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 activity 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, CD_4, 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.

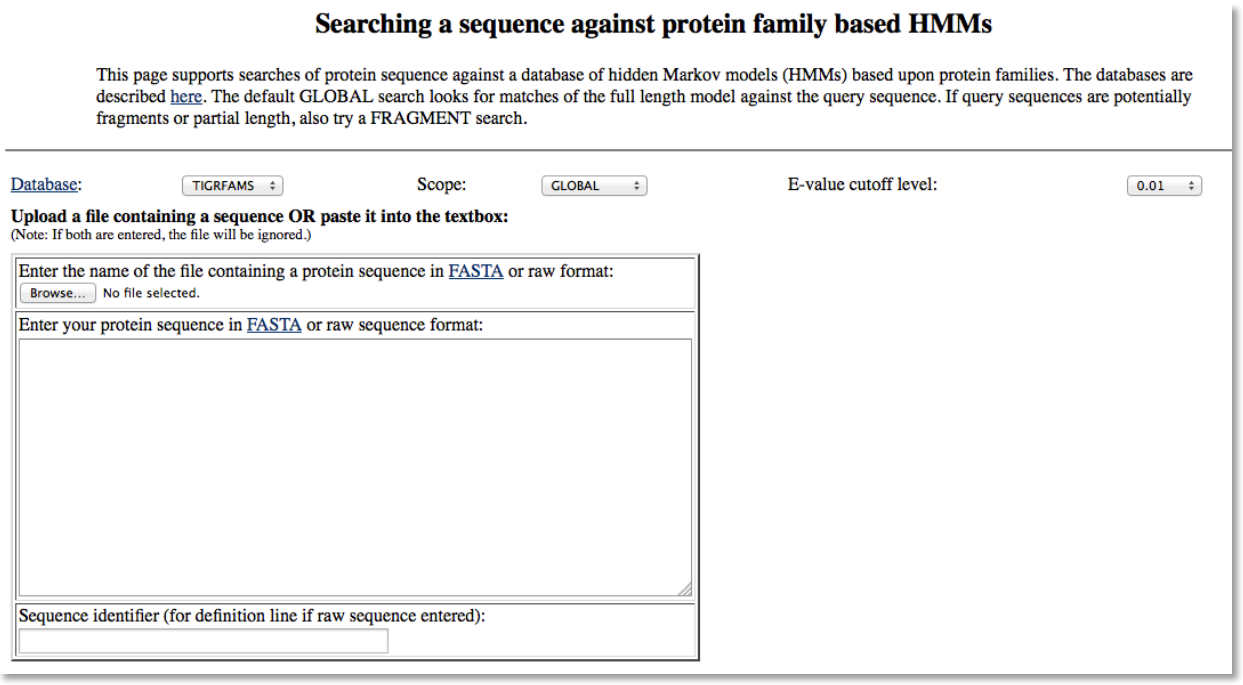


Figure 3.1. The TIGRFAM search entry page with parameters set as described in the text.

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).

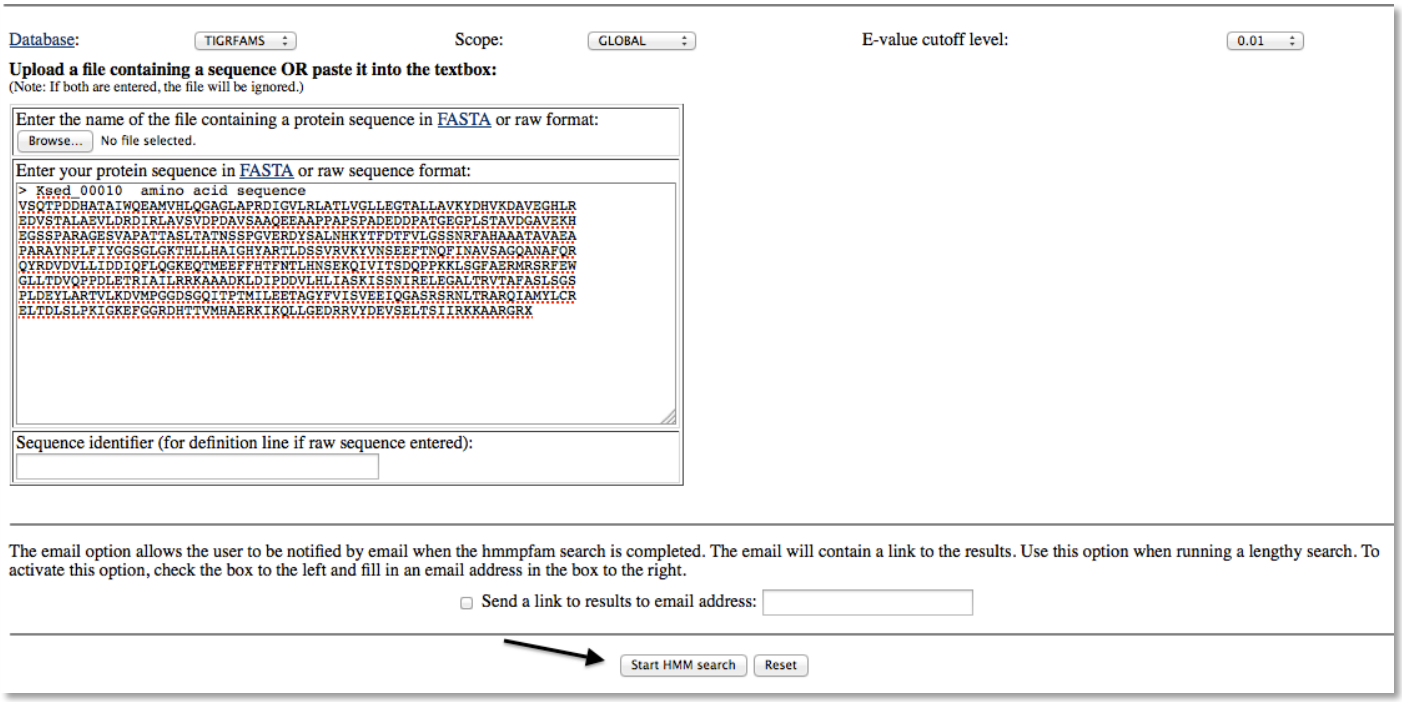


Figure 3.2. The TIGRFAM start page with the Ksed_00010 amino acid sequence entered. The start HMM search button is indicated by the arrow.

