Module 2: Sequence-Based Similarity Data

Objective

The objective of this module is:

1. To determine if the protein you are annotating is similar to other known proteins using BLAST, CDD, T-COFFEE and WebLogo applications.
2. To document your search results in the Module 2 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.
- To have completed the Basic Information Module (Module 1).

General Background to Sequence-Based Similarity Module

In this module we will be looking to see the level of similarity of your protein to that of other proteins or protein domains (regions) in various databases. The idea behind this investigation is that the more similar your protein is to proteins described in those databases, the more likely it is that your protein will have a function similar to that of the protein in the database. This module is one of the most powerful and important of all the modules in determining the identity and function of your protein.

We will be working with the amino acid sequence of your protein in this module. The reason that we do not use the DNA sequence to look for similar genes, rather than proteins, is that there is **redundancy in the genetic code** (review the basic information document provided to you with your lab manual to see what this means). Thus DNA sequences can vary and yet encode the same amino acids. By looking at the amino acids themselves we completely remove the redundancy issue. The BLAST search you will perform is called BLASTP, or a protein-protein BLAST (meaning the amino acid query sequence you enter will be compared to all amino acid sequences in the database). Other BLAST searches are possible for other applications. For example, BLASTN can perform a nucleotide-nucleotide sequence search.

The BLAST and CDD parts of the module will find similarities between your protein and others in the database and give you data to interpret about the level of similarity. The T-COFFEE and WebLogo parts of the module will allow multiple alignments of your protein to others that have been identified in BLAST so that you can directly see the extent of similarity among all the matches. T-COFFEE and

WebLogo will allow you to identify which portions of your sequence are most conserved among all the sequences identified by your BLAST search.

Procedures

Log in To GENI-ACT

1. Log in to GENI-ACT page (http://GENI-ACT.org/) using specific user name and password assigned.

2. On the GENI-ACT page, shift-click the locus tag of your gene at the top of the page to open the gene information page and then click on the notebook link to open the lab notebook for the same gene.

3. Click on the "Module: Sequence-Based Similarity" tab to open the Sequence-Based Similarity section of the lab notebook.

Basic Local Alignment Search Tool (BLAST)

Background: BLAST is used to rapidly identify amino acid sequences that are related to a query sequence submitted by an investigator. Because the Genbank data base is so large and it would take so much time to do so, BLAST does not attempt to align the complete query amino acid sequence with every other sequence in the database. Rather it takes what is called a heuristic (http://en.wikipedia.org/wiki/Heuristic) approach to looking for regions of similarity in between the query sequence and those in the database. There regions are at first very small and are built outward. Only the sequences which continue to have a good match with the query sequence at one level continue to matched with the next "bigger" series of amino acids and so on. Though less accurate than matching the query sequence completely with every sequence in the database, the speed with which BLAST performs the search makes it a much more practical way to search a large set of sequences. More detail about BLAST can be found at http://en.wikipedia.org/wiki/BLAST.

1. Open the module 1 (Basic Information) notebook and copy the sequence for your gene along with the FASTA header containing the locus tag of your gene.
2. Navigate to the NCBI BLAST (http://www.ncbi.nlm.nih.gov/blast)
3. Paste your sequence into the Enter Query Sequence Box (Figure 2.1) and then select a database to search from the **Database** dropdown menu. The **non-redundant protein sequence (NR)** database is a massive repository of protein sequences derived primarily from sequenced genomes. The vast majority of the sequences in NR have never been manually annotated and do not have experimental evidence to support their function. It is likely that your query sequence will hit closely related sequence in NR, but the value of these hits in terms of identifying function of your protein may not be high. **SwissProt** is a much smaller sequence database that contains only curated sequences (meaning sequences are only added to this database once wet lab experimental evidence supports the function of the sequence). Hence, while it may be less likely that your sequence will match a very similar sequence in SwissProt, the prediction of the gene product can be taken with much higher confidence from SwissProt than from NR. It is recommended to run BLAST against ***BOTH*** the Swiss-Prot (as indicated in Figure 2.1***) AND*** the nr databases in two separate BLAST searches and then compare the results of each You should set up the nr blast first, as it takes longer to run, and then then run the Swiss-Prot in a second window**.** ***Compare the results obtained from both the nr***

and Swis-Prot searches. Things to keep in mind as you compare the results ar (described more fully in the paragraphs that follow):

a. Do both searches give significant results (as indicated by low E-values and high scores described below)?

