search with existing results creates a profile HMM of the 715 significant hits and uses this profile HMM as the query to perform a second iteration against the sequence database. It should be noted here that opting for the second iteration automatically performs the search using HMMSEARCH. The advanced input parameters discussed above are reset to the defaults during the second iteration.

Download results

13. The download results options are similar to that previously described in Basic Protocol 1 (step 11).

PROTEIN SEQUENCE DATABASES

SUPPORT PROTOCOL 1

This protocol explains how to obtain a protein amino acid sequence to use as an input for Basic Protocol 1 or Alternate Protocol 1. There are many sequence resources through which one can obtain protein amino acid sequences. Two of the most well-known protein sequence databases—UniProtKB (*UNIT 1.27*; Pundir et al., 2015) and NCBI Protein (*UNIT 1.3*; Gibney & Baxevanis, 2011)—are discussed below.

Necessary Resources

An up-to-date Web browser such as Firefox, Safari, or Chrome

To select protein sequence of interest from UniProtKB

1a. Visit the UniProtKB home page at *http://www.uniprot.org/*.

2a. Type the protein name `Argonaute` in the search tab and press Search.

3a. From the result table, select entry with protein name Protein Argonaute-1 from *Homo sapiens* by clicking on the UniProt entry (accession) Q9UL18.

In UniProt, a protein entry can be searched by using its name or its gene name or using its UniProt accession or identifier. Some of the advanced search options include restricting the search to an organism and/or manually curated high-quality sequences (Swiss-Prot). The protein entry Q9UL18 is manually curated, as denoted by the Reviewed status.

4a. In the page for entry Q9UL18, click on the Sequence option at the left to directly jump to the sequence results or scroll down until you find the sequence entry. Click on the option FASTA at the top to view the amino acid sequence in FASTA format.

Copy and paste the sequence in the PHMMER sequence search box as explained in Basic Protocol 1, or save the sequence in a text file and upload it to the PHMMER sequence search option as explained in Alternate Protocol 1.

To select protein sequence of interest from NCBI

1b. Visit NCBI home page at *http://www.ncbi.nlm.nih.gov/*.

2b. Select Protein from the drop-down menu at the top. Type `Argonaute-1 AND` "`Homo sapiens`" in the search box and press Search.

In NCBI Protein, a protein entry can also be accessed using the accession or GenInfo Identifier (GI).

3b. From the hits, select the entry by clicking on 'protein argonaute-1 isoform 1 [Homo sapiens]' with accession NP_036331.1 and length 857 amino acids.

4b. Click on the FASTA option to view the amino acid sequence in FASTA format. Copy and paste the sequence in the PHMMER sequence search box as explained in Basic Protocol 1 or save the sequence in a text file and upload it to PHMMER sequence search option as explained in Alternate Protocol 1.

BASIC PROTOCOL 2

QUICK PROFILE SEARCH USING HMMSCAN

This protocol describes how to query a protein sequence against profile HMM-based protein family databases, such as Pfam.

Necessary Resources

An up-to-date Web browser such as Firefox, Safari, or Chrome

Input amino acid sequence

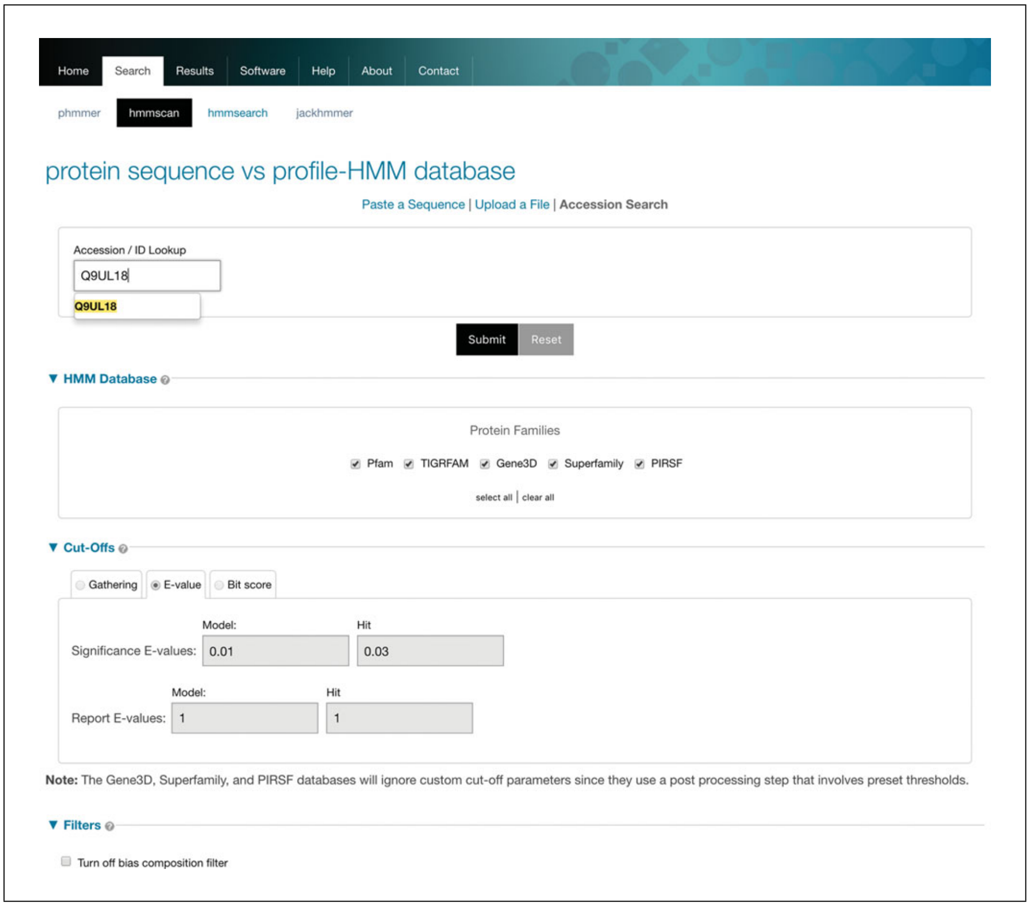
1. Visit the HMMER home page *http://www.ebi.ac.uk/Tools/hmmer/*.
2. Click on the Search tab at the top and select hmmscan from the options.
3. Input the amino acid sequence of human protein Argonaute-1 by either pasting the sequence or uploading the text file or using the UniProt accession (Q9UL18) as described in Basic Protocol 1 or Alternate Protocol 1.