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} .**

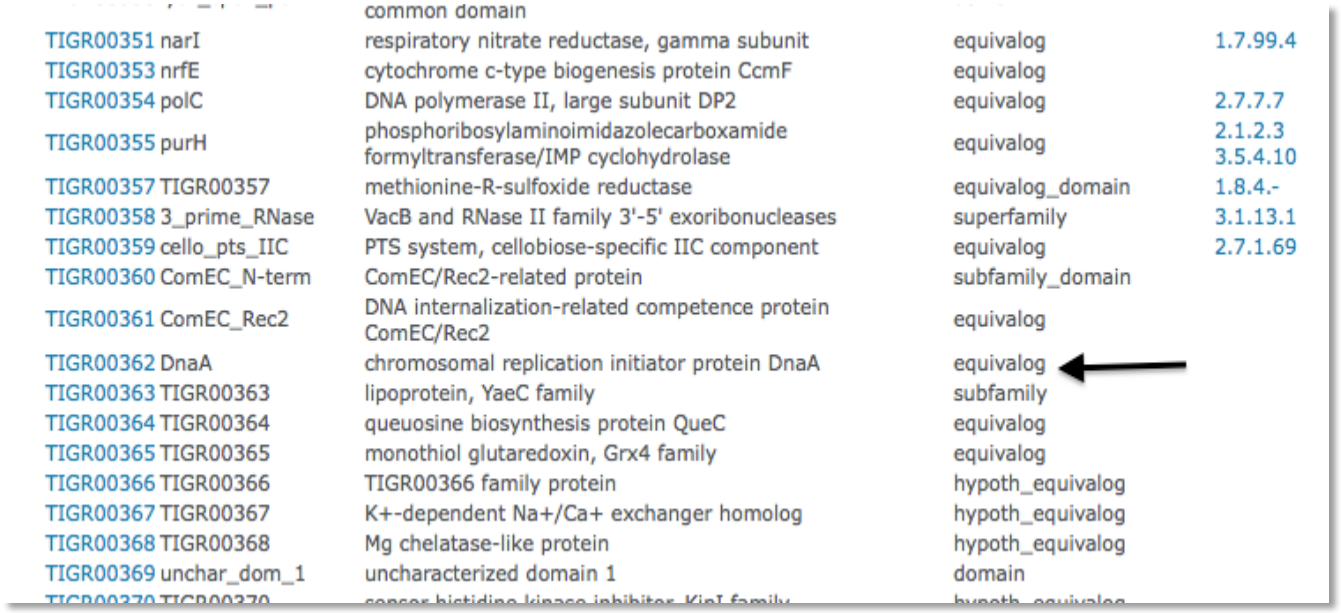
5.2. After searching the TIGRFAM database, raw text results will show which TIGRFAMs match. The name of the TIGRFAM ('Description' column) may be cut off (see caption to figure 3.3). To find the entire name, identify the TIGRFAM number (e.g. TIGR by the code found in the 'Model' column.

J. Craig Venter INSTITUTE

```
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.elseqefntWikplkllsiegnttlilsvPneFvkd
                    Wq++  +L+   l +++    ++ ++l+++ ++ t++l+v    vkd
  Ksed_00010    13    WQEAMVHLQGaGLAPRDIG-VLRLATLVGLLEG-TALLAVKYDHVKD 57

                   wieknyldlIeellqeltgeeieieftvgdsetpapaeeelepask.pep
                    +e ++++ + ++l+e+ +++i++++ v       p+  + +++ + p++
  Ksed_00010    58 AVEGHLREDVSTALAEVLDRDIRLAVSVD------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.

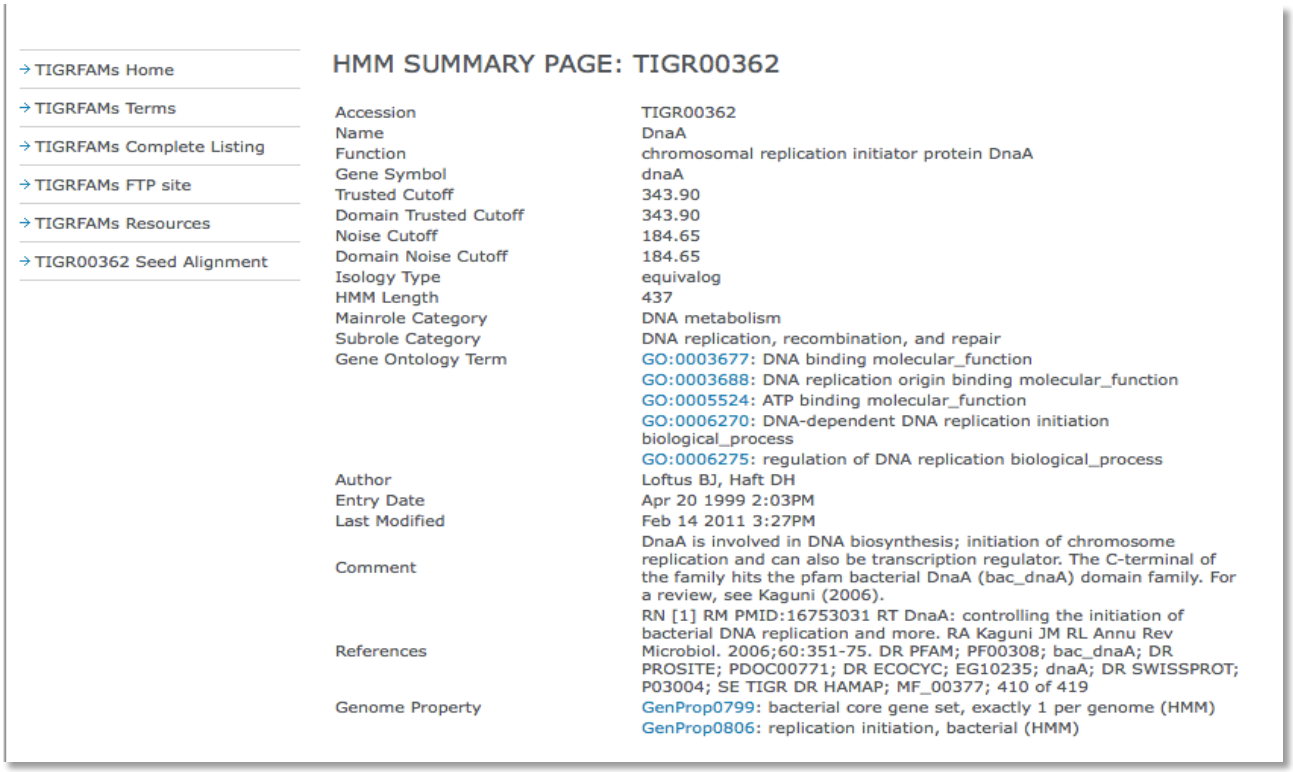
5.3. To find the full description/name of the TIGRFAM, navigate to: http://www.jcvi.org/cgi-bin/tigrfams/Listing.cgi and scroll down the list to find the complete TIGRFAM name and the category term (Figure 3.4). You can also get to the full description of all TIGFAMs by opening a new TIGRFAM search window and then first clicking on the TIGRFAMs Page tab, and then on the link for TIGRFAMS Complete Listing. You must then scroll down the list of TIGRFAMs (numbered from low to high) to find the TIGRFAM hit obtained during the search.



	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	

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).



- → TIGRFAMs Home
- → TIGRFAMs Terms
- → TIGRFAMs Complete Listing
- → TIGRFAMs FTP site
- → TIGRFAMs Resources
- → TIGR00362 Seed Alignment

HMM SUMMARY PAGE: TIGR00362

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).
References	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
Genome Property	GenProp0799: bacterial core gene set, exactly 1 per genome (HMM) GenProp0806: replication initiation, bacterial (HMM)

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. It is also useful to copy and paste any details from the comment section of the detailed TIGRFAM listing (see Figure 3.5) to give you additional background information about your protein.

