

Module 3: Structure Based Evidence

Objective

The objectives of this module are:

1. To determine if the protein you are annotating is functionally similar to other known proteins, or has domains of known function, using TIGRFAM, PFAM and PDB applications.
2. To document your search results in the Structure Based Evidence Module lab notebook.

Materials

To perform this exercise you will need:

- Access to the internet on a computer equipped with the most recent version of Firefox (preferred), Chrome or Safari.
- To have completed the sign up for GENI-ACT described in the Signing Up for GENI-ACT section of the manual.

Background

This module is an extension of the sequence-based similarity module, where sequence homology was looked at to determine relatedness of proteins. One limit to the solely sequenced based approach in determining function of a protein is that proteins can share some regions of sequence homology and yet have widely varied function. In module 3, the relatedness of sequence to functional domains or structures will be investigated. Three tools will be used for this purpose: TIGRFAM, PFAM and PDB. **Note that this module is being done “out of the order” of the module list on GENI-ACT.** This is because it is more related to Module 2 and we feel it makes more sense to perform this module immediately after Module 2 has been completed.

TIGRFAM and Pfam tools are based on Hidden Markov Modeling (HMM). A Hidden Markov Model is a probabilistic model developed from observed sequences of proteins of a known function. The profile HMM is used to score the alignment of the amino acid sequence entered to other proteins based on amino acid identity and position. Proteins, or domains of proteins, of known function are aligned to create the profile HMM, against which sequences of proteins with unknown function are compared. The software then predicts whether amino acid sequence of the protein of unknown function matches that of the profile

HMM. If it does, then it is assumed that the function of the protein, or protein domain, matching the profile HMM will have the same function as defined by the profile HMM.

TIGRFAM

TIGRFAM is a manually curated database (meaning that the database is constructed based on some sort of supporting evidence) of protein families known to have similar functions. Each TIGRFAM model is assigned to a category which describes the type of functional relationship the proteins in the model have to each other. The models have the following hierarchy:

- **equivalog** - one specific function, e.g. “ribokinase”
- **subfamily** - group of related functions generally with different substrate specificities, e.g. “carbohydrate kinase”
- **superfamily** - many different functions that are related in a very general way, e.g. “kinase”
- **domain** – not necessarily full-length of the protein, contains one functional part or structural feature of a protein, may be fairly specific or may be very general, e.g. “ATP-binding domain”

When an amino acid sequence is searched against the TIGRFAM database and a good hit is found, there is likely to be a functional relationship between the query and the hit in the database.

Pfam

The Pfam database contains information about protein domains and families. Pfam-A is the manually curated portion of the database that contains over 10,000 entries. For each entry a protein sequence alignment and a hidden Markov model is stored. Because the entries in Pfam-A do not cover all known proteins, an automatically generated supplement is provided called Pfam-B. Pfam-B contains a large number of small families derived from clusters produced by an algorithm. Although of lower quality, Pfam-B families can be useful when no Pfam-A families are found (<http://en.wikipedia.org/wiki/Pfam>). The following nomenclature is used in describing PFAM results:

- **Domain:** A structural unit which can be found in multiple protein contexts.
 - e.g., zinc finger, leucine zipper
- **Family:** A collection of related proteins containing the same domain.
 - e.g., immunoglobulins, CD4, MHC, TCR, etc.
- **Clan:** A collection of multiple protein families. The relationship may be defined by similarity of sequence, structure, or profile-HMM.
 - e.g., ATPase functioning in ETC vs. ATPase functioning in DNA replication.

Protein Data Bank (PDB)

The Protein Data Bank (PDB) is a repository for the three-dimensional structural data of large biological molecules. It is, by definition, a curated databank that has information from researchers who have experimentally determined 3 dimensional structures of proteins in the databank. These researchers will also often provide evidence for the function a particular structural domain has in a protein. You will look for matches to your protein sequence in this databank. If you find a match you are likely to find significant information about the function of your protein.

Procedures

Standard Operating Procedure: Structure-based Evidence Module

TIGRFAM:

1. Go to <http://blast.jcvi.org/web-hmm/>.
2. Select TIGRFAMS in the database pull down menu, leave the scope set to GLOBAL and change the E-value cutoff limit to '0.01' from pull down menu as shown in Figure 3.1. The E-value cutoff limit may be changed depending on how well the sequence is conserved.

Searching a sequence against protein family based HMMs

This page supports searches of protein sequence against a database of hidden Markov models (HMMs) based upon protein families. The databases are described [here](#). The default GLOBAL search looks for matches of the full length model against the query sequence. If query sequences are potentially fragments or partial length, also try a FRAGMENT search.

Database: Scope: E-value cutoff level:

Upload a file containing a sequence OR paste it into the textbox:
(Note: If both are entered, the file will be ignored.)

Enter the name of the file containing a protein sequence in [FASTA](#) or raw format:
 No file selected.

Enter your protein sequence in [FASTA](#) or raw sequence format:

Sequence identifier (for definition line if raw sequence entered):

Figure 3.1. The TIGRFAM search entry page with parameters set as described in the text.

3. Open your GENI-ACT basic information module, copy the FASTA formatted amino acid sequence of the protein encoded by your gene and paste it into the search text box of the TIGRFAM entry page (Figure 3.2).
4. Click on the Start HMM search (arrow, Figure 3.2).

Database: TIGRFAMS Scope: GLOBAL E-value cutoff level: 0.01

Upload a file containing a sequence OR paste it into the textbox:
 (Note: If both are entered, the file will be ignored.)