Select target profile HMM database

4. Select one or more target HMM databases by clicking on the check boxes. For this example, all target databases are selected (Fig. 3.15.7).

The target databases—Pfam, TIGRFAM, Gene3D, Superfamily, and PIRSF—are collections of protein families modeled using profile HMMs. The Pfam database is a collection of profile HMMs of over 16,700 protein domains (UNIT 2.5, Cogill et al., 2008; also see Finn et al., 2016); TIGRFAM comprises multiple sequence alignments and HMMs of mostly prokaryotic proteins (Haft et al., 2013); Gene3D is a collection of domain annotations of protein sequences predicted using the HMM libraries of CATH superfamilies (UNIT 1.28; Sillitoe, Lewis, & Orengo, 2015); Superfamily comprises domain annotations of protein sequences predicted using the HMM libraries based on SCOP superfamilies (Wilson et al., 2009); and the PIRSF database comprises HMMs for comprehensive non-overlapping clustering of protein sequences (Wu et al., 2004).

5. If target databases Gene3D, Superfamily, or PIRSF are selected (including via the select all button), the custom cut-off E-values are ignored.



protein sequence vs profile-HMM database

Accession / ID Lookup: Q9UL18

Q9UL18

▼ HMM Database

Protein Families

☑ Pfam ☑ TIGRFAM ☑ Gene3D ☑ Superfamily ☑ PIRSF

▼ Cut-Offs

○ Gathering ● E-value ○ Bit score

	Model:	Hit
Significance E-values:	0.01	0.03
Report E-values:	1	1

Note: The Gene3D, Superfamily, and PIRSF databases will ignore custom cut-off parameters since they use a post processing step that involves preset thresholds.

▼ Filters

□ Turn off bias composition filter

Figure 3.15.7 The HMMSCAN search input page as discussed in Basic Protocol 2.

Advanced option

6. The advanced input option consists of the bias composition filter. By default this option is turned on.

7. Click Submit to run the job.

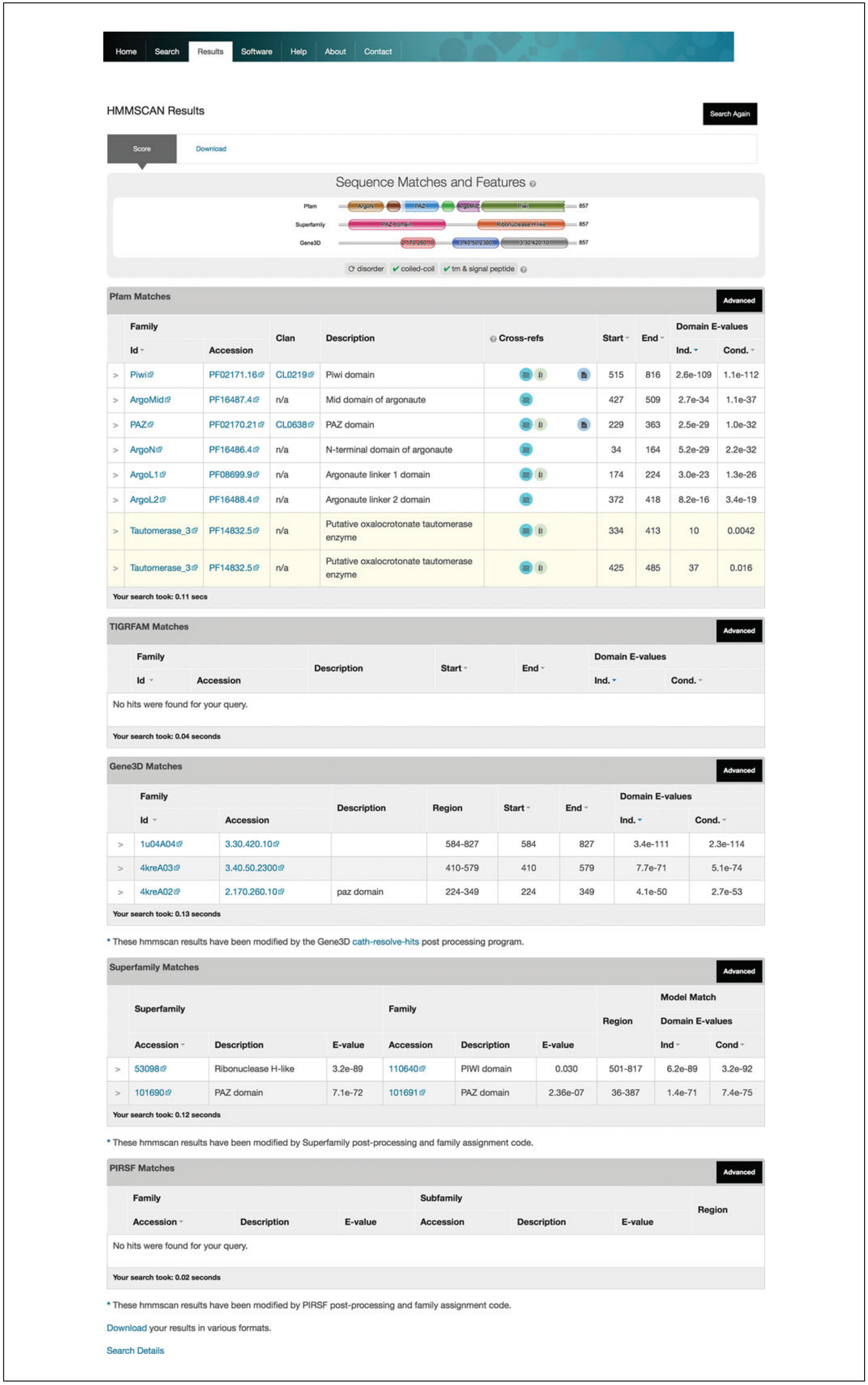
Interpreting search results

8. The Score tab displays the results from the target databases (Fig. 3.15.8). The domain architectures of the query sequence is shown at the top.