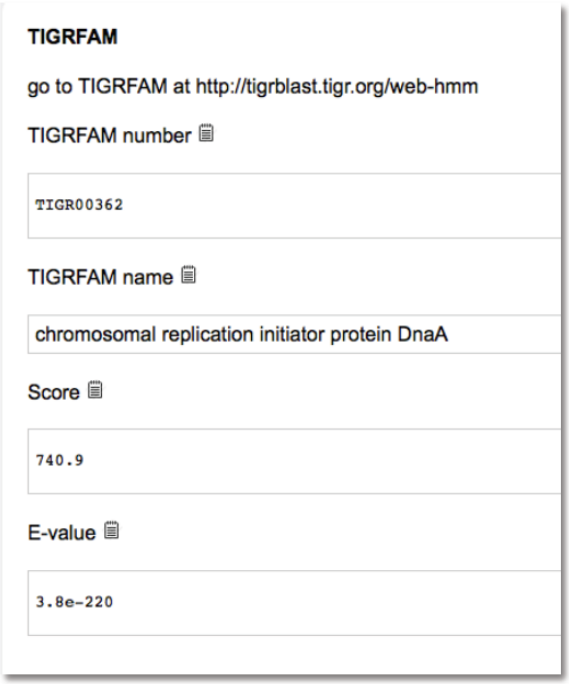


Figure 3.6. The TIGRFAM geni-act notebook page with the information from the top TIGRFAM hit (TIGR00362) entered. Note the name is DnaA, but additional functional information has been added to remind the annotator that DnaA acts as a chromosomal replication initiator protein.

Protein Family (Pfam):

1. Go to http://xfam.org/ and select the Pfam option as shown by the arrow in figure 3.7. **Note that the link in the GENI-ACT notebook my not take you directly to the xfam.org site, but there is a hyperlink on the page you will be taken to that will take you to the xfam site.**

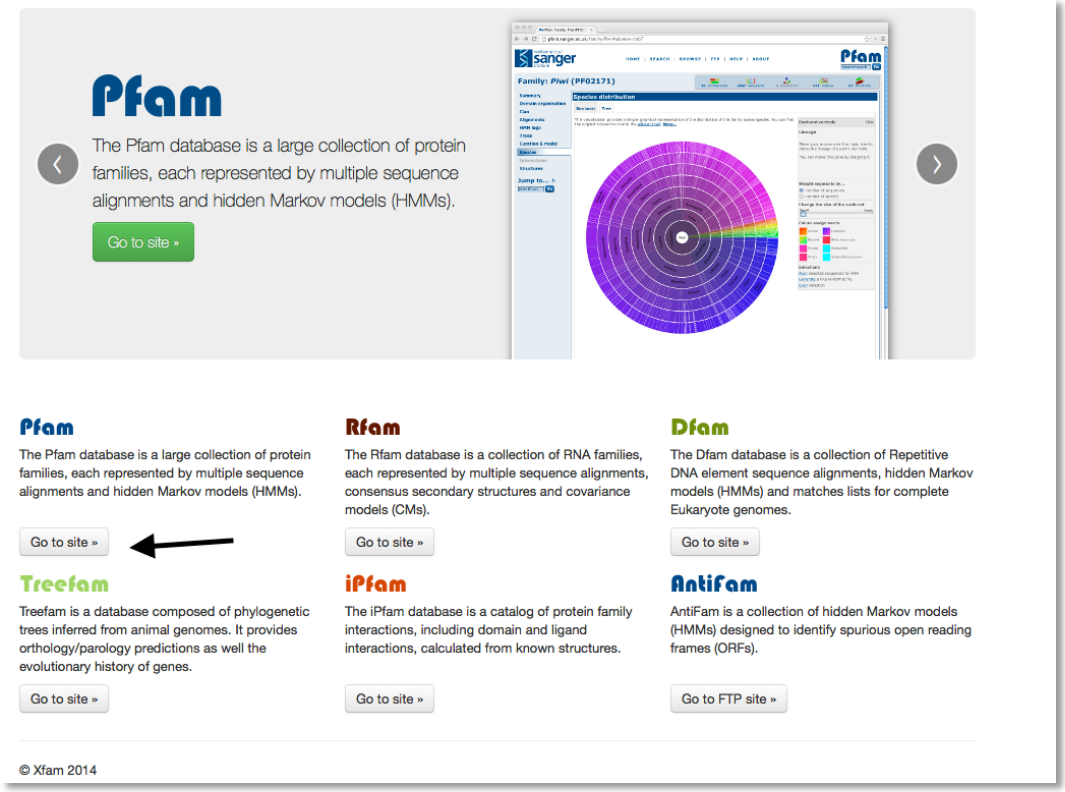


Figure 3.7. The Pfam entry page. Clicking the button indicated by the arrow opens the Pfam site.

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

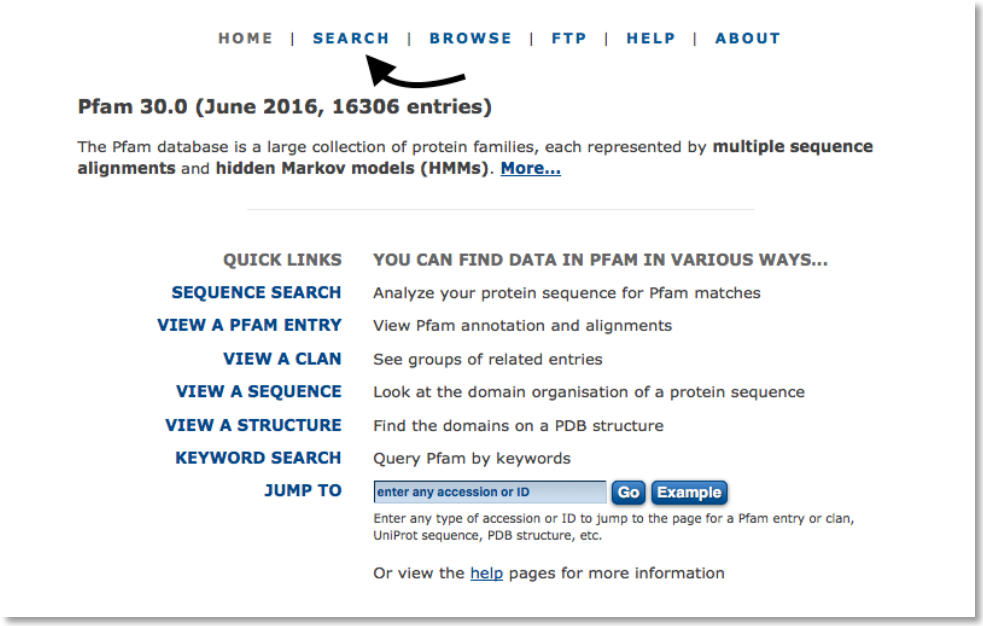


Figure 3.8. The Pfam search entry page. Click on the search tab as indicated by the arrow to begin the search.

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.