Enter the name of the file containing a protein sequence in [FASTA](#) or raw format:
 No file selected.

Enter your protein sequence in [FASTA](#) or raw sequence format:

```
> Ksed_00010 amino acid sequence
VSQTPDDHATAIWQEMVHLOGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDVKDAVEGHLR
EDVSTALAEVLRDRIRLAVSVDPDAVSAAQEEAAPAPSPADEDDPATGEGPLSTAVDGAVEKH
EGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDFVLGSSNRFAHAATAVAEA
PARAYNPLFIYGGSGLGKTHLLHAIGHYARTLSSVVRVKYVNSEEFTNQFINAVSAGQANAFOR
QYRDVVDVLLDDIQFLQCKEQTMEEFPHFTNTHNSEKQIVITSDQPPKLSGFAERMRSFEW
GLLTDVQPPDLETRIALRRKAAADKLDIPDDVLHLIASKISSNIRELEGALTRVTAFASLSGS
PLDEYLARTVLKDVMPGGDSGQITPTMILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLGR
ELTDLSLPXIKGKFGGRDHTVMHAERKIKOLLGEDARVYDEVSELTSLIRKKAARGRX
```

Sequence identifier (for definition line if raw sequence entered):

The email option allows the user to be notified by email when the hmmpfam search is completed. The email will contain a link to the results. Use this option when running a lengthy search. To activate this option, check the box to the left and fill in an email address in the box to the right.

Send a link to results to email address:

Figure 3.2. The TIGRFAM start page with the Ksed_00010 amino acid sequence entered. The start HMM search button is indicated by the arrow.

5. The TIGRFAM results page will look something like that shown in Figure 3.3. Ksed_00010 has two TIGRFAM hits. Record the TIGRFAM name, number, score and E-value from the results obtained in your GENI-ACT notebook.
 - 5.1. Only record results with a positive score and an E-value < 10⁻³.

- 5.2. After searching the TIGRFAM database, raw text results will show which TIGRFAMs match. The name of the TIGRFAM ('Description' column) may be cut off (see caption to figure 3.3). To find the entire name, identify the TIGRFAM number (e.g. TIGR by the code found in the 'Model' column).

J. Craig Venter™
 I N S T I T U T E

```

hmmpfam - search one or more sequences against HMM database
HMMER 2.3.2 (Oct 2003)
Copyright (C) 1992-2003 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
-----
HMM file:          TOGA_LIB_bin.HMM
Sequence file:     hmmpfam-search-15283-1404415246.in
-----

Query sequence: Ksed_00010
Accession:      [none]
Description:    amino acid sequence

Scores for sequence family classification (score includes all domains):
Model      Description                               Score      E-value     N
-----
TIGR00362  DnaA: chromosomal replication initiator pro  740.9      3.8e-220    1
TIGR03420  DnaA_homol_Hda: DnaA regulatory inactivator  13.9       1.2e-13     1

Parsed for domains:
Model      Domain  seq-f  seq-t  hmm-f  hmm-t  score  E-value
-----
TIGR00362  1/1     13    498  ..    1    449  []    740.9  3.8e-220
TIGR03420  1/1     163   396  ..    1    231  []    13.9   1.2e-13

Alignments of top-scoring domains:
TIGR00362: domain 1 of 1, from 13 to 498: score 740.9, E = 3.8e-220
      *->WqrvlerLek.elseqefntWikipkllsiegnttlilsvPneFvkd
      Wq++  +L+  1  +++  ++  ++l+++  ++  t++l+v  vkd
Ksed_00010  13      WQEAMVHLQGaGLAPRDIG-VLRLATLVGLLEG-TALLAVKYDHSV 57

      wienknydlIeellqeltgeeieieftvgdsetpapaelelepask.pep
      +e  ++++  +  ++l+e+  +++i++++  v      p+  +  +++  +  p++
Ksed_00010  58  AVEGHLREDVSTALAEVLDLDRDIRLAVSVD-----PDAVSAAQEEAaPPA 101
  
```

←Figure 3.3. TIGRFAM results for Ksed_00010. Two hits are seen with significant E values, but the 1st hit, TIGR00362 has a much higher score and lower E value than the 2nd hit. Note the name of TIGR00362 is truncated (column labeled Description). See the text for the way to get the full name.

5.3. To find the full description/name of the TIGRFAM, navigate to: <http://www.jevl.org/cgi-bin/tigrfams/Listing.cgi> and scroll down the list to find the complete TIGRFAM name and the category term (Figure 3.4).

	common domain		
TIGR00351	narI	respiratory nitrate reductase, gamma subunit	equivalog 1.7.99.4
TIGR00353	nrfE	cytochrome c-type biogenesis protein CcmF	equivalog
TIGR00354	polC	DNA polymerase II, large subunit DP2	equivalog 2.7.7.7
TIGR00355	purH	phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase	equivalog 2.1.2.3 3.5.4.10
TIGR00357	TIGR00357	methionine-R-sulfoxide reductase	equivalog_domain 1.8.4.-
TIGR00358	3_prime_RNase	VacB and RNase II family 3'-5' exoribonucleases	superfamily 3.1.13.1
TIGR00359	cello_pts_IIC	PTS system, cellobiose-specific IIC component	equivalog 2.7.1.69
TIGR00360	ComEC_N-term	ComEC/Rec2-related protein	subfamily_domain
TIGR00361	ComEC_Rec2	DNA internalization-related competence protein ComEC/Rec2	equivalog
TIGR00362	DnaA	chromosomal replication initiator protein DnaA	equivalog ←
TIGR00363	TIGR00363	lipoprotein, YaeC family	subfamily
TIGR00364	TIGR00364	queuosine biosynthesis protein QueC	equivalog
TIGR00365	TIGR00365	monothiol glutaredoxin, Grx4 family	equivalog
TIGR00366	TIGR00366	TIGR00366 family protein	hypoth_equivalog
TIGR00367	TIGR00367	K+-dependent Na+/Ca+ exchanger homolog	hypoth_equivalog
TIGR00368	TIGR00368	Mg chelatase-like protein	hypoth_equivalog
TIGR00369	unchar_dom_1	uncharacterized domain 1	domain
TIGR00370	TIGR00370	cancer histidine kinase inhibitor KinI family	hypoth_equivalog