   The query sequence is run in the background against all the databases. In this example, there are no matches of the query sequence to the TIGRFAM and PIRSF databases. The domain architectures of the query sequence as shown by Pfam, Superfamily, and Gene3D are shown. The domain architectures shown are different due to the way the profile HMMs of each domain are defined in these databases. In the graphical representation, round corners of a domain denote a complete match, while jagged edges indicate partial match of the query to the profile HMM of the respective domain family.

   Scroll over individual domain features to see their descriptions, sequence, and profile HMM alignment coordinates.

9. The Pfam Matches table lists domain families whose profile HMMs were matched to the query sequence. For each domain in the sequence, their identifier, Pfam accession, clan, cross-reference links to external databases, the start and end sequence coordinates respective to that domain and E-values are listed. Matches to domain profile HMMs that are below the set threshold limits are highlighted with a yellow background.



HMMSCAN Results

Sequence Matches and Features

Pfam Matches

	Family		Clan	Description	Cross-refs	Start	End	Domain E-values	
	Id	Accession						Ind.	Cond.
>	Piwi	PF02171.16	CL0219	Piwi domain		515	816	2.6e-109	1.1e-112
>	ArgoMid	PF16487.4	n/a	Mid domain of argonaute		427	509	2.7e-34	1.1e-37
>	PAZ	PF02170.21	CL0638	PAZ domain		229	363	2.5e-29	1.0e-32
>	ArgoN	PF16486.4	n/a	N-terminal domain of argonaute		34	164	5.2e-29	2.2e-32
>	ArgoL1	PF08699.9	n/a	Argonaute linker 1 domain		174	224	3.0e-23	1.3e-26
>	ArgoL2	PF16488.4	n/a	Argonaute linker 2 domain		372	418	8.2e-16	3.4e-19
>	Tautomerase_3	PF14832.5	n/a	Putative oxalocrotonate tautomerase enzyme		334	413	10	0.0042
>	Tautomerase_3	PF14832.5	n/a	Putative oxalocrotonate tautomerase enzyme		425	485	37	0.016

Your search took: 0.11 secs

TIGRFAM Matches

Family		Description	Start	End	Domain E-values	
Id	Accession				Ind.	Cond.

No hits were found for your query.

Your search took: 0.04 seconds

Gene3D Matches

	Family		Description	Region	Start	End	Domain E-values	
	Id	Accession					Ind.	Cond.
>	1u04A04	3.30.420.10		584-827	584	827	3.4e-111	2.3e-114
>	4kreA03	3.40.50.2300		410-579	410	579	7.7e-71	5.1e-74
>	4kreA02	2.170.260.10	paz domain	224-349	224	349	4.1e-50	2.7e-53

Your search took: 0.13 seconds

* These hmmscan results have been modified by the Gene3D cath-resolve-hits post processing program.

Superfamily Matches

	Superfamily			Family			Region	Model Match	
								Domain E-values	
	Accession	Description	E-value	Accession	Description	E-value		Ind	Cond
>	53098	Ribonuclease H-like	3.2e-89	110640	PIWI domain	0.030	501-817	6.2e-89	3.2e-92
>	101690	PAZ domain	7.1e-72	101691	PAZ domain	2.36e-07	36-387	1.4e-71	7.4e-75

Your search took: 0.12 seconds

* These hmmscan results have been modified by Superfamily post-processing and family assignment code.

PIRSF Matches

Family			Subfamily			Region
Accession	Description	E-value	Accession	Description	E-value	

No hits were found for your query.

Your search took: 0.02 seconds

* These hmmscan results have been modified by PIRSF post-processing and family assignment code.

Download your results in various formats.

Search Details

Figure 3.15.8 The Score tab from the HMMSCAN search Results page showing domain annotations.

Click on the Advanced button on the left to see details of start and end coordinates of the alignment, the start and end coordinates of the model, and profile HMM length and the Bit-score.

Click on the link with the identifier name of a domain or the accession to visit the summary page in Pfam. Click on the Back button of the browser to return to the result page.

10. Similarly the ‘Gene3D Matches and the Superfamily Matches tables list hits to the domain HMMs from the CATH and Superfamily databases, respectively, along with their descriptions, sequence coordinates, and alignment scores. These scores are different from the native HMMER output because of the post-processing step performed by these databases. Click on the ‘Id’ or ‘Accession’ links to see their respective CATH and Superfamily database entries.

11. As indicated earlier, there are no matches to the query sequence in the TIGRFAM and PIRSF databases.

Download results

12. Click on the Download tab at the top to download the HMMscan search results in text, tab-delimited, XML, or JSON formats. The Job link can be used to share the results or to retrieve later.

ITERATIVE SEARCHING WITH JACKHMMER

BASIC PROTOCOL 3

This basic protocol demonstrates how to search for protein homologs using the JACKHMMER search algorithm. This protocol should be used when there is a need to identify more distantly related sequences than can be found using Basic Protocol 1.

JACKHMMER is an iterative search method for searching protein sequence databases. When a protein sequence is used as an input, in the first iteration, a PHMMER search is performed against the target sequence database. The results of this search are aligned, and a profile-hidden Markov model (HMM) is constructed. In the second iteration, this profile HMM is used as the input for a HMMSEARCH search against the same target database. In subsequent iterations, a new profile HMM is constructed from the result sets, and a new HMMSEARCH search is performed. If a sequence alignment or profile HMM is used as an input to JACKHMMER, the first, and all subsequent iterations, use HMMSEARCH. Using a sequence alignment or profile HMM as input is only possible using the JACKHMMER Web interface and not possible using the command line version. The user is able to continue searching through as many iterations as is desired, or until the search result set no longer changes between iterations. This method is able to find more distant homologs than a single HMMSEARCH or PHMMER search.

Necessary Resources

An up-to-date Web browser such as Firefox, Safari, or Chrome

Input amino acid sequence

1. Visit the HMMER home page at *http://www.ebi.ac.uk/Tools/hmmer/*, click the Search tab, and then select ‘jackhmmer’.

2. Paste an amino acid sequence of interest in FASTA format into the search box, upload a text file containing a sequence in FASTA format, or provide a UniProt accession or identifier.

   If you do not have a protein sequence at hand, use the example sequence provided or see Support Protocol 1 to obtain one from a protein sequence database. In this protocol, *the amino acid sequence of an uncharacterized protein with the UniProtKB accession R6FBL8, is used.*

3. Select a target database. Modify the E-value cutoffs if desired.

   See Basic Protocol 1, step 2, for a short description of the various target databases. The database used in this example is Reference Proteomes. Modifying the Significance and Report E-value cut-offs can change the protein sequences that are used to construct the profile HMM and the sequences that are displayed in the search results respectively.