   3.1. Change the E-value to 0.001 and click submit.

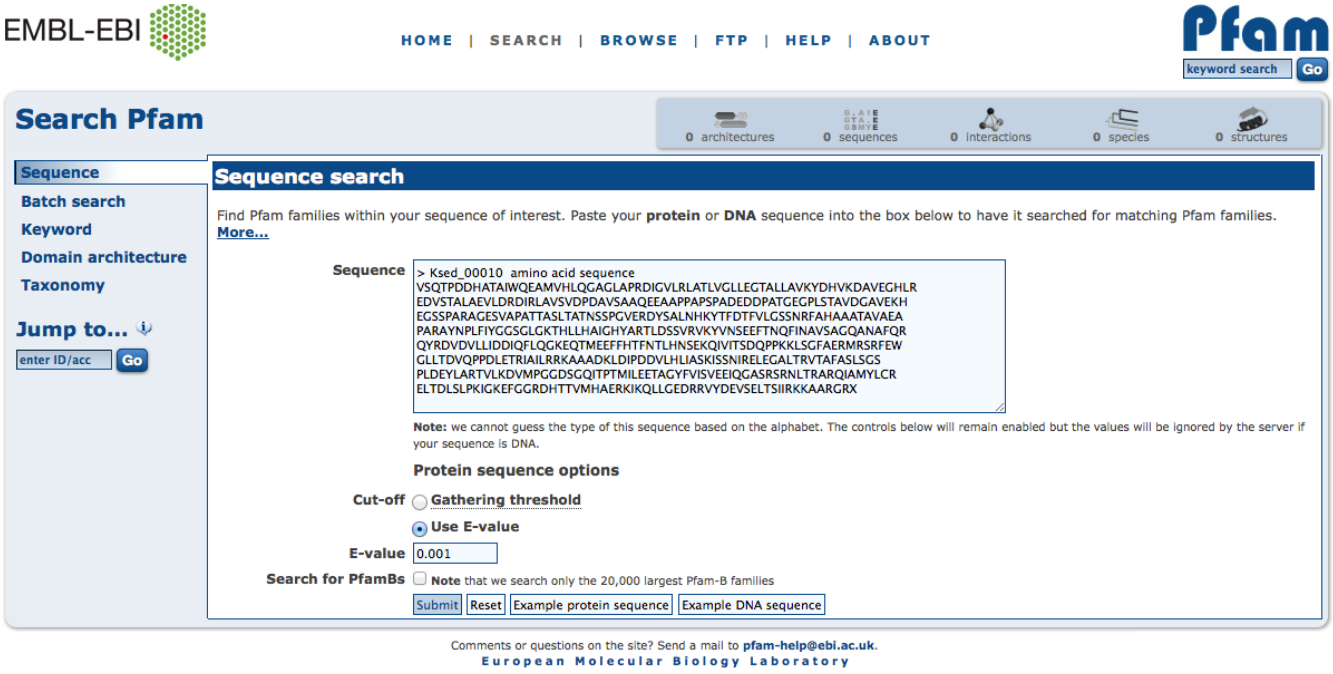


Figure 3.9. The Pfam search start page with the amino acid sequence of Ksed_00010 pasted into the text window.

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.

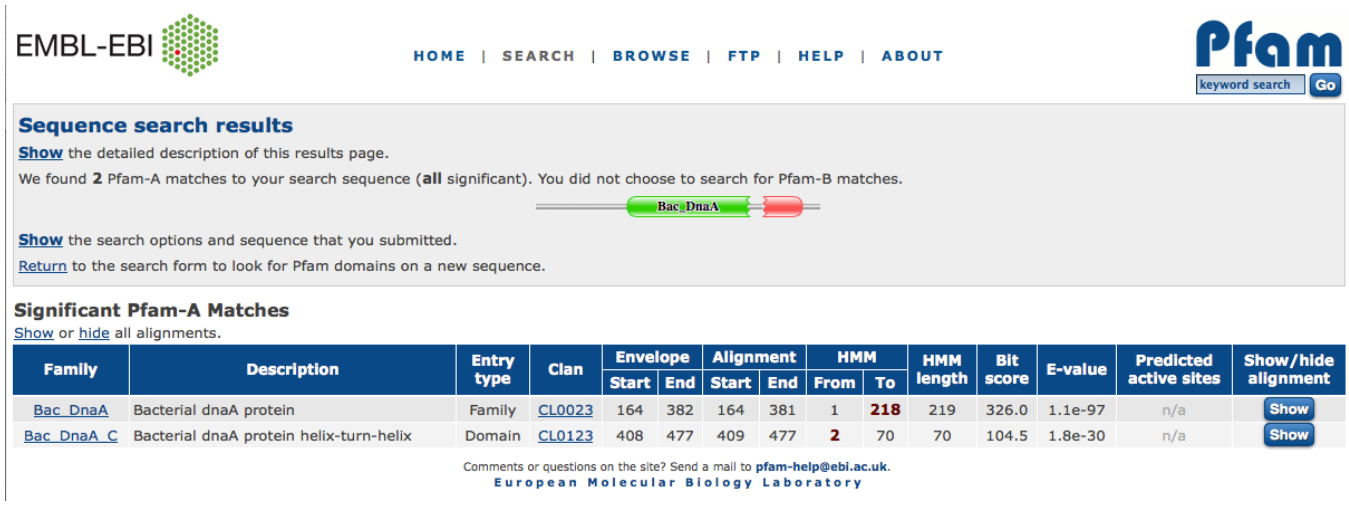
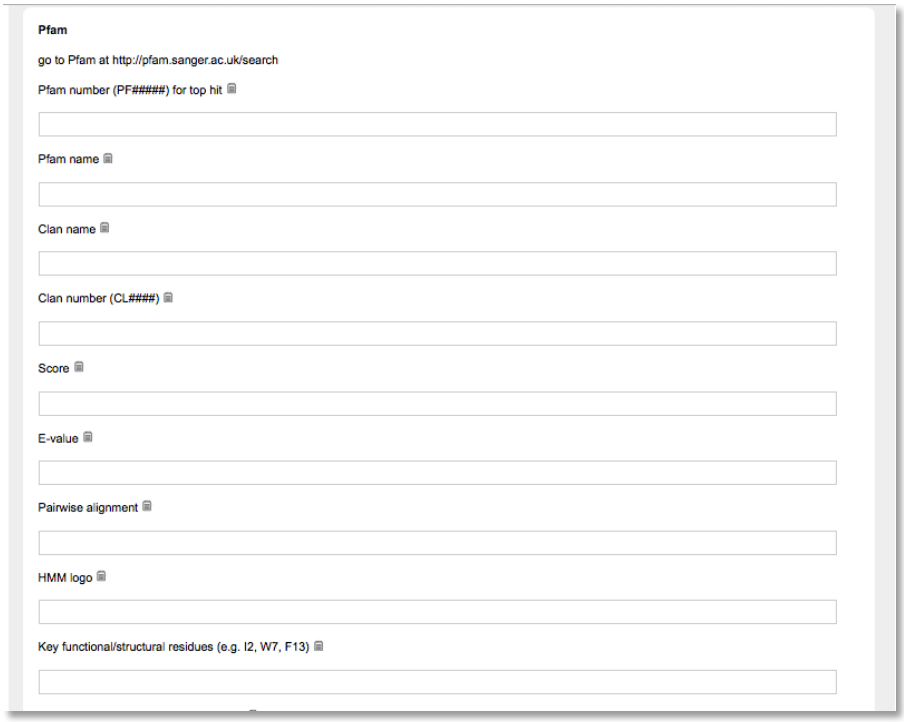


Figure 3.10. Pfam results for Ksed_00010. See the text for an explanation of the results.

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.



←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.