Figure 3.4. The TIGRFAM listing page. The arrow points to the TIGRFAM with the truncated name described in figure 3.3. Note that it is classified as an equivalog. Clicking on the hyperlinked TIGRFAM number will open a more detailed description page as shown in figure 3.5.

5.4. Clicking on the hyperlink for the TIGRFAM number will open a more complete description of the HMM that will over insight about the function of your protein (Figure 3.5).

<p>→ TIGRFAMs Home</p> <p>→ TIGRFAMs Terms</p> <p>→ TIGRFAMs Complete Listing</p> <p>→ TIGRFAMs FTP site</p> <p>→ TIGRFAMs Resources</p> <p>→ TIGR00362 Seed Alignment</p>		<h3>HMM SUMMARY PAGE: TIGR00362</h3>	
Accession		TIGR00362	
Name		DnaA	
Function		chromosomal replication initiator protein DnaA	
Gene Symbol		dnaA	
Trusted Cutoff		343.90	
Domain Trusted Cutoff		343.90	
Noise Cutoff		184.65	
Domain Noise Cutoff		184.65	
Isology Type		equivalog	
HMM Length		437	
Mainrole Category		DNA metabolism	
Subrole Category		DNA replication, recombination, and repair	
Gene Ontology Term		GO:0003677 : DNA binding molecular_function GO:0003688 : DNA replication origin binding molecular_function GO:0005524 : ATP binding molecular_function GO:0006270 : DNA-dependent DNA replication initiation biological_process GO:0006275 : regulation of DNA replication biological_process	
Author		Loftus BJ, Haft DH	
Entry Date		Apr 20 1999 2:03PM	
Last Modified		Feb 14 2011 3:27PM	
Comment		DnaA is involved in DNA biosynthesis; initiation of chromosome replication and can also be transcription regulator. The C-terminal of the family hits the pfam bacterial DnaA (bac_dnaA) domain family. For a review, see Kaguni (2006). RN [1] RM PMID:16753031 RT DnaA: controlling the initiation of bacterial DNA replication and more. RA Kaguni JM RL Annu Rev Microbiol. 2006;60:351-75. DR PFAM; PF00308; bac_dnaA; DR PROSITE; PDOC00771; DR ECOCYC; EG10235; dnaA; DR SWISSPROT; P03004; SE TIGR DR HAMAP; MF_00377; 410 of 419 GenProp0799: bacterial core gene set, exactly 1 per genome (HMM) GenProp0806: replication initiation, bacterial (HMM)	
References			
Genome Property			

Figure 3.5. The detailed description page for TIGR00362

- 5.5. Record the information requested for any significant TIGRFAM hits you find in your GEN-ACT notebook as shown in figure 3.6.

TIGRFAM
go to TIGRFAM at http://tigrblast.tigr.org/web-hmm
TIGRFAM number 
TIGR00362
TIGRFAM name 
chromosomal replication initiator protein DnaA
Score 
740.9
E-value 
3.8e-220

Figure 3.6. The TIGRFAM geni-act notebook page with the information from the top TIGRFAM hit (TIGR00362) entered. Note the name is DnaA, but additional functional information has been added to remind the annotator that DnaA acts as a chromosomal replication initiator protein.

Protein Family (Pfam):