   In HMMER, search hits have both a sequence and a domain significance score. The sequence significance score is derived from the probability that any part of the sequence is homologous with the query sequence or profile HMM. The domain or hit significance score is derived from the probability that a specific region of the protein is homologous with the query sequence or profile HMM. For a sequence to be included in the next profile HMM, both the sequence E-value and the hit E-value for the region that the previous profile HMM matches against must be less than their respective cut offs.

   The Report E-value cut-off affects which protein sequences are displayed in the search results, and has no effect on the profile HMM. In general, the Report E-values should be higher than the significance E-values, as it can be desirable to see protein sequences that are borderline for inclusion in the profile HMM.

4. The search can be started by clicking the Submit button. By default, this will start only the first iteration of the search.

   If desired, up to five iterations can be run automatically by specifying the 'Number of iterations' under the 'Batch Options', which are shown by clicking Advanced. Further iterations can be started manually.

Interpreting results

5. Result table features are described in Basic Protocol 1. If multiple iterations were specified as described in the annotation to step 4, above, then the Results Summary page will be displayed immediately (see step 6, below).

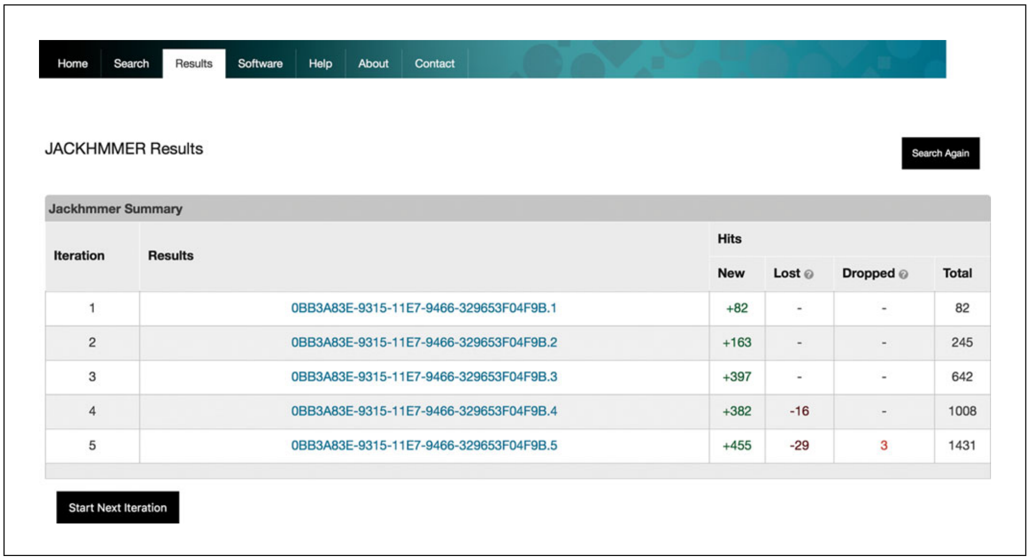
   In the Sequence Matches and Features boxes, we can see that R6FBL8 does not match any Pfam families. By looking at the Domain tab, we can investigate whether other sequences in the search results match any Pfam families. The majority of sequences in the search results do not have any Pfam hits along the length of their sequence. The remaining sequences match against various Pfam families, but not in the region matched by the search's profile HMM. In order to find more distantly related homologous proteins, we can continue iterating the search.

Iterative search

6. From the results page for the first iteration, the second iteration can be started by clicking 'Start iteration 2'. This will display the JACKHMMER Results Summary page, from which further iterations can be started.

   By default, all regions that exceed the cut-offs, as described in step 3, will be included in the profile HMM. If desired, sequences can be manually included or excluded from the profile HMM using the checkboxes on the far right of the matches table. All automatically included sequences can be removed from the profile HMM by selecting 'unselect all' in the Sequence selection options at the top of the page. If the 'unselect all' option is used, manually select some sequences to include in the subsequent round.

7. The summary table displays information about the number of sequences that were added or removed from the profile HMM (Fig. 3.15.9). The New column is the number of sequences in the profile HMM which were not in the profile HMM for the previous iteration. The Lost column is the number of sequences that were in the previous profile HMM but which are no longer reported in the results. The Dropped column is the number of sequences that were in the previous profile HMM but which