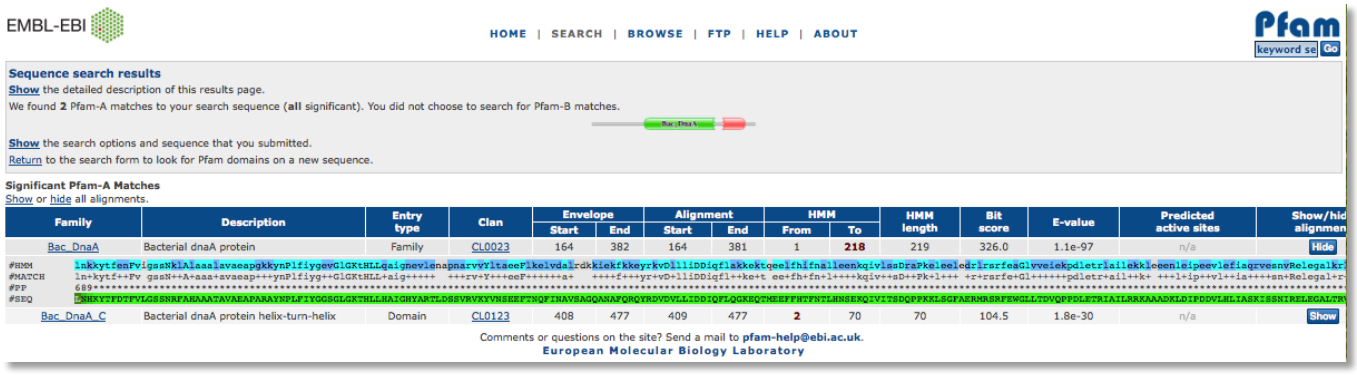


Figure 3.12. The Pfam results page with the show alignment option selected. See text section 3.5 above for an explanation.

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.

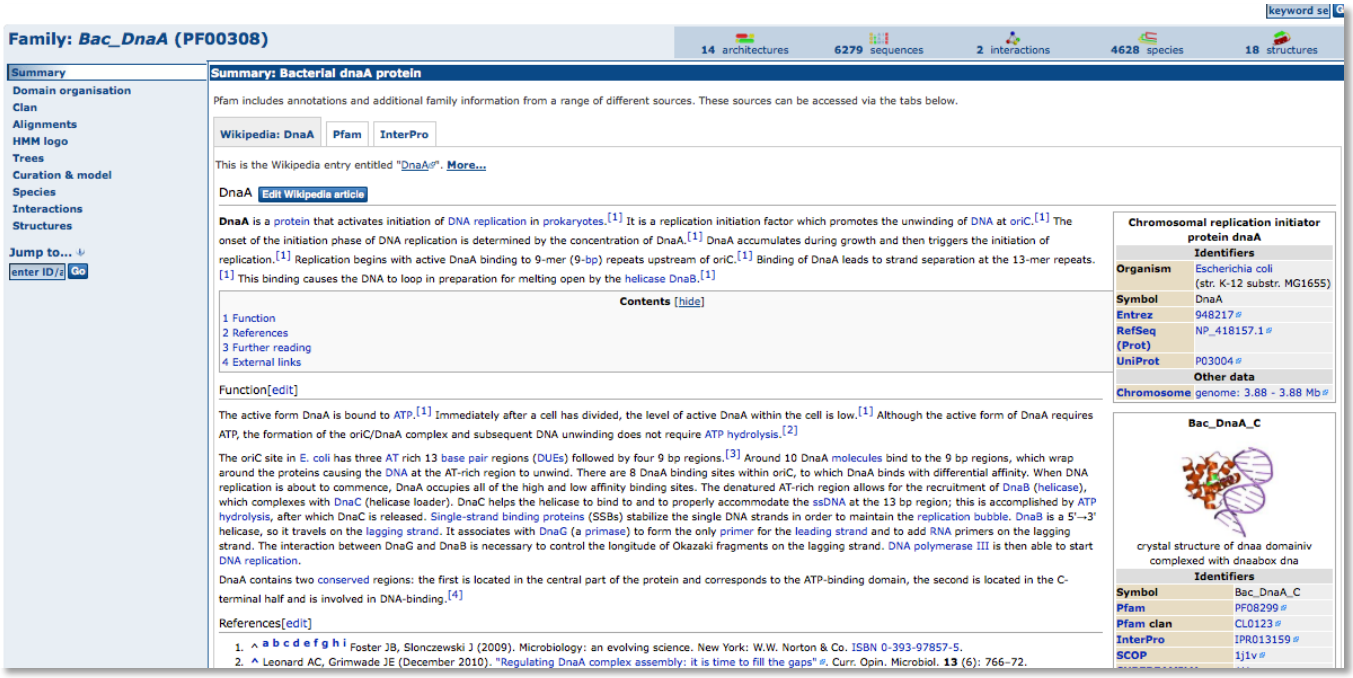


Figure 3.13. The Domain Summary page for Bac_DnaA.

6. 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.

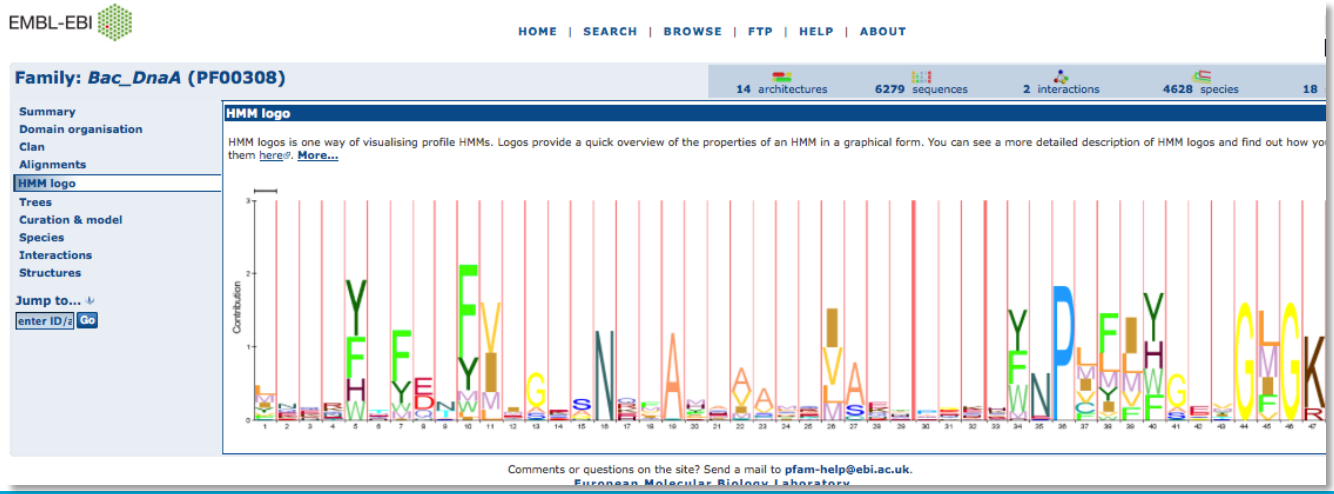


Figure 3.14. The HMM logo page

7. To find potential key functional residues (amino acids), we need to compare the pairwise alignment results with those of the HMM logo. This is illustrated in figure 3.15 below, where the arrows link amino acids in the HMM logo to the pairwise alignment of the query protein to the HMM. There are a number of amino acids that are predominant in the model HMM as indicated by the large wide letters at various positons in the logo (L1, F7, F10, P36 etc; amino acid position indicated by numbers at the tip of the HMM logo). Comparing the pairwise alignment to the HMM logo, one can see that L1, F7 (left arrow) , F10 and P36 (right arrow) are exact matches of the Ksed_00010 amino acid sequence and the major amino acid found at the same position in the HMM logo. All amino acids in the sequence meeting this criterion should be listed as being possible key functional residues. Note that there appears to be significant variability among proteins used to create this model HMM, as indicated by the large number of smaller letters under the predominant amino acids at those positions.