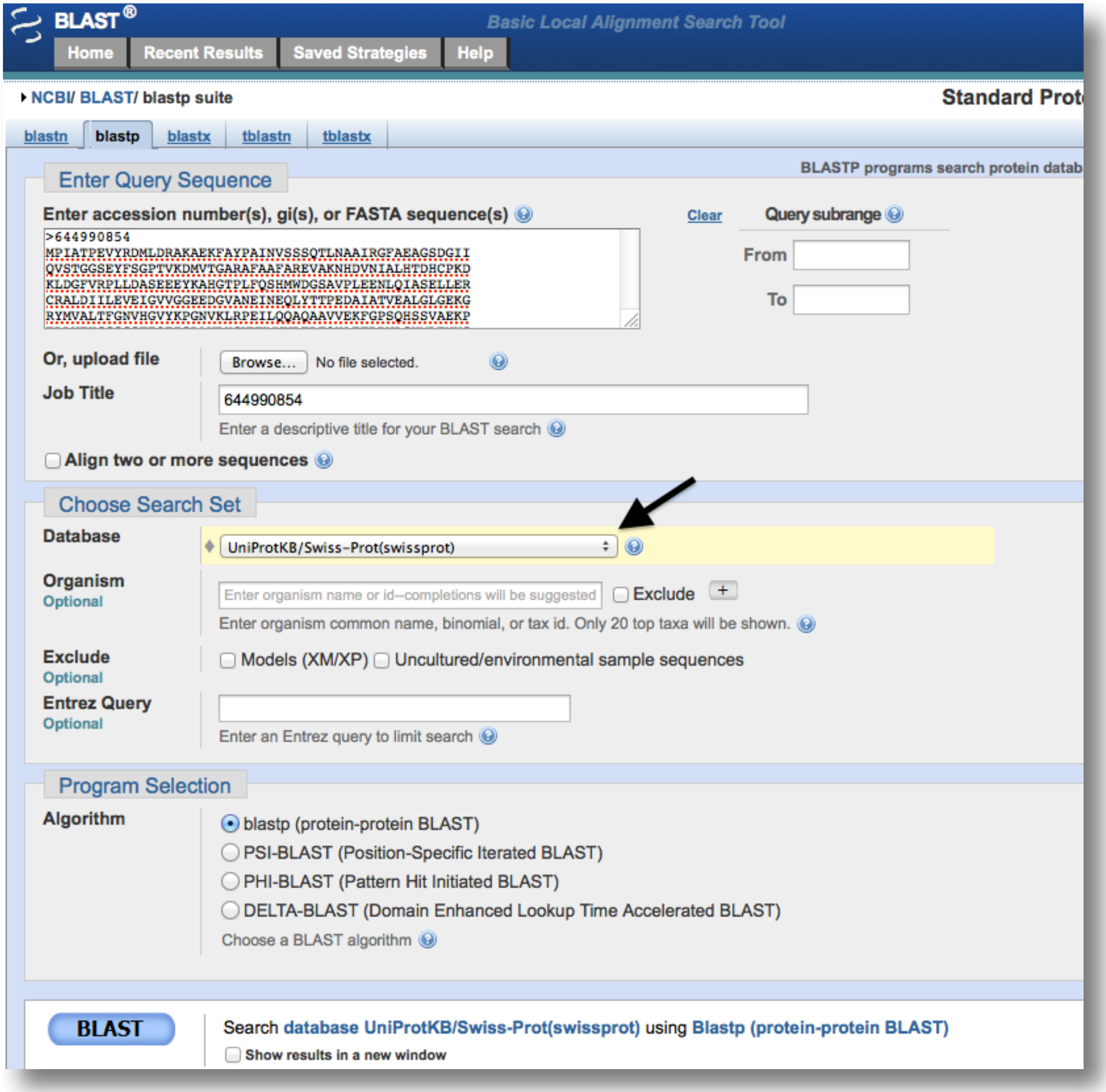
b. Are the names of the significant hits in both searches identical or very similar?

i. If the answer to both a and b above are yes, then you should use only the Swiss-Prot results to record in your notebook.

ii. If no significant hits (see 8b below) are found using SwissProt, but are found in nr, record that fact in your notebook and use the nr database.

iii. If significant hits are found in BOTH databases, but the names given to each seem to be different, then you should record results for the top 2 BLAST hits in Swiss-Prot and nr in the lab notebook as shown in the example notebook for Ksed_00010 available in your assignment.

4. Leave the other settings in Default.



BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite
Standard Prot
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein datab