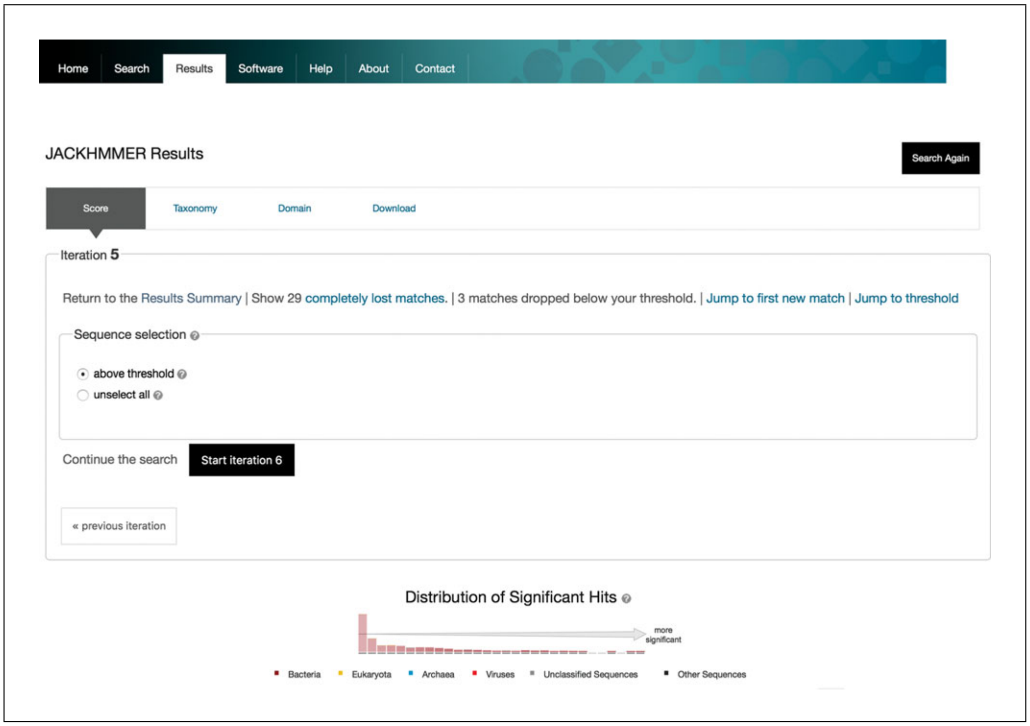


JACKHMMER Results

Jackhmmer Summary

Iteration	Results	Hits			
		New	Lost	Dropped	Total
1	0BB3A83E-9315-11E7-9466-329653F04F9B.1	+82	-	-	82
2	0BB3A83E-9315-11E7-9466-329653F04F9B.2	+163	-	-	245
3	0BB3A83E-9315-11E7-9466-329653F04F9B.3	+397	-	-	642
4	0BB3A83E-9315-11E7-9466-329653F04F9B.4	+382	-16	-	1008
5	0BB3A83E-9315-11E7-9466-329653F04F9B.5	+455	-29	3	1431

Figure 3.15.9 JACKHMMER Result Summary, after five iterations of the search. The total number of sequences found by each iteration of the search are shown, alongside the number of sequences gained, lost, and dropped.



JACKHMMER Results

Iteration 5

Return to the Results Summary | Show 29 completely lost matches. | 3 matches dropped below your threshold. | Jump to first new match | Jump to threshold

Sequence selection

● above threshold
○ unselect all

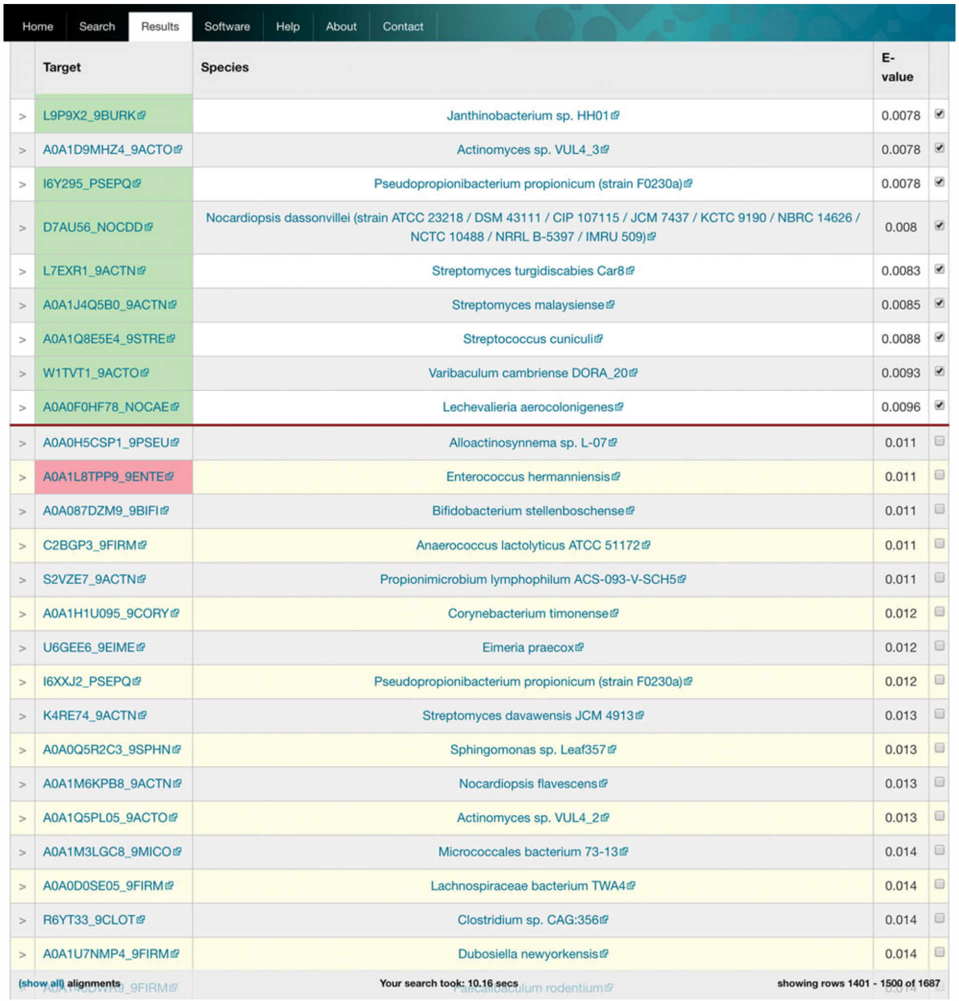
Continue the search

Distribution of Significant Hits

Figure 3.15.10 JACKHMMER Result page for the fifth iteration of a search. The next iteration of the search can be started by clicking ‘Start iteration 6′.

have dropped below the significance level but are still reported. The comparison is made only with the previous iteration. A sequence which is present in the first iteration, lost from the second, and returns in the third will be counted as new.

8. Iterations can be inspected in more detail by clicking the identifier of the iteration in the summary table. Beyond the first iteration, the results table is similar to the table produced by PHMMER (see Basic Protocol 1), but has some additions. The box above the table allows us to see which sequences have been lost from the search, to jump to the most significant new match, and to jump to significance threshold (Fig. 3.15.10). Sequences above the significance threshold will be included in the