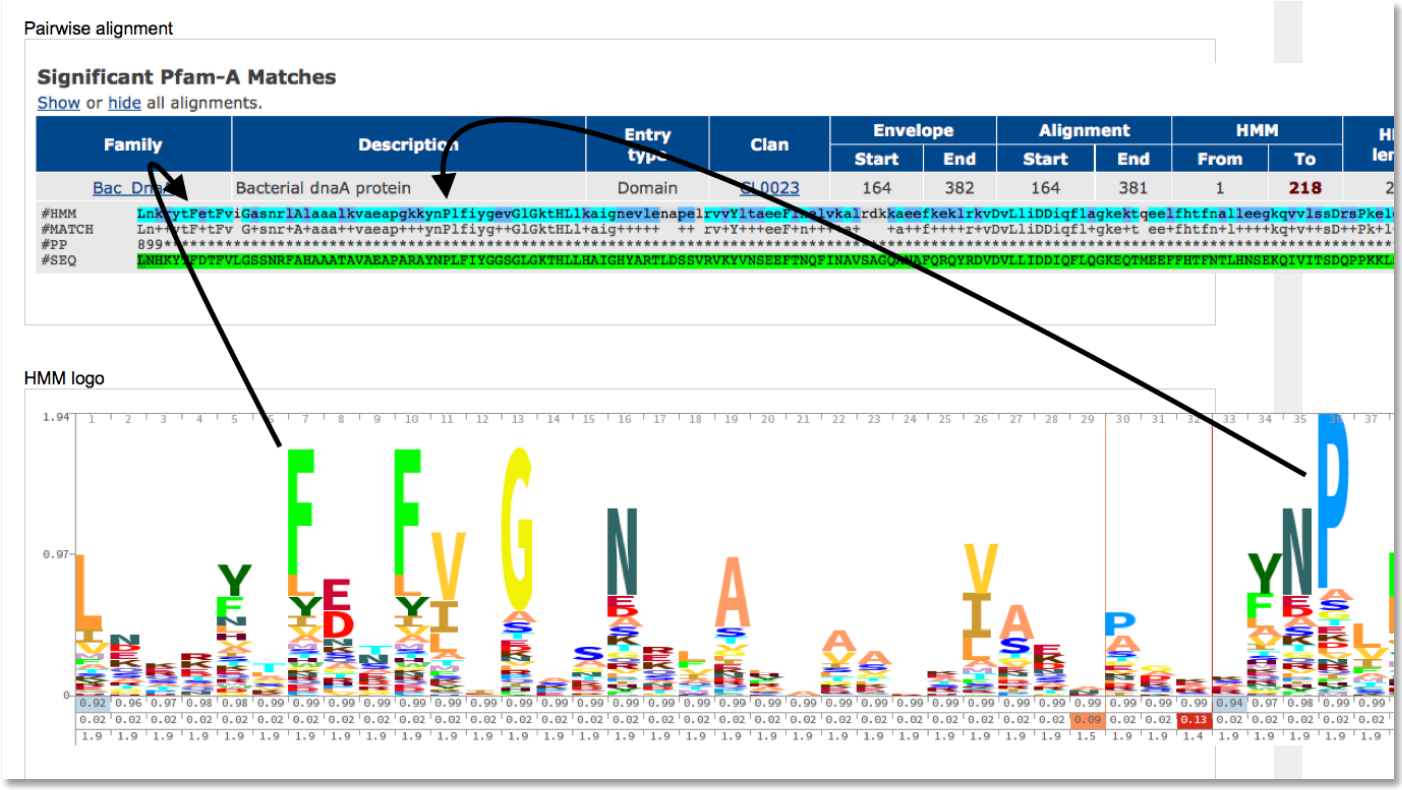


Figure 3.15. Comparison of the pairwise alignment and HMM logo results for Ksed_00010 and Bac DnaA.

8. Return to the Pfam results summary page to record Clan Information. Click in the Clan hyperlink for the Pfam hit on which you are working and then record the Clan Name and Number from this page in your Notebook.

9. Repeat procedure for all significant PFAM hits.

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.

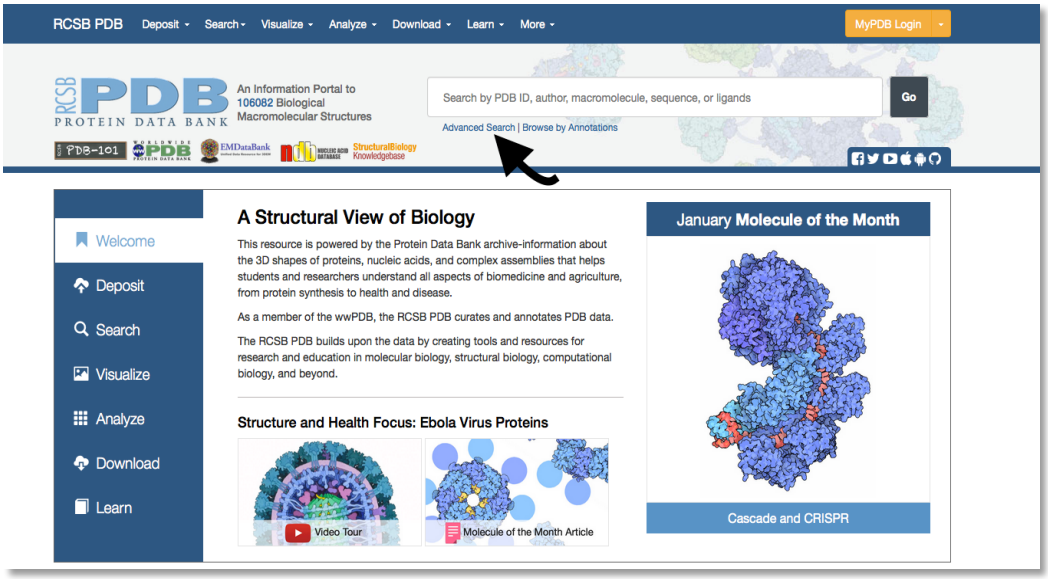


Figure 3.16. The Protein Data Bank start page. Click on Advanced (see arrow)

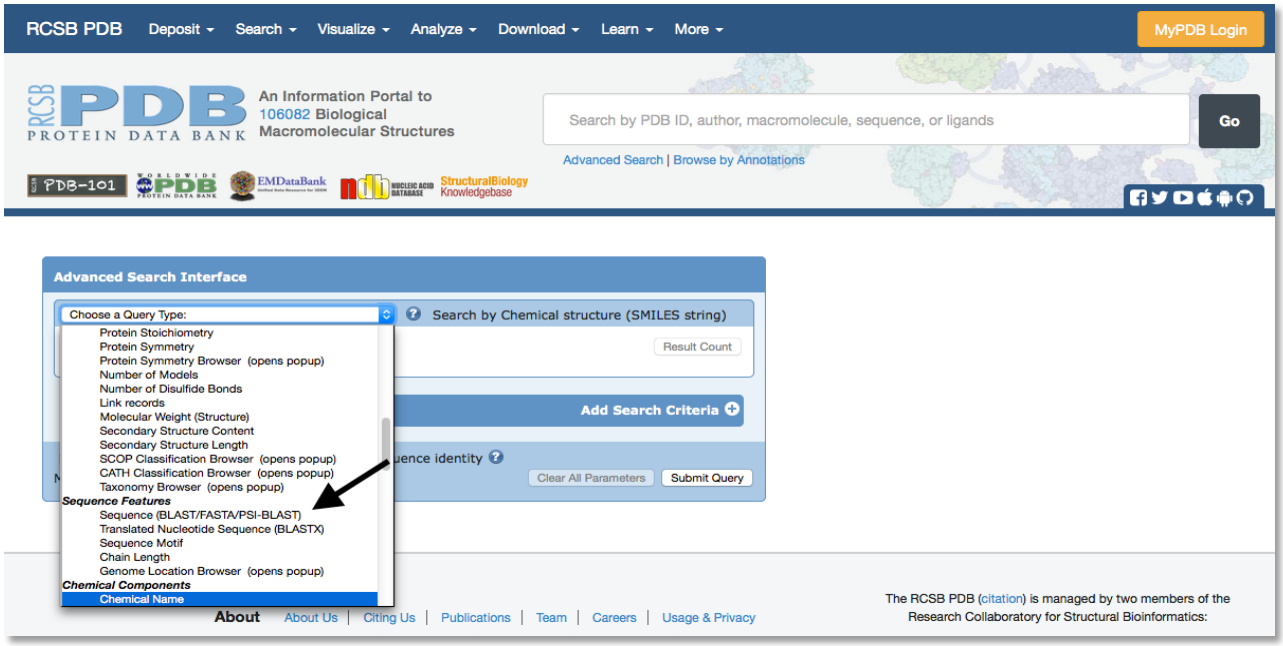


Figure 3.17. Selection of a BLAST search in PDB. Select the Choose Query Type pull down menu and select "Sequence (BLAST/FASTA/PSI-BLAST)" option

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).