1. Go to <http://xfam.org/> and select the Pfam option as shown by the arrow in figure 3.7.

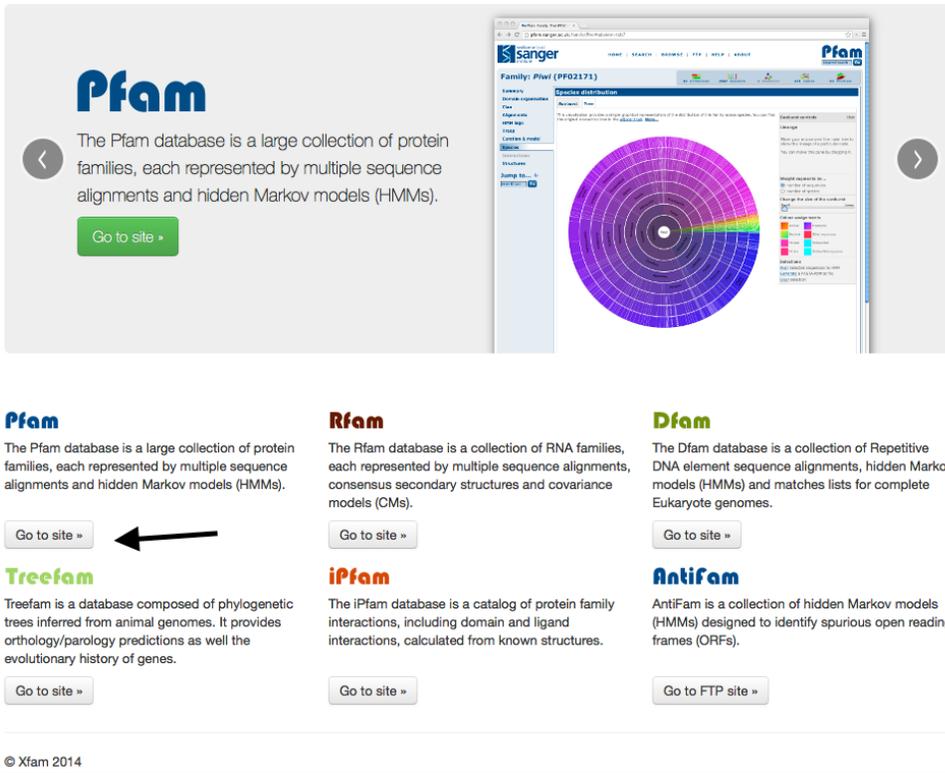


Figure 3.7. The Pfam entry page. Clicking the button indicated by the arrow opens the Pfam site.

2. Click the “Search” tab at the top of the page as shown in figure 3.8 to access the main Pfam search page.

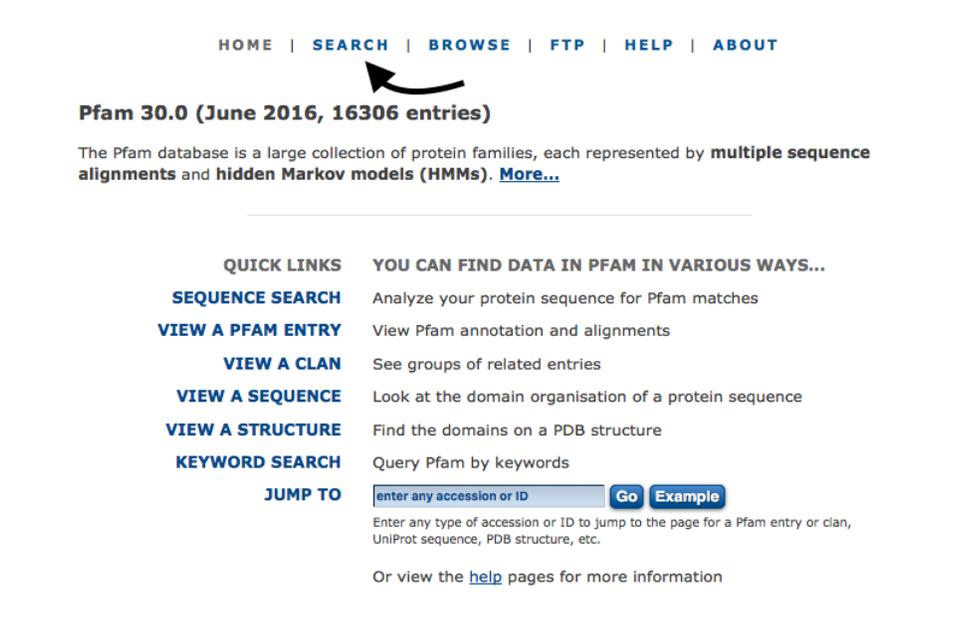


Figure 3.8. The Pfam search entry page. Click on the search tab as indicated by the arrow to begin the search.

3. Click on the “Sequence” option and paste the FASTA formatted sequence of the protein under investigation in the sequence text box as shown in Figure 3.9.
 - 3.I. Change the E-value to 0.001 and click submit.

EMBL-EBI  HOME | SEARCH | BROWSE | FTP | HELP | ABOUT  keyword search Go

Search Pfam

0 architectures 0 sequences 0 interactions 0 species 0 structures

Sequence search

Find Pfam families within your sequence of interest. Paste your **protein** or **DNA** sequence into the box below to have it searched for matching Pfam families. [More...](#)

Sequence

```
> Ksed_00010 amino acid sequence
VSQTPDDHATAIWQAMVHLQAGLAPRDIGVLRRLATLVGLLEGTTALLAVKYDHYKDAVEGHLR
EDVSTALAEVLDRLAVSVPDAVSAAEAAAPPAPSPADEDDPATGEGPLSTAVDCAVEKH
EGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYYFTDFVLGSSNRFHAAATAVAEA
PARAYNPLFIYGGGLGKTHLLHAIHYARTLDSSRVKYNSEFFTNQFINAVSAGQANAFQR
QYRDVVDLLIDDIQFLGQKEQTMEEFFHTNTLHNSKQIVTSDQPKKLSGFAERMRSRFEW
GLLTDVQPPDLETRIALRRKAAADKLDIPDDVLHLIASKISSNIREGALTRVTFASLSGS
PLDEYLARTVLKDVMPGGDSGQITPTMILEETAGYFVSVVEIQGASRSRNLTRARQIAMYLCR
ELTDLSPKIGKEFGGRDHTTVMHAERKIKQLLGEDRRVYDEVSELSIIRKKAARGRX
```

Note: we cannot guess the type of this sequence based on the alphabet. The controls below will remain enabled but the values will be ignored by the server if your sequence is DNA.

Protein sequence options

Cut-off Gathering threshold Use E-value

E-value

Search for PfamBs **Note** that we search only the 20,000 largest Pfam-B families

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Figure 3.9. The Pfam search start page with the amino acid sequence of Ksed_00010 pasted into the text window.

4. A results window will appear similar to the one shown in figure 3.10 if you have significant Pfam-A hits. If you do not get any hits return to the PFAM start page, click the Search for Pfam-Bs check box and repeat the search.