	Target	Species	E-value	
>	L9P9X2_9BURK	Janthinobacterium sp. HH01	0.0078	☑
>	A0A1D9MHZ4_9ACTO	Actinomyces sp. VUL4_3	0.0078	☑
>	I6Y295_PSEPQ	Pseudopropionibacterium propionicum (strain F0230a)	0.0078	☑
>	D7AU56_NOCDD	Nocardiopsis dassonvillei (strain ATCC 23218 / DSM 43111 / CIP 107115 / JCM 7437 / KCTC 9190 / NBRC 14626 / NCTC 10488 / NRRL B-5397 / IMRU 509)	0.008	☑
>	L7EXR1_9ACTN	Streptomyces turgidiscabies Car8	0.0083	☑
>	A0A1J4Q5B0_9ACTN	Streptomyces malaysiense	0.0085	☑
>	A0A1Q8E5E4_9STRE	Streptococcus cuniculi	0.0088	☑
>	W1TVT1_9ACTO	Varibaculum cambriense DORA_20	0.0093	☑
>	A0A0F0HF78_NOCAE	Lechevalieria aerocolonigenes	0.0096	☑
>	A0A0H5CSP1_9PSEU	Alloactinosynnema sp. L-07	0.011	☐
>	A0A1L8TPP9_9ENTE	Enterococcus hermanniensis	0.011	☐
>	A0A087DZM9_9BIFI	Bifidobacterium stellenboschense	0.011	☐
>	C2BGP3_9FIRM	Anaerococcus lactolyticus ATCC 51172	0.011	☐
>	S2VZE7_9ACTN	Propionimicrobium lymphophilum ACS-093-V-SCH5	0.011	☐
>	A0A1H1U095_9CORY	Corynebacterium timonense	0.012	☐
>	U6GEE6_9EIME	Eimeria praecox	0.012	☐
>	I6XXJ2_PSEPQ	Pseudopropionibacterium propionicum (strain F0230a)	0.012	☐
>	K4RE74_9ACTN	Streptomyces davawensis JCM 4913	0.013	☐
>	A0A0Q5R2C3_9SPHN	Sphingomonas sp. Leaf357	0.013	☐
>	A0A1M6KPB8_9ACTN	Nocardiopsis flavescens	0.013	☐
>	A0A1Q5PL05_9ACTO	Actinomyces sp. VUL4_2	0.013	☐
>	A0A1M3LGC8_9MICO	Micrococcales bacterium 73-13	0.014	☐
>	A0A0D0SE05_9FIRM	Lachnospiraceae bacterium TWA4	0.014	☐
>	R6YT33_9CLOT	Clostridium sp. CAG:356	0.014	☐
>	A0A1U7NMP4_9FIRM	Dubosiella newyorkensis	0.014	☐

(show all) alignments   Your search took: 10.16 secs   showing rows 1401 - 1500 of 1687

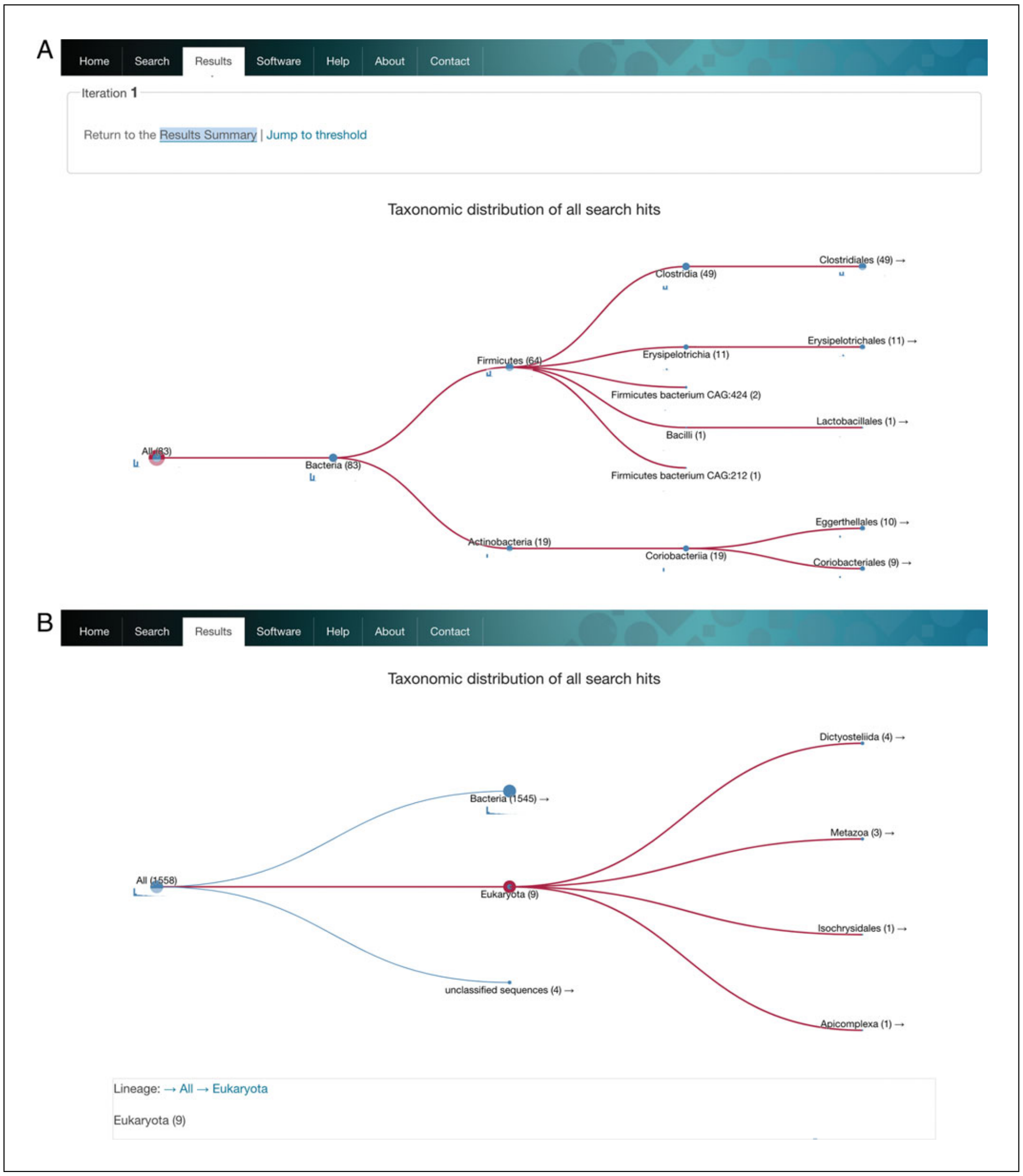
Figure 3.15.11 JACKHMMER query match table showing the significance threshold. Sequences highlighted in green are new additions to the next iteration's profile HMM. Sequences highlighted in red will be dropped from the next profile HMM.

search, and sequences below it will be excluded (this can be modified by manually including or excluding sequences). Targets highlighted in green in the table are sequences that were not included in the profile HMM for the previous iteration. Targets highlighted in red were previously above the significance threshold, but have now fallen below it (Fig. 3.15.11).