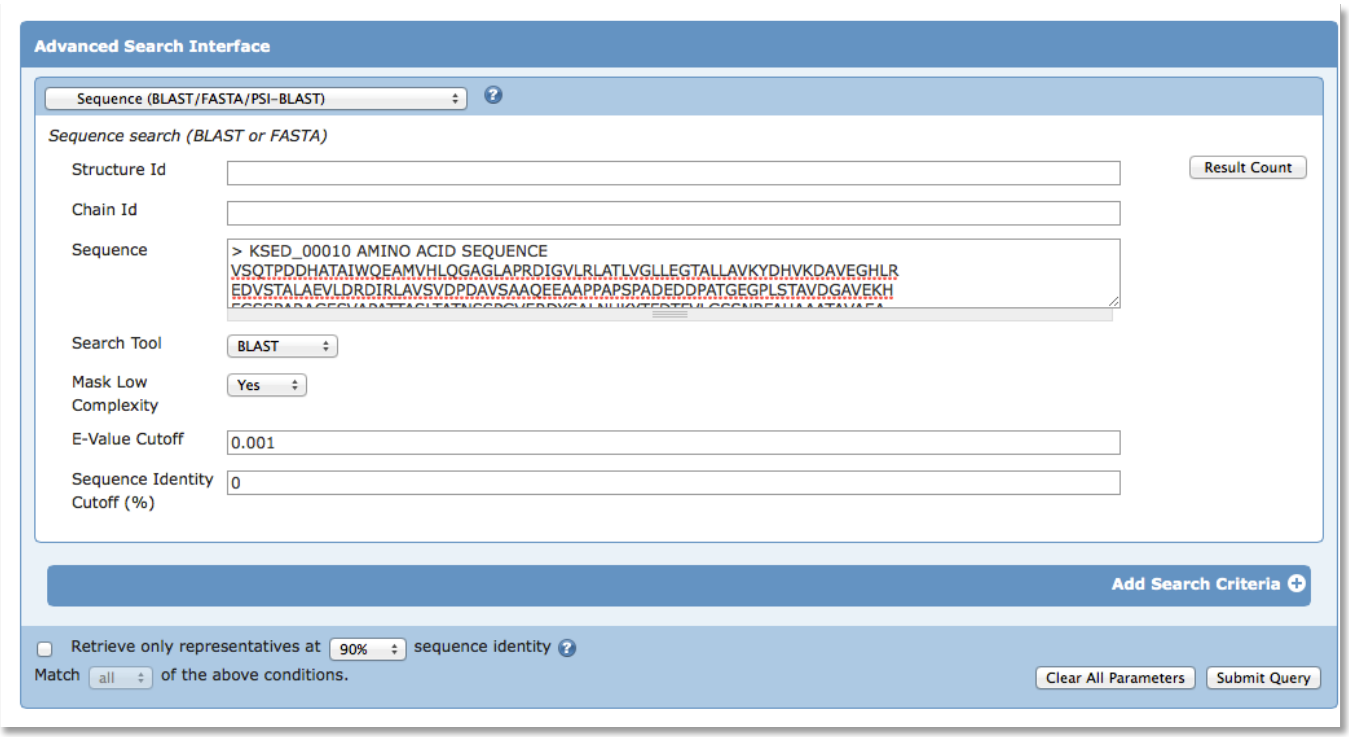


Figure 3.18. The PDB Advanced search page. The FASTA formatted Ksed_00010 amino acid sequence has been pasted into the sequence text box and the E-Value Cutoff has manually been changed to 0.001.

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).

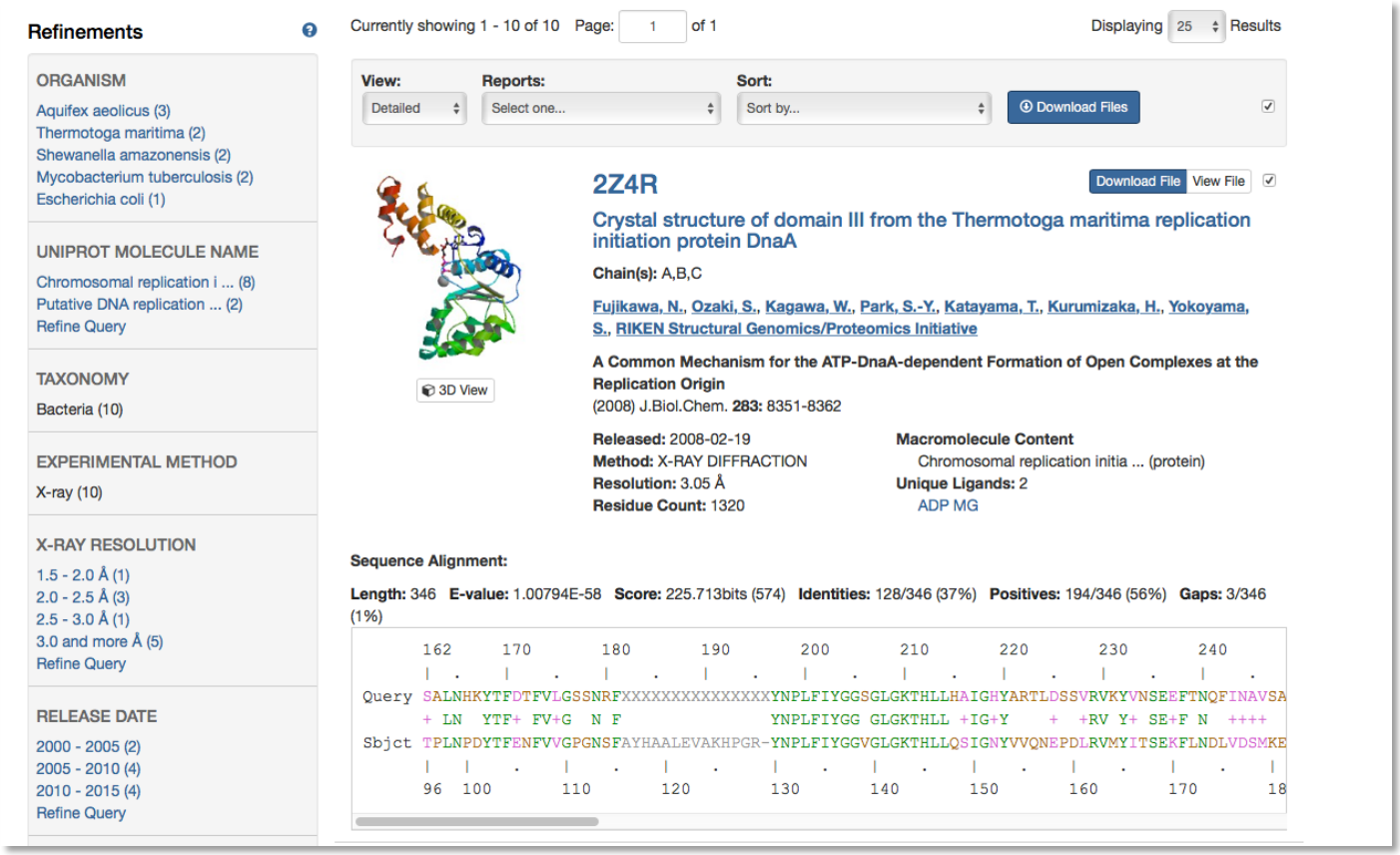


Figure 3.19. The top hit for the PDB search of Ksed_00010. The PDB code is 2Z4R, the name is "Crystal structure of domain III from…." and the E-value and score are as shown. The alignment of your sequence with that of the hit is also shown (using the slider at the bottom will allow the full alignment to be viewed from beginning to end.