EMBL-EBI  HOME | SEARCH | BROWSE | FTP | HELP | ABOUT  keyword search Go

Sequence search results

[Show](#) the detailed description of this results page.

We found **2** Pfam-A matches to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.



[Show](#) the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Bac_DnaA	Bacterial dnaA protein	Family	CL0023	164	382	164	381	1	218	219	326.0	1.1e-97	n/a	Show
Bac_DnaA_C	Bacterial dnaA protein helix-turn-helix	Domain	CL0123	408	477	409	477	2	70	70	104.5	1.8e-30	n/a	Show

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Figure 3.10. Pfam results for Ksed_00010. See the text for an explanation of the results.

5. Interpretation of Pfam results (taken from <http://pid.nci.nih.gov/2011/110913/full/pid.2011.3.shtml> - f6).
- 5.1. The columns labeled family (domain), description, entry type and clan give the specifics of these for each of the Pfam hits. The predicted active site column will indicate if an active enzyme site is included in the HMM. The GENI-ACT notebook will have entries for a number of items that follow (Figure 3.11) and you should enter information in the notebook as you encounter it. However, there is much additional information important for determining the function of your protein than is requested in your online notebook, and it will likely to record additional findings in a notebook you keep for yourself.

The image shows a screenshot of the Pfam search interface. It contains several input fields with labels and a small square icon to the right of each label. The labels are: 'Pfam number (PF####) for top hit', 'Pfam name', 'Clan name', 'Clan number (CL####)', 'Score', 'E-value', 'Pairwise alignment', 'HMM logo', and 'Key functional/structural residues (e.g. I2, W7, F13)'. Each label is followed by a horizontal input field.

← Figure 3.11.
The GENI-ACT
notebook for the
Structure Based
Evidence Module

- 5.2. Scores: Using the search parameters described above will result in matches being reported that have an E-value less than or equal to 10^{-3} (or whatever threshold you have set on the search page).
- 5.3. Alignment and envelope coordinates: Each sequence match to a Pfam HMM will have two sets of coordinates: the alignment coordinates and the envelope coordinates. The envelope coordinates indicate the region on the sequence over which the match lies, whereas the alignment coordinates indicate the region over which the alignment confidence is high.
- 5.4. Graphically, the alignment coordinates are depicted with a solid color and the envelope coordinates in a lighter shade of the same color. When the region within the envelope coordinates does not match the entire length of a HMM, the match is said to be partial; graphically, this is drawn with a jagged edge at the N or C terminal or both, depending on which region of the match is incomplete.
- 5.5. The option to display the residue-by-residue scores is also available via the show/hide button (Figure 3.12). When the alignment is 'shown', the #HMM line shows the consensus amino acid sequence of the model, with capital letters representing the most conserved (high information content) positions, and dots (.) indicating insertions in the query sequence with respect to the model. Identical residues are colored cyan, and similar residues are colored dark blue; the #MATCH line indicates matches between the model and the query sequence, where a + indicates positive score, interpretable as "conservative substitution" with respect to what the model expects at

that position; the #PP line represents the posterior probability (essentially the expected accuracy) of each aligned residue, where a 0 means 0–5%, 1 means 5–15%, and so on to 9 meaning 85–95% and a * meaning 95–100% posterior probability (pp); the #SEQ line is the query sequence, colored according to the pp for each residue match on a scale from bright green for * through paler green and pale red down to bright red for 0.

The screenshot shows the Pfam search results page. At the top, there are navigation links: HOME | SEARCH | BROWSE | FTP | HELP | ABOUT. The Pfam logo is in the top right corner. Below the navigation, there is a section for "Sequence search results" with links to "Show the detailed description of this results page" and "We found 2 Pfam-A matches to your search sequence (all significant). You did not choose to search for Pfam-B matches." Below this, there is a "Show the search options and sequence that you submitted" link and a "Return to the search form to look for Pfam domains on a new sequence" link. The main section is "Significant Pfam-A Matches" with a "Show or hide all alignments" link. A table lists the matches:

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Bac_DnaA	Bacterial dnaA protein	Family	CL0022	164	382	164	381	1	218	219	326.0	1.1e-97	n/a	Hide
Bac_DnaA_C	Bacterial dnaA protein helix-turn-helix	Domain	CL0123	408	477	409	477	2	70	70	104.5	1.8e-30	n/a	Show

Below the table, there is a "Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk. European Molecular Biology Laboratory" link.

Figure 3.12. The Pfam results page with the show alignment option selected. See text section 3.5 above for an explanation.

- 5.6. Clicking on the hyperlink for the top hit Family name will open a family page similar to the one shown in Figure 3.13 for the top Pfam hit for Ksed_00010. The family page for a Pfam-A family contains the functional annotation at the top of the page, derived either directly from the Wikipedia entry for that family if one is available or from Pfam or InterPro. At the side of the page are a number of tabs, each relating to a different set of data, some of which we will discuss below. Read any text written on this page carefully. Since each Pfam domain has been manually curated, this information can be extremely useful in predicting the function of the query gene that contains the domain. If ever a GO or EC number is given, record that number in the Lab Notebook as it will aid in predicting the function of the gene product.