9. From the second iteration, a profile HMM logo (Wheeler, Clements, & Finn, 2014) is displayed at the bottom of the results table. This is a visual representation of the probability of a particular residue occurring at a particular coordinate in the sequence. This probability is encoded in the profile HMM, and was used to retrieve the results in this iteration of the search.

10. As with PHMMER searches, domain and taxonomy summaries can be displayed for each iteration of the search by selecting the appropriate tab.

The taxonomy summary demonstrates the ability of JACKHMMER to find distant homologs to the query sequence. With the query sequence R6FBL8, the first iteration of the search identifies only bacterial sequences. By the fifth iteration, the search identifies eukaryotic sequences that are homologous with the query sequence (Fig. 3.15.12).



Lineage: → All → Eukaryota

Eukaryota (9)

Figure 3.15.12 Taxonomy summary page for the first and fifth iterations of a JACKHMMER search. (**A**) In the first iteration, only bacterial sequences are found. (**B**) By the fifth, homologous eukaryotic sequences are identified.

Download results

11. By selecting the Download tab, each iteration can be downloaded, as described in Basic Protocol 1.

JACKHMMER USING PROFILE HMM OR MULTIPLE SEQUENCE ALIGNMENT AS INPUT

ALTERNATE PROTOCOL 2

Iterative JACKHMMER search can also be performed using a profile hidden Markov model (HMM) as input. This performs a search as described in Basic Protocol 3, but with all iterations using HMMSEARCH. The profile HMM can either be uploaded, or chosen from a protein classification database, such as Pfam. Alternatively, a multiple sequence alignment can be used as input. In this case, the first iteration of the search will be performed by constructing a profile HMM from the alignment.

Necessary Resources

An up-to-date Web browser such as Firefox, Safari, or Chrome

Input profile HMM

1a. Visit the HMMER home page at *http://www.ebi.ac.uk/Tools/hmmer/*, click the Search tab, and then select 'jackhmmer'.

2a. Paste a profile HMM in HMMER format into the search box, or upload it by selecting Upload a File. Alternatively, you can use the HMM accession or identifier to fetch the HMM. Click on Accession Search. Type the Pfam, TIGRfam, PIRSF, or Gene3D accession in the format `<accession>@<database>`.

For example, to use the Pfam globin family as the query HMM, enter `PF00042@pfam` *into the lookup box.*

3a. Follow advanced input options as described in Alternate Protocol 1 (steps 4 to 7).

Input multiple sequence alignment

1b. Visit the HMMER home page at *http://www.ebi.ac.uk/Tools/hmmer/*, click the Search tab, and then select 'jackhmmer'.

2b. Paste a multiple sequence alignment into the search box, or upload a multiple sequence alignment by selecting Upload a File. The alignment can be in the following formats: Aligned FASTA, Clustal, PSI-BLAST, PHYLIP, Selex, GCG/MSF, STOCKHOLM, or UC Santa Cruz A2M.

A method to generate a multiple sequence alignment is described in Support Protocol 3. Multiple sequence alignment are also discussed further in UNITS 2.3 AND 3.8.

3b. Follow advanced input options as described in Alternate Protocol 1 (steps 4 to 7).

Further steps

4b. Follow Basic Protocol 3 from step 5 onwards to interpret search results, perform further iterations, and download results.

SUPPORT PROTOCOL 2

GENERATING A MULTIPLE SEQUENCE ALIGNMENT

This protocol demonstrates how to generate a multiple sequence alignment from three or more protein amino acid sequences to use as input in Alternate Protocol 3. JACKHMMER accepts input multiple sequence alignments in various formats, which include Clustal, Aligned FASTA, PSI-BLAST, PHYLIP, Selex, GCG, MSF, STOCKHOLM, and UC Santa Cruz A2M (alignment to model).

Necessary Resources

An up-to-date Web browser such as Firefox, Safari, or Chrome

Multiple sequence alignment

1. Visit *http://www.ebi.ac.uk/Tools/msa/*. Select Clustal Omega by clicking on Launch Clustal Omega.

2. Select Protein and paste three or more sequences of homologous proteins in FASTA format along with their headers. Alternatively, a text file containing amino acid sequences in FASTA format can be uploaded.

3. Select any one of the output formats that is supported by JACKHMMER. In this example, select output to be displayed in STOCKHOLM format. Click Submit.

4. In the result page, click on the Download Alignment File to view alignment. Save this as a text file or copy and paste the alignment as described in Alternate Protocol 3 (step 2).

COMMENTARY

Background Information

Sequence similarity search methods offer a powerful means to identify evolutionarily related proteins based on their amino acid composition (*UNIT 3.1*; Pearson, 2013a). Historically, efforts to identify evolutionarily related proteins, or homologs, using sequence similarity approaches were enhanced by the development of amino acid substitution matrices such as PAM (Point Accepted Mutation) by Margaret Dayhoff in 1978 and BLOSUM (BLOck SUbstitution Matrix) by Steven and Jorja Henikoff in 1992 (Henikoff & Henikoff, 1992). The similarity between two protein sequences can be assessed by looking at their sequence divergence—these substitution matrices describe the rate at which one amino acid within a protein sequence is changed to another over the evolutionary time scale. The PAM and BLOSUM matrices indicate the log-odds ratio of the observed versus expected frequency of individual amino acids. Depending on the divergence of protein sequences, different substitution matrices can be used to score pairwise sequence alignments. The sequence similarity search algorithms such as BLAST (Altschul et al., 1997), FASTA (Pearson & Lipman et al., 1988), and PHMMER, discussed here, employ variations of the PAM and BLOSUM substitution matrices to find homologs using protein amino acid sequence as query to search against sequence databases.

The next major improvement in sensitivity of sequence similarity search algorithms was achieved via the use of position-based scoring [e.g., Position Specific Scoring Matrix (PSSM) and profile HMMs], where the substitution rate of each amino acid is scored independently, based on a set of evolutionarily related and aligned sequences. Two methods that use such profiles are PSI-BLAST (Altschul et al., 1997) and JACKHMMER (Finn, Clements, & Eddy, 2011), which use PSSMs and profile HMMs respectively. Both are iterative searches that are typically initiated with a single sequence, with subsequent searches based on a profile derived from the set of sequences found in the previous search. However, comparing the results of equivalent searches between the two algorithms shows that profile HMM–based methods are, more often than not, more sensitive. This is because profile HMMs employ full probabilistic modeling (without the use of mathematical approximations) and not only represent the divergence in protein sequences through amino acid substitutions but also parameterize rates of insertion or deletion observed in the input alignment (whereas PSSMs use affine gap open and extend penalties).