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>644990854
MPIATPEVYRDMLDRAKAEKFAYPAINVSSSQTLNAAIRGFAEAGSDGII
QVSTGGSEYFSGPTVKDMVTGARAFAAFAREVAKNHDVNIALHTDHCPKD
KLDGFVRPLLDASEEEYKAHGTPLFQSHMWDGSAVPLEENLQIASELLER
CRALDIILEVEIGVVGGEEDGVANEINEQLYTTPEDAIATVEALGLGEKG
RYMVALTFGNVHGVYKPGNVKLRPEILQQAQAAVVEKFGPSQHSSVAEKP
```

Or, upload file: Browse... No file selected.

Job Title: 644990854

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: UniProtKB/Swiss-Prot(swissprot)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional:

Enter an Entrez query to limit search

Program Selection

Algorithm

● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)

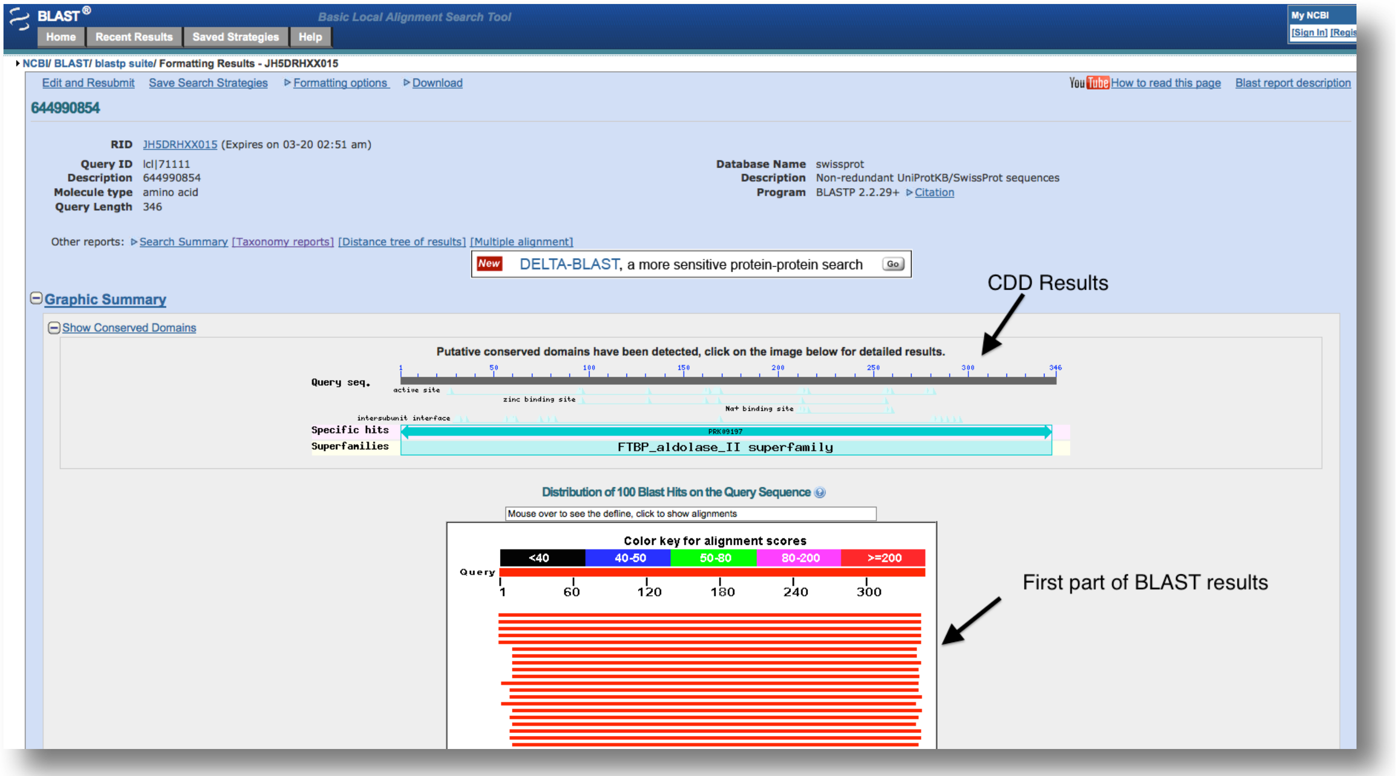
□ Show results in a new window

Figure 2.1. The BLAST search start page. Paste your sequence into the Enter Query Sequence box. The information in the FASTA header will automatically populate the job title box. You should select the UnitProtKB/Swiss-Prot(swissprot) data base from the dropdown menu indicated by the arrow as described in the text.

5. Click the "BLAST" button to search for the best protein sequence match. It may take seconds to minutes for the search to complete. When it does you will see results as illustrated in Figure 2.2 – 2.6.

6. On the BLAST result page, you will see results for both BLAST and CDD searches (CDD will be described later in this section) (Figure 2.2). Scroll down to the section labeled **Distribution of 100 Blast Hits on the Query Sequence**. This section is expanded in Figure 2.3. Across the top of the

alignment distribution you will ranges of scores that are in colored boxes. An alignment that falls within the range of scores indicated by a box will have a line of that same color. This allows you to quickly scan your results and determine whether there are a large number of good alignments (highest score color lines). In addition, the lines give you a visual representation of the coverage of the alignment with your query sequence. The number of residues in your sequence are indicated in a scale below the red line labeled query. If an alignment line extends nearly the full length of the scale, you can conclude that the alignment covers most of the sequence you submitted. A high score and close to 100% coverage would indicate a high quality alignment. The example distribution of alignments shown in figure 2.3 shows a large number of alignments with high scores and nearly complete coverage, suggesting this sequence is highly conserved in a number of different organisms.



644990854

RID JH5DRHXX015 (Expires on 03-20 02:51 am)

Query ID lcl|71111
Description 644990854
Molecule type amino acid
Query Length 346

Database Name swissprot
Description Non-redundant UniProtKB/SwissProt sequences
Program BLASTP 2.2.29+ Citation

Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

New DELTA-BLAST, a more sensitive protein-protein search Go

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

CDD Results

First part of BLAST results

Figure 2.2. The initial BLAST results page. Both a Conserved Domain Database (CDD Results) and BLAST searches are done simultaneously. The CDD results will be discussed below. BLAST result interpretations are discussed in the text.

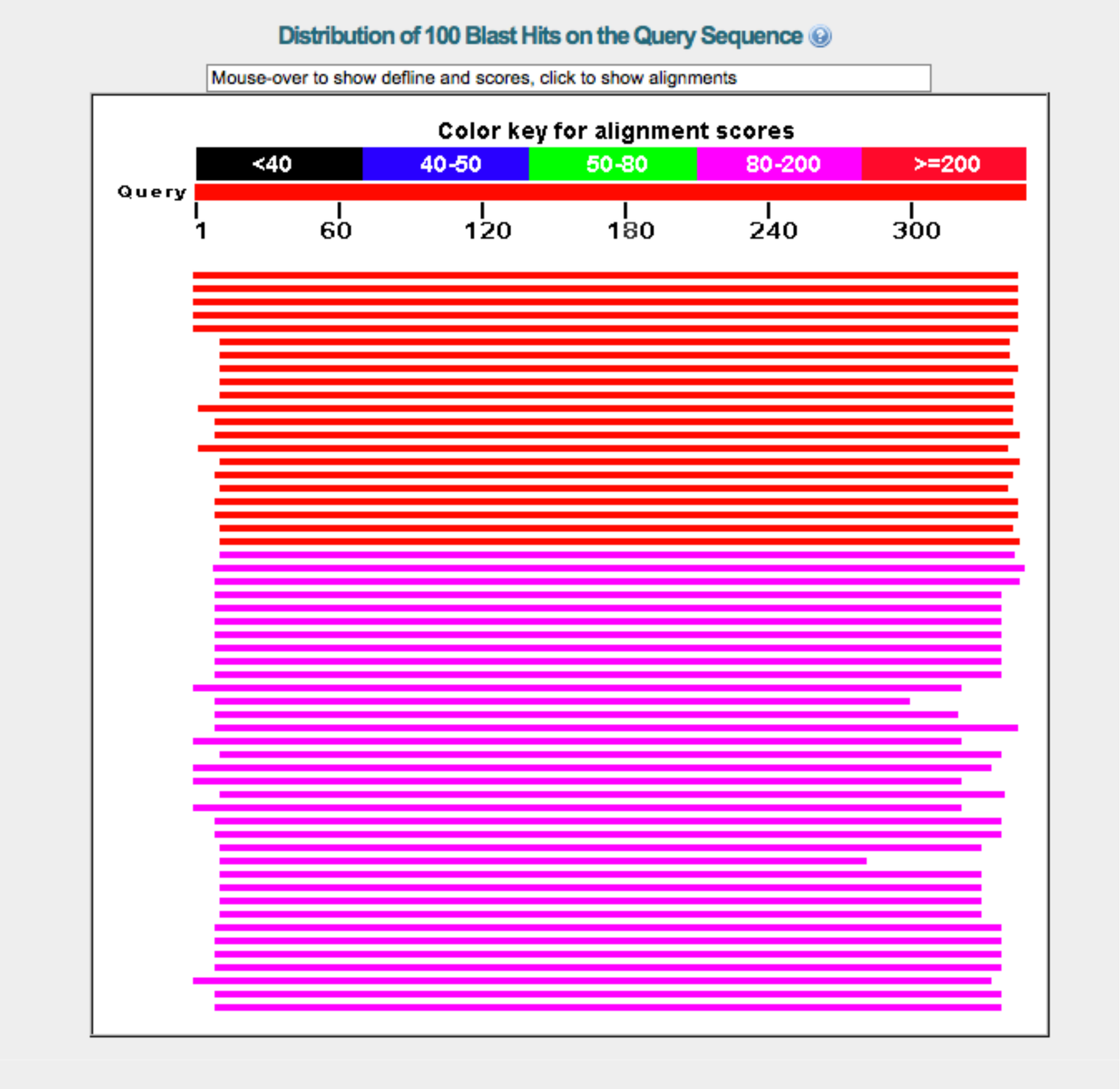
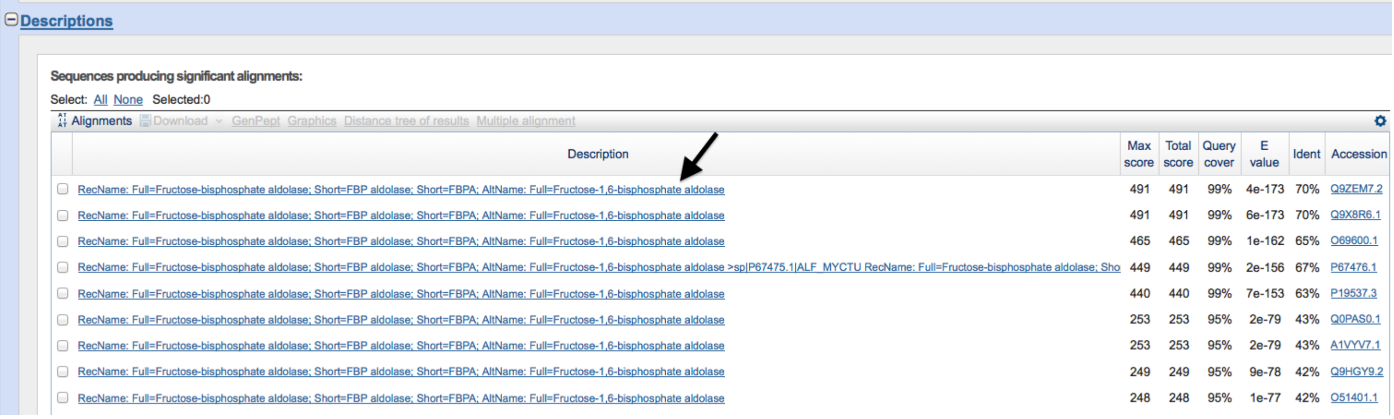

Distribution of 100 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
60
120
180
240
300

Figure 2.3. Distribution of BLAST hits resulting from the query sequence. The coverage of the hits are indicated by the length of the lines relative to the scale at the top of the figure. The different color lines indicate the score (discussed in text) as keyed above the scale of the query sequence. Clicking on any one of the lines will take you to the alignment for that hit. If you scroll down the page from this image you will see hyperlinks to the alignments in the same order as the lines in this figure. There you will find the actual statistics for the alignment.

7. Either clicking on the first alignment in the **Distribution of 100 Blast Hits on the Query Sequence** or scrolling down to the page to the section that looks like Figure 2.4 will allow you to collect quantitative data about the alignment that you will enter into your notebook.



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase	491	491	99%	4e-173	70%	Q9ZEM7.2
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase	491	491	99%	6e-173	70%	Q9X8R6.1
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase	465	465	99%	1e-162	65%	O69600.1
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase >sp|P67475.1|ALF_MYCTU RecName: Full=Fructose-bisphosphate aldolase; Sho	449	449	99%	2e-156	67%	P67476.1
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase	440	440	99%	7e-153	63%	P19537.3
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase	253	253	95%	2e-79	43%	Q0PAS0.1
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase	253	253	95%	2e-79	43%	A1VYV7.1
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase	249	249	95%	9e-78	42%	Q9HGY9.2
□	RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase	248	248	95%	1e-77	42%	O51401.1

Figure 2.4. The top BLAST hits found below the color alignments illustrated in Figure 2.3. The score, % coverage of the query and E value are shown, along with a hyperlink to the Genbank file describing the hit (Accession column). Clicking on the hyperlink indicated by the arrow in this figure will show the actual alignment (shown in Figure 2.5).

8. The items you will want to look at on the right side of the page are:

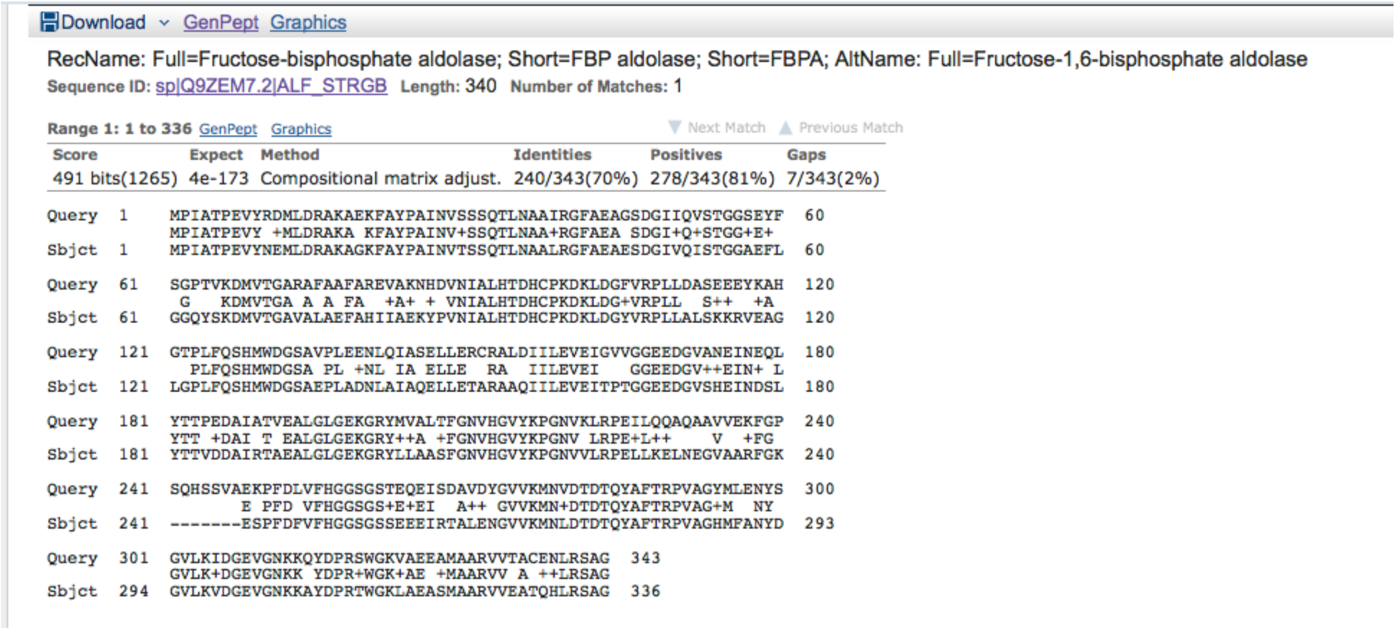