Figure 3.20. The PDB notebook page.

5. Insert the Alignment into the Lab Notebook. It cannot be downloaded as an image directly, and copy and paste results in the alignment being distorted in the notebook. There are two ways to efficiently capture the alignment for your notebook. One is to Grab or Snip the image in sections (with some overlap between sections), save them as .png files and upload them to the notebook one at a time. **You should upload the images in REVERSE order in the notebook to make sure the 1st part of the alignment is at the top of the notebook page and the subsequent sections follow in order. Each upload appears at the top of the page by default (Figure 3.21).**

Figure 3.21. The pairwise alignment notebook page with the alignment pasted in as 2 separate png files.

6. A second, and easier method that can be used if the option is shown in your results, is to click on the report tab and select the "BLAST/FASTA/PSI-BLAST Results" tab as shown in figure 3.22 below. In the page that results, scroll down to the first pairwise alignment in the list (the top PDB hit) and SNIP or Grab the alignment as you did in your BLAST search in Module 2 (Figure 3.23)

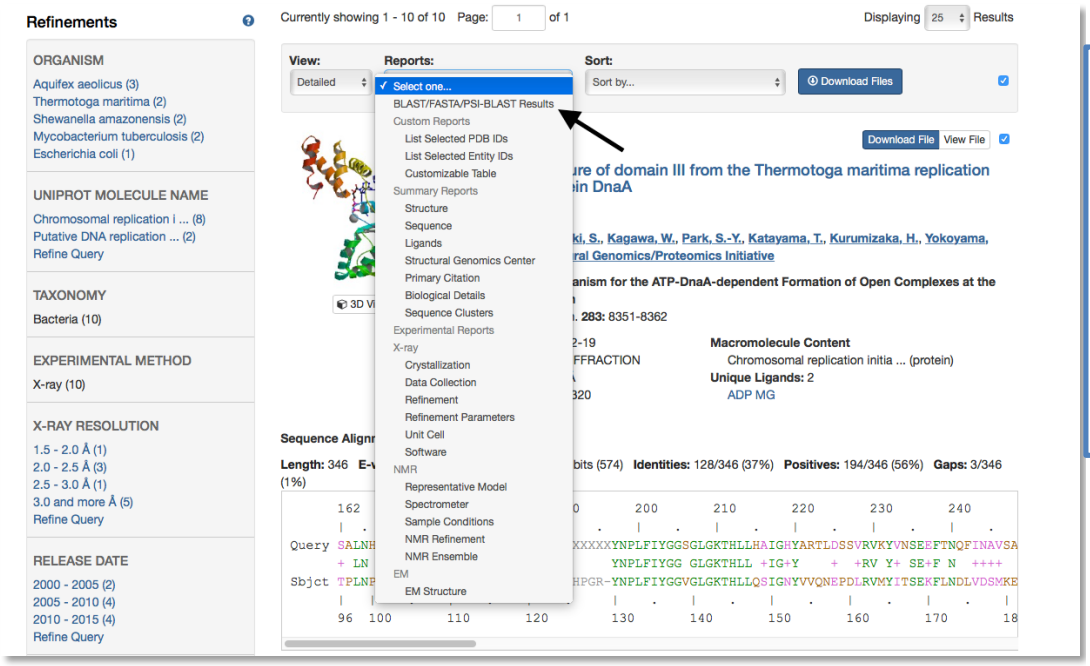


Figure 3.22. The report section pull down menu. The arrow points to the "BLAST/FASTA …" report that should be selected to obtain the pairwise alignment.

```
>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 SALNHKYTFDTFVLGSSNRFXXXXXXXXXXXXXXXYNPLFIYGGSGLGKTHLLHAIGHYA 221
           + LN  YTF+ FV+G  N F                YNPLFIYGG GLGKTHLL +IG+Y
Sbjct: 96  TPLNPDYTFENFVVGPGNSFAYHAALEVAKHPGR-YNPLFIYGGVGLGKTHLLQSIGNYV 154

Query: 222 RTLDSSVRVKYVNSEEFTNQFINAVSAGQANAFQRQYRD-VDVLLIDDIQFLQGKEQTME 280
              +  +RV Y+ SE+F N  ++++  G+ N F+ +YR  VD+LLIDD+QFL GK
Sbjct: 155 VQNEPDLRVMYITSEKFLNDLVDSMKEGKLNEFREKYRKKVDILLIDDVQFLIGKTGVQT 214

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

Query: 341 KAAADKLDIPDDVLHLIASKISSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMP 400
                +  ++P++VL+ +A  +  N+R L GA+ ++  +  ++  +G  +D   A  +LKD +
Sbjct: 275 MLEIEHGELPEEVLNFVAENVDDNLRRLRGAIIKLLVYKETTGKEVDLKEAILLLKDFIK 334

Query: 401 GGDSGQITPT-MILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIG 459
                + P    ++E  A     +  EEI    SR+      AR+I MY+ +      SL  I
Sbjct: 335 PNRVKAMDPIDELIEIVAKVTGVPREEILSNSRNVKALTARRIGMYVAKNYLKSSLRTIA 394

Query: 460 KEFGGRDHTTVMHAERKIKQLLGEDRRVYDEVSELTSIIRKKAARG 505
           ++F       V   ++    LL  ++++   + E+   I ++A  G
Sbjct: 395 EKFNRSHPVVVDSVKKVKDSLLKGNKQLKALIDEVIGEISRRALSG 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.