Prior to HMMER3, profile HMM–based methods were only widely used in protein family databases, such as Pfam, as the searches were too slow, taking hours to search a large sequence collection. HMMER3 represented a major increment in speed—a Pfam release that took months to calculate using HMMER2 could be completed in a matter of hours on the same size database with the same amount of compute resources. Furthermore, the speed improvements in HMMER3 allowed the routine use of the forward-backward algorithm that increases sensitivity by scoring the sequence similarity based on the sum of all possible alignments. As query and target sequences become more divergent, there can be multiple ways of aligning them. HMMER3 is able to calculate the sum over all possible alignments, rather than only considering an optimal alignment (used in the HMMER Viterbi method and in BLAST). As well as speeding up the releases for Pfam, the increase in speed allowed interactive searches, as implemented in the HMMER Web server (*http://hmmer.org*). Spreading searches across around multiple CPUs (80 in the case of the HMMER Web server) allows nearly immediate results for the average query against a large sequence collection such as UniProtKB.

The HMMER Web server offers search against a comprehensive collection of protein sequences, which is represented by database such as UniProtKB. The current version of UniProtKB (release 2017_08) has 555,426 protein sequences in Swiss-Prot and 89,396,316 protein sequences in TrEMBL. The sequence searches can also be limited to a subset of UniProtKB database such as Swiss-Prot, which comprises high-quality manually curated protein entries, or Reference Proteome, which provides a good coverage of sequence space while reducing the size of the search database at the same time. The HMMER Web server also allows searches of sequences from Ensembl database. The current version of Ensembl (release 90) comprises sequences and annotations for 114 species. Other specialized target databases such Pfam, TIGRFAM, GENE3D, Superfamily, and PIRSF, which comprise hidden Markov models, can be used to annotate the input query sequence.

The search protocol for HMMSEARCH is not discussed in this article, since the search method is similar to PHMMER, but only differs in the input format. While a protein sequence is used as query in PHMMER, the input for HMMSEARCH is a multiple sequence alignment or a profile HMM. The target search database for both algorithms is the same, i.e., one or more protein sequence databases.

Unlike the output from various BLAST servers, the HMMER output is organized into separate logical sections that include pairwise sequence alignments, taxonomic distribution of resultant hits, and their ordering based on similar domain architectures. This format makes it easier to navigate, interpret, filter, and download results. In addition, the resultant sequence hits are linked to external databases, which provide annotations on structure, functions and interactions.

Apart from using the Web service, the full suite of sequence search algorithms available on the HMMER Web server can be implemented locally using the stand-alone HMMER software package. The source code and binaries of the current version HMMER 3.1b2 (released 5th March 2015) for Linux, MacOSX, and Windows machines can be downloaded from *http://hmmer.org/download.html*. The complete documentation for locally installing the HMMER package and the user's guide to it can be found at *http://hmmer.org/documentation.html*. The HMMER Web server can also be accessed through the API. API scripts allow queries to be sent and results, including those from the domain architecture and taxonomy tabs, to be retrieved programmatically.

In conclusion, the protocols discussed in this article provide an overview of how to access HMMER suite of programs, using the HMMER Web server interface, to identify close and remote protein homologs and to navigate the output to help in protein function annotation.

Literature Cited

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*, 3389–3402. doi: 10.1093/nar/25.17.3389.

Coggill, P., Finn, R. D., & Bateman, A. (2008). Identifying protein domains with the Pfam database. *Current Protocols in Bioinformatics*, *23*, 2.5:2.5.1–2.5.17. doi: 10.1002/0471250953.bi0205s23.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*, W29–37. doi: 10.1093/nar/gkr367.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., ... Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*, D279–285. doi: 10.1093/nar/gkv1344.

Gibney, G., & Baxevanis, A. D. (2011). Searching NCBI databases using Entrez. *Current Protocols in Bioinformatics*, *34*, 1.3:1.3.1–1.3.25. doi: 10.1002/0471250953.bi0103s34.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Research*, *41*, D387–395. doi: 10.1093/nar/gks1234.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*, 10915–10919. doi: 10.1073/pnas.89.22.10915.

Mills, L. (2014). Common file formats. *Current Protocols in Bioinformatics*, *1*, 1B:A.1B.1–A.1B.18. doi: 10.1002/0471250953.bia01bs45.

Pearson, W. R. (2013a). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, *42*, 3.1:3.1.1–3.1.8. doi: 10.1002/0471250953.bi0301s42.

Pearson, W. R. (2013b). Selecting the right similarity-scoring matrix. *Current Protocols in Bioinformatics*, *3*, 3.5:3.5.1–3.5.9.

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, *85*, 2444–2448. doi: 10.1073/pnas.85.8.2444.

Pundir, S., Magrane, M., Martin, M. J., O'Donovan, C., & The UniProt Consortium. (2015). Searching and navigating UniProt databases. *Current Protocols in Bioinformatics*, *50*, 1.27.1-1.27.10. doi: 10.1002/0471250953.bi0127s50.

Schuster-Böckler, B., & Bateman, A. (2007). An introduction to hidden Markov models. *Current Protocols in Bioinformatics*, *18*, A.3A.1–A.3A.9. doi: 10.1002/0471250953.bia03as18.

Sillitoe, I., Lewis, T., & Orengo, C. (2015). Using CATH-Gene3D to analyze the sequence, structure, and function of proteins. *Current Protocols in Bioinformatics*, *50*, 1.28.1-1.28.21. doi: 10.1002/0471250953.bi0128s50.

Wheeler, T. J., Clements, J., & Finn, R. D. (2014). Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, *15*, 7. doi: 10.1186/1471-2105-15-7.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., ... Gough, J. (2009). SUPERFAMILY—Sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research*, *37*, D380–386. doi: 10.1093/nar/gkn762.

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L. S., Natale, D. A., Vinayaka, C. R., ... Barker, W. C. (2004). PIRSF: Family classification system at the Protein Information Resource. *Nucleic Acids Research*, *32*, D112–114. doi: 10.1093/nar/gkh097.