Family: Bac_DnaA (PF00308)

14 architectures 6279 sequences 2 interactions 4628 species 18 structures

Summary: Bacterial dnaA protein

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: DnaA Pfam InterPro

This is the Wikipedia entry entitled "DnaA". [More...](#)

DnaA [Edit Wikipedia article](#)

DnaA is a protein that activates initiation of DNA replication in prokaryotes.^[1] It is a replication initiation factor which promotes the unwinding of DNA at *oriC*.^[1] The onset of the initiation phase of DNA replication is determined by the concentration of DnaA.^[1] DnaA accumulates during growth and then triggers the initiation of replication.^[1] Replication begins with active DnaA binding to 9-mer (9-bp) repeats upstream of *oriC*.^[1] Binding of DnaA leads to strand separation at the 13-mer repeats.^[1] This binding causes the DNA to loop in preparation for melting open by the helicase DnaB.^[1]

Contents [\[hide\]](#)

- 1 Function
- 2 References
- 3 Further reading
- 4 External links

Function [\[edit\]](#)

The active form DnaA is bound to ATP.^[1] Immediately after a cell has divided, the level of active DnaA within the cell is low.^[1] Although the active form of DnaA requires ATP, the formation of the *oriC*/DnaA complex and subsequent DNA unwinding does not require ATP hydrolysis.^[2]

The *oriC* site in *E. coli* has three AT rich 13 base pair regions (DUEs) followed by four 9 bp regions.^[3] Around 10 DnaA molecules bind to the 9 bp regions, which wrap around the proteins causing the DNA at the AT-rich region to unwind. There are 8 DnaA binding sites within *oriC*, to which DnaA binds with differential affinity. When DNA replication is about to commence, DnaA occupies all of the high and low affinity binding sites. The denatured AT-rich region allows for the recruitment of DnaB (helicase), which complexes with DnaC (helicase loader). DnaC helps the helicase to bind to and to properly accommodate the ssDNA at the 13 bp region; this is accomplished by ATP hydrolysis, after which DnaC is released. Single-strand binding proteins (SSBs) stabilize the single DNA strands in order to maintain the replication bubble. DnaB is a 5'→3' helicase, so it travels on the lagging strand. It associates with DnaG (a primase) to form the only primer for the leading strand and to add RNA primers on the lagging strand. The interaction between DnaG and DnaB is necessary to control the longitude of Okazaki fragments on the lagging strand. DNA polymerase III is then able to start DNA replication.

DnaA contains two conserved regions: the first is located in the central part of the protein and corresponds to the ATP-binding domain, the second is located in the C-terminal half and is involved in DNA-binding.^[4]

References [\[edit\]](#)

- ¹ ^a ^b ^c ^d ^e ^f ^g ^h ⁱ Foster JB, Slonczewski J (2009). Microbiology: an evolving science. New York: W.W. Norton & Co. ISBN 0-393-97857-5.
- ² ³ Leonard AC, Grimwade JE (December 2010). "Regulating DnaA complex assembly: it is time to fill the gaps" *#*. *Curr. Opin. Microbiol.* **13** (6): 766–72.

Chromosomal replication initiator protein dnaA

Identifiers

Organism	Escherichia coli (str. K-12 substr. MG1655)
Symbol	DnaA
Entrez	948217 #
RefSeq	NP_418157.1 #
(Prot)	
UniProt	P03004 #
Other data	
Chromosome	genome: 3.88 - 3.88 Mb #

Bac_DnaA_C

crystal structure of dnaa domain in complex with dnaabox dna

Identifiers

Symbol	Bac_DnaA_C
Pfam	PF08299 #
Pfam clan	CL0123 #
InterPro	IPR013159 #
SCOP	1j1v #

Figure 3.13. The Domain Summary page for Bac_DnaA.

- On the left menu of the Domain Summary page, click "HMM Logo" to access this very useful way of visualizing amino acid conservation among the sequences used to build the Pfam domain (Figure 3.14.) HMM Logos provide the researcher with a quick overview of the features of a profile HMM while conserving as much information as possible. Similar to the sequence logo generated by WebLogo earlier, on the HMM logo, the larger the letter, the more conserved this residue is in the protein family. Colors correspond to different amino acid types (e.g. neutral, acidic, etc.). Letters are sorted in descending order depending on their probability of occurring at a given position in a sequence that contains the domain. Right click on the HMM logo and choose "copy image". You should then paste the image into Paint (PC) or Preview (Mac) and save the image file as a png. You can then sections) upload it to the Lab Notebook.

Family: Bac_DnaA (PF00308)

14 architectures 6279 sequences 2 interactions 4628 species 18

HMM logo

HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you them [here](#) <#>. [More...](#)

Contribution

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Figure 3.14. The HMM logo page

Protein Data Bank (PDB)

1. Go to <http://www.rcsb.org/pdb/home/home.do> and click on Advanced to begin your search (Figure 3.16)
2. Choose the BLAST/FASTA/PSI-BLAST option from the query type pull down menu. There are a number of choices sorted heading and subheading in the pull down. You will need to scroll down to the Sequence Features heading in the list, under which you will find the BLAST/FASTA/PSI-BLAST option, as indicated by the arrow in Figure 3.17.