   a. **Score.** The score is a numerical representation of the quality of the alignment. It is calculated base on how well the sequences match, with higher numerical values assigned for exact matches, lower scores for "similar" amino acids and penalties assigned for gaps (see below) that are introduced to construct the alignment and for mismatches. The sum of these numbers is the score. The higher the score, the more likely the alignment is significant. You can see a more detailed explanation at the following link: http://en.wikipedia.org/wiki/BLAST.

   b. **Expect or E-value.** The E-value is the probability that this alignment could have occurred randomly. In general, we will consider an E-value significant if it is less than E-03. Note that this notation is the same as saying the E-value is 1 x 10^{-3}. Lower E-values indicate a lower probability that the observed match is due to random chance rather than actual similarity. Note the first alignment has an E-value of 4-173 or 4 x 10^{-173}. This is a VERY small number and a good indication that the match is significant. A low E-value should not be taken by itself as being an indicator of the quality of the alignment. ***If you do not have any significant BLAST hits (no hits or hits with E-values of greater than 1 x 10^{-3} using either SwissProt or the NR database searches), you should make that notation in your notebook and move onto the next module. A finding of no significant BLAST hits would indicate that no other sequence in the database has homology to your protein. The interpretation would thus be that you are dealing with a newly discovered protein or that that your protein has been called in error and does not really exist.***

c. **Query Coverage.** This value is shown as a percentage in the column. You will want to look at this value in combination with the Score and E-value to determine the quality of the alignment. The best alignments will have a highly significant E-value and a high percentage of coverage.

d. **Identities.** This value is given as a percentage as well, telling you what percentage of the amino acids in the alignment are an identical (see alignment below).

9. You will eventually record the score and E-value in your notebook, but before you do we will get some more information that you will need to record as well. To the left of the page shown in figure 2.4 you will find a hyperlink to click on that will show you the actual alignment of your query with the hit from the database (Figure 2.5).



RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase

Sequence ID: sp|Q9ZEM7.2|ALF_STRGB Length: 340 Number of Matches: 1

Range 1: 1 to 336 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
491 bits(1265)	4e-173	Compositional matrix adjust.	240/343(70%)	278/343(81%)	7/343(2%)