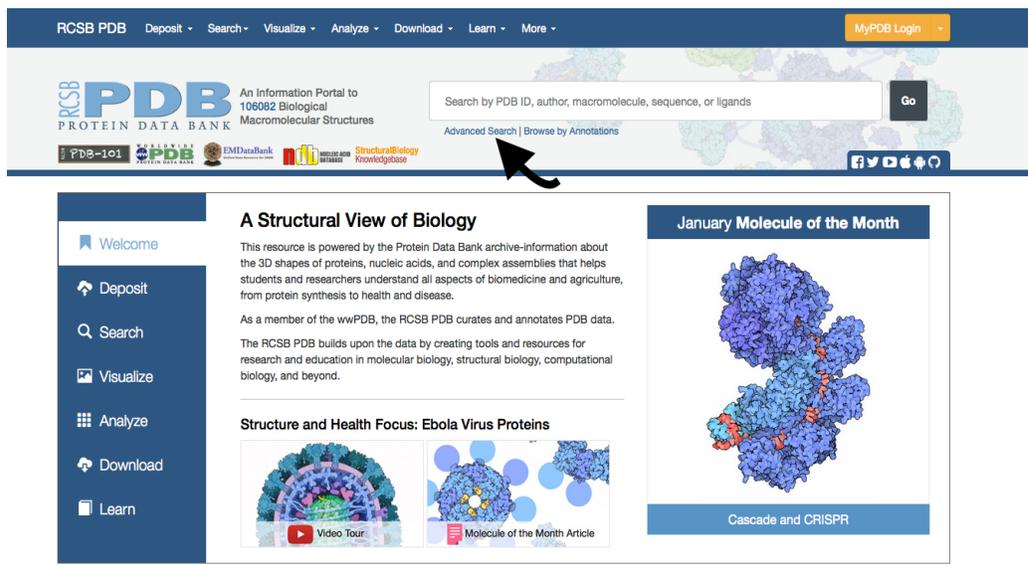


Figure 3.16. The Protein Data Bank start page. Click on Advanced (see arrow)

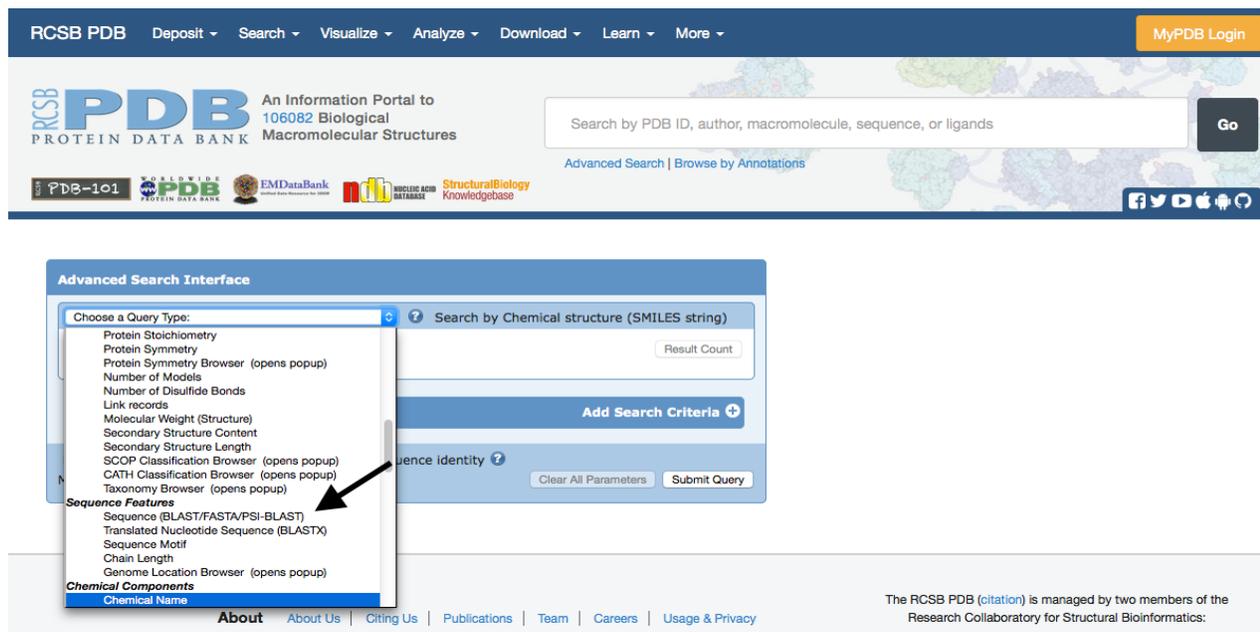


Figure 3.17. Selection of a BLAST search in PDB. Select the Choose Query Type pull down menu and select “Sequence (BLAST/FASTA/PSI-BLAST)” option

3. Paste the FASTA formatted sequence of your protein into the sequence text box. Change the "E Cut-Off Value" to 0.001 (Figure 3.18). Click the "Search" button and review the results. This runs a BLAST search just like you did in NCBI BLAST in the Sequence-based Similarity module. However, this is searching the query gene against all of the gene sequences that have solved structures in PDB (Figure 3.19).

The screenshot shows the "Advanced Search Interface" for the Protein Data Bank. The search tool is set to "BLAST". The E-Value Cutoff is manually set to "0.001". The sequence text box contains the following FASTA formatted sequence:

```
> KSED_00010 AMINO ACID SEQUENCE
VSQTPDDHATAIWQEQAMVHLOGAGLAPRDIGVLRATLVGLLEG TALLAVKYDHSV KDAVEGHLR
EDVSTALAEVLDRLIRLAVSVDPPDAVSAAQEEAAPSPADEDDPATGEGPLSTAVDGA VEKH
FCGCRADAGEQLAATSTACLTATACGCRDQLEDFVCAIILWQVETDLEILGCGAIDFLLAATATAEA
```

Other visible parameters include: Mask Low Complexity set to "Yes", and Sequence Identity Cutoff (%) set to "0". Buttons for "Add Search Criteria", "Retrieve only representatives at 90% sequence identity", "Match all of the above conditions", "Clear All Parameters", and "Submit Query" are also present.