```
Query  1    MPIATPEVYRDMLDRAKAEKFAYPAINVSSSQTLNAAIRGFAEAGSDGIIQVSTGGSEYF  60
            MPIATPEVY +MLDRAKA KFAYPAINV+SSQTLNAA+RGFAEA SDGI+Q+STGG+E+
Sbjct  1    MPIATPEVYNEMLDRAKAGKFAYPAINVTSSQTLNAALRGFAEAESDGIVQISTGGAEFL  60

Query  61   SGPTVKDMVTGARAFAAFAREVAKNHDVNIALHTDHCPKDKLDGFVRPLLDASEEEYKAH  120
             G   KDMVTGA A A FA  +A+ + VNIALHTDHCPKDKLDG+VRPLL  S++  +A
Sbjct  61   GGQYSKDMVTGAVALAEFAHIIAEKYPVNIALHTDHCPKDKLDGYVRPLLALSKKRVEAG  120

Query  121  GTPLFQSHMWDGSAVPLEENLQIASELLERCRALDIILEVEIGVVGGEEDGVANEINEQL  180
              PLFQSHMWDGSA PL +NL IA ELLE  RA  IILEVEI   GGEEDGV++EIN+ L
Sbjct  121  LGPLFQSHMWDGSAEPLADNLAIAQELLETARAAQIILEVEITPTGGEEDGVSHEINDSL  180

Query  181  YTTPEDAIATVEALGLGEKGRYMVALTFGNVHGVYKPGNVKLRPEILQQAQAAVVEKFGP  240
            YTT +DAI T EALGLGEKGRY++A +FGNVHGVYKPGNV LRPE+L++    V  +FG
Sbjct  181  YTTVDDAIRTAEALGLGEKGRYLLAASFGNVHGVYKPGNVVLRPELLKELNEGVAARFGK  240

Query  241  SQHSSVAEKPFDLVFHGGSGSTEQEISDAVDYGVVKMNVDTDTQYAFTRPVAGYMLENYS  300
                   E PFD VFHGGSGS+E+EI  A++ GVVKMN+DTDTQYAFTRPVAG+M  NY
Sbjct  241  -------ESPFDFVFHGGSGSSEEEIRTALENGVVKMNLDTDTQYAFTRPVAGHMFANYD  293

Query  301  GVLKIDGEVGNKKQYDPRSWGKVAEEAMAARVVTACENLRSAG  343
            GVLK+DGEVGNKK YDPR+WGK+AE +MAARVV A ++LRSAG
Sbjct  294  GVLKVDGEVGNKKAYDPRTWGKLAEASMAARVVEATQHLRSAG  336
```

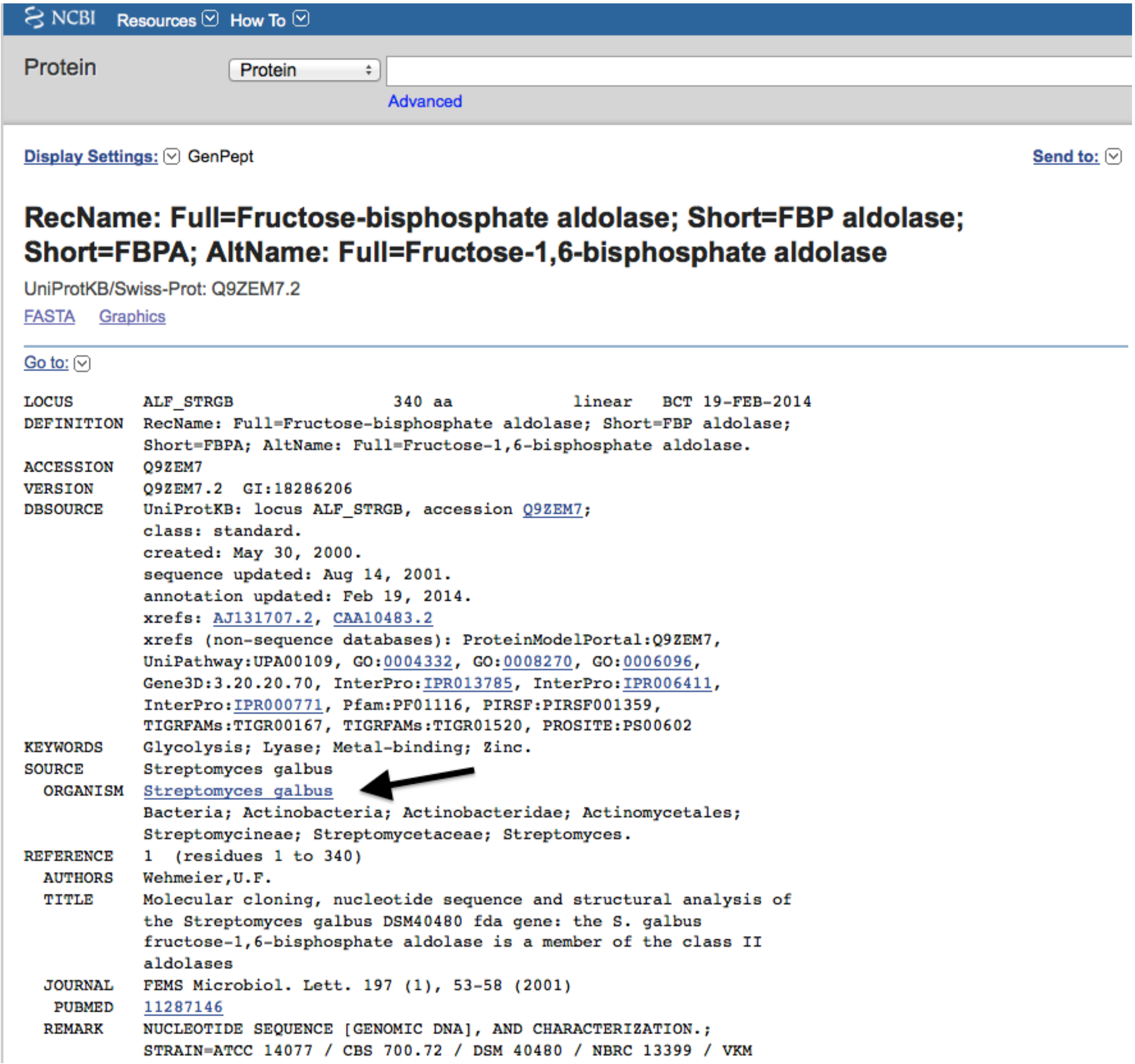
Figure 2.5. The alignment resulting when the first hyperlink in figure 2.4 was clicked. You will see more information in the alignment than from the list of hits. The score (491 in the example above), the Expect or E value (4e-173 or 4 x 10^{-173} in the example above) are the same as in the tabular form. We also see the number and percentages of identical amino acids, of positives (amino acids paired with amino acids of similar biochemical properties) and gaps. See the text for further explanation.

a. As you look at the alignment you will see the amino acid sequence of your protein in single letter code labeled "query" and the amino acid sequence of the match as "sbjct".

b. The line between these two sequences will tell you the extent of match. If the amino acid at a given position matches exactly between the query and subject, you will see that amino acid indicated. If there are amino acids with similar biochemical properties at a given position you will see a + indicated. No letter indicates a total mismatch between the query and subject. BLAST can also introduce gaps, indicated by a series of – symbols in the query or subject to get a better alignment between the two. In the example shown in figure 2.5 you will see a series of 7 gaps in the line beginning at 241 in the subject. You can think of these gaps as either insertion or deletion mutations that have occurred over evolutionary time in one or the other of the proteins.

c. Above the alignment you will see the score and E-values again, as well as the number of identical amino acids, the number of positives (matches between biochemical similar amino acids) and the number of gaps.

d. You will also see a hyperlink to take you to the Genbank record for the subject. Shift click on the hyperlink to open the record and you will see a page similar to figure 2.6. The Genbank record gives you information about how the subject sequence was entered into the database, the names of the scientists who submitted the sequence and any publication that resulted from their work. Of interest to you will be the name of the organism whose genome contained the subject sequence (see arrow in figure 2.6).



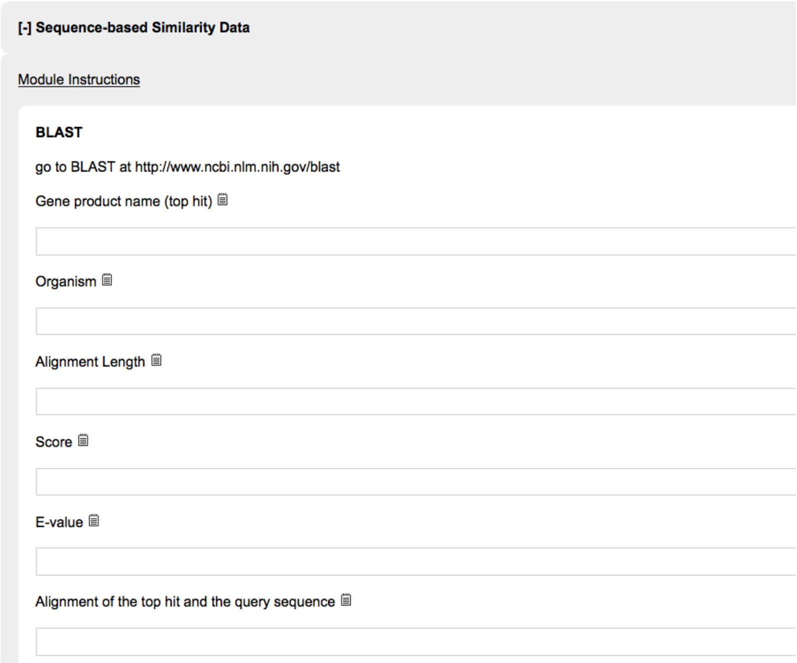
RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase; Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase

UniProtKB/Swiss-Prot: Q9ZEM7.2