Figure 3.18. The PDB Advanced search page. The FASTA formatted Ksed_00010 amino acid sequence has been pasted into the sequence text box and the E-Value Cutoff has manually been changed to 0.001.

4. Examine the quality of the alignments between the query gene and the BLAST hits in the Protein Data Bank. If the E-value meets the cutoff set by your instructor, and a significant length of the protein is aligned, then this is a good hit. If two proteins are very similar in sequence and have approximately the same length, it is highly probable that they fold very similarly. Therefore, the structure that corresponds to the PDB BLAST hit likely resembles how the query gene product folds.

If your query results in one or more good matches, record the PDB Code, Name, Length, Score, Alignment length, and E-value into the Lab Notebook (Figure 3.20).

The screenshot shows the 'Refinements' section of a web application. On the left, there are filters for ORGANISM, UNIPROT MOLECULE NAME, TAXONOMY, EXPERIMENTAL METHOD, X-RAY RESOLUTION, and RELEASE DATE. The main area displays search results for 'Structure of domain III from the Thermotoga maritima replication origin DnaA'. A dropdown menu is open over the 'Reports' section, with 'BLAST/FASTA/PSI-BLAST Results' highlighted and an arrow pointing to it. Other options include 'Select one...', 'Custom Reports', 'List Selected PDB IDs', 'List Selected Entity IDs', 'Customizable Table', 'Summary Reports', 'Structure', 'Sequence', 'Ligands', 'Structural Genomics Center', 'Primary Citation', 'Biological Details', 'Sequence Clusters', 'Experimental Reports', 'X-ray', 'Crystallization', 'Data Collection', 'Refinement', 'Refinement Parameters', 'Unit Cell', 'Software', 'NMR', 'Representative Model', 'Spectrometer', and 'Sample Conditions'. Below the dropdown, a 'Sequence Alignment' section shows a pairwise alignment between a query and a subject sequence.

Figure 3.22. The report section pull down menu. The arrow points to the “BLAST/FASTA ...” report that should be selected to obtain the pairwise alignment.

```
>2Z4S:1:A|pdbid|entity|chain(s)|sequence
Length = 440

Score = 225 bits (574), Expect = 1e-58, Method: Compositional matrix adjust.
Identities = 128/346 (36%), Positives = 194/346 (56%), Gaps = 3/346 (0%)

Query: 162 SALNHKYTFDFTVLGSSNRFXXXXXXXXXXXXXXXXXNPLFIYGGSLGKTHLLHAIGHYA 221
      + LN YTF+ FV+G N F YNPLFIYGG GLGKTHLL +IG+Y
Sbjct: 96 TPLNPDYTFENFVVGPGNSFAYHAALVAKHPGR-YNPLFIYGGVGLGKTHLLQSIGNV 154

Query: 222 RTLDSSVRVKYVNSEEF TNQFINAVSAGQANAFQROYRD-VDVLLIDDIQFLQKEQTE 280
      + +RV Y+ SE+F N ++++ G+ N F+ +YR VD+LLIDD+QFL GK
Sbjct: 155 VQNEPDLRVMYITSEKFLNDLVDSMKEGKLNFEFREKYRKKVDILLIDDVQFLIGKTGVQT 214

Query: 281 EFFHTFNTLHNSKQIVITSDQPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRR 340
      E FHTFN LH+S KQIVI SD+ P+KLS F +R+ SRF+ GL+ ++PPD ETR +I R+
Sbjct: 215 ELFHTFNBELHDSGKQIVICSDREPQKLEFQDRLVSRFQMLVAKLEPPDEETRKSIAK 274

Query: 341 KAADKLDIPDDVLHLIASKISSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMP 400
      + ++P++VL+ +A + N+R L GA+ ++ + +G +D A +LKD +
Sbjct: 275 MLEIEHGELPEEVLNFAENVDDNLRRLRGAI IKLLVYKETTQGEVDLKEAILLKDFIK 334

Query: 401 GGDGQITPT-MILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSPKIG 459
      + P ++E A + EEI SR+ AR+I MY+ + SL I
Sbjct: 335 PNRVKAMDPIDELIEIVAKVTGVPREEILSNRNVKALTARRIGMYVAKNYLKSRLRTIA 394

Query: 460 KEFGGRDHTVMHAERKIKQLLGEDRRVYDEVSELTSIIRKKAARG 505
      ++F V ++ LL ++++ + E+ I ++A G
Sbjct: 395 EKFNRSHPVVVDVSVKVKDSLKGNKQLKALIDEVIGEISRRALSG 440
```

Figure 3.23. The pairwise alignment of the top PDB hit for Ksed_00010 from the BLAST/FASTA/PSI-BLAST Results report.

7. If there is a literature reference that corresponds with the protein structure, it may be beneficial to read it. When a structure is published, the structural biologists frequently characterize the function of the protein and functional residues in the protein structure, and if these residues are present in your query gene, this may confirm the identity and function of your gene product.
8. The required information for this module should now be in your notebook. Between this module and the Sequence Based Similarity Module you may be able to hypothesize the name and function of your protein. However, do not be disappointed if you still cannot hypothesize a name or function for your protein, particularly if it is “hypothetical”. Other modules that follow may lead you to determine whether the gene you are working on has been called correctly.
9. There is also other information in PDB that you might find useful for the research poster you will prepare for the capstone, including literature citations as described above. You can also look at the 3D structure of the hit in PDB either as a static image or as a dynamic model that you can manipulate. Click on the 3D image of the top hit to explore options.