```
LOCUS       ALF_STRGB               340 aa            linear   BCT 19-FEB-2014
DEFINITION  RecName: Full=Fructose-bisphosphate aldolase; Short=FBP aldolase;
            Short=FBPA; AltName: Full=Fructose-1,6-bisphosphate aldolase.
ACCESSION   Q9ZEM7
VERSION     Q9ZEM7.2  GI:18286206
DBSOURCE    UniProtKB: locus ALF_STRGB, accession Q9ZEM7;
            class: standard.
            created: May 30, 2000.
            sequence updated: Aug 14, 2001.
            annotation updated: Feb 19, 2014.
            xrefs: AJ131707.2, CAA10483.2
            xrefs (non-sequence databases): ProteinModelPortal:Q9ZEM7,
            UniPathway:UPA00109, GO:0004332, GO:0008270, GO:0006096,
            Gene3D:3.20.20.70, InterPro:IPR013785, InterPro:IPR006411,
            InterPro:IPR000771, Pfam:PF01116, PIRSF:PIRSF001359,
            TIGRFAMs:TIGR00167, TIGRFAMs:TIGR01520, PROSITE:PS00602
KEYWORDS    Glycolysis; Lyase; Metal-binding; Zinc.
SOURCE      Streptomyces galbus
  ORGANISM  Streptomyces galbus
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Streptomycineae; Streptomycetaceae; Streptomyces.
REFERENCE   1  (residues 1 to 340)
  AUTHORS   Wehmeier,U.F.
  TITLE     Molecular cloning, nucleotide sequence and structural analysis of
            the Streptomyces galbus DSM40480 fda gene: the S. galbus
            fructose-1,6-bisphosphate aldolase is a member of the class II
            aldolases
  JOURNAL   FEMS Microbiol. Lett. 197 (1), 53-58 (2001)
   PUBMED   11287146
  REMARK    NUCLEOTIDE SEQUENCE [GENOMIC DNA], AND CHARACTERIZATION.;
            STRAIN=ATCC 14077 / CBS 700.72 / DSM 40480 / NBRC 13399 / VKM
```

Figure 2.6. Sequence ID information for BLAST hit in figure 2.6. This information will appear after clicking on the sequence ID hyperlink above the alignment in Figure 2.6. You will need to perform this action to find the name of the organism from which this sequence was retrieved. In this case the organism name is *Streptpmyces galbus*.

10. Record the Gene product name, Organism, Alignment length (equals the number of the last residue from the query sequence aligned minus the number of the first residue aligned plus 1), Score, and E-value for the top hit found in the **Lab Notebook** for that gene (Figure 2.7).
11. Use either the Grab tool on a Mac or the Snip tool on a PC to capture the alignment (ask your instructor how to perform this manipulation if you are not aware of how to do it). Copy and paste the Alignment of the top hit with the query sequence into the **Lab Notebook.** Figures 2.8-2.12 show how to upload an image to the notebook. Comment on the E-value and compare the length of the top hit to the query sequence.
12. Repeat steps 10 and 11 for the next best BLAST hit.



[-] Sequence-based Similarity Data

Module Instructions

BLAST

go to BLAST at http://www.ncbi.nlm.nih.gov/blast

Gene product name (top hit)

Organism

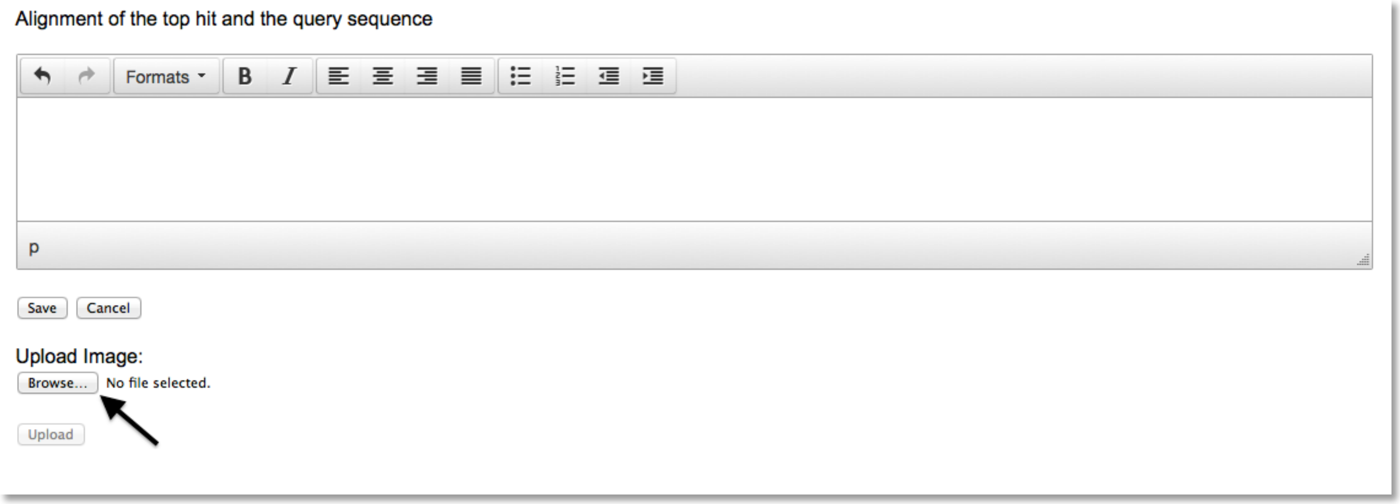
Alignment Length

Score

E-value

Alignment of the top hit and the query sequence

Figure 2.7. The sequence based similarity notebook page in module 2 of GENI-ACT.

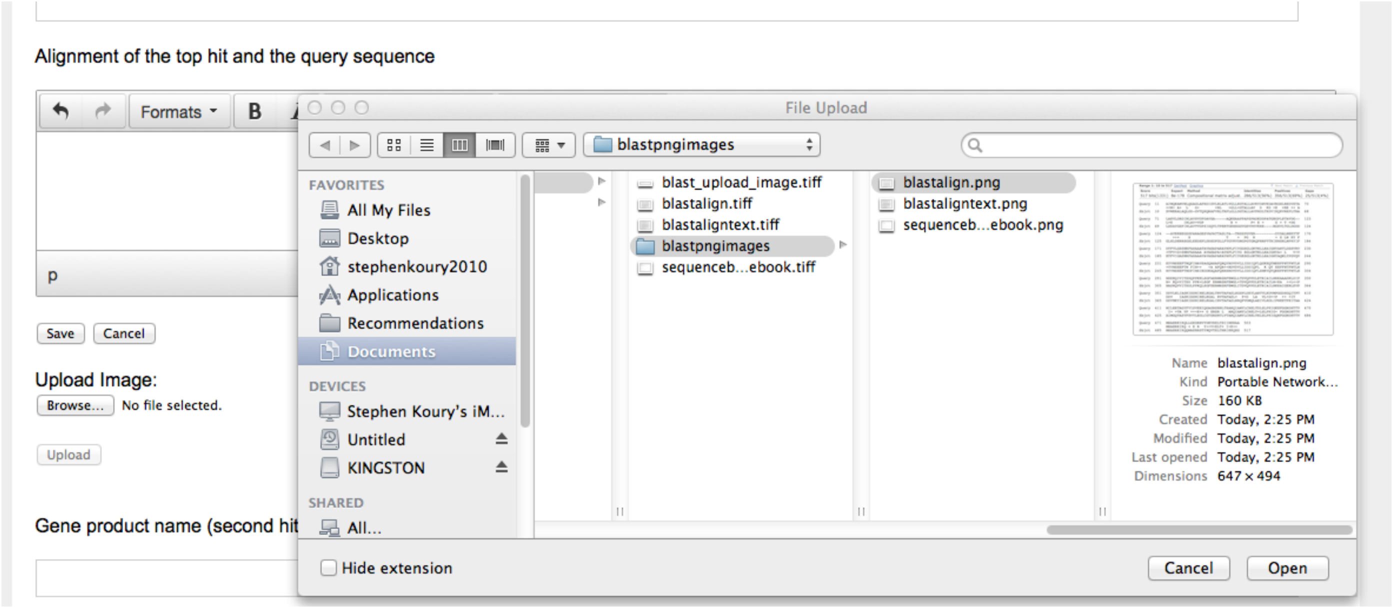


Alignment of the top hit and the query sequence

Upload Image:

No file selected.

Figure 2.8. After “snipping” or “grabbing” the image of the alignment and saving a copy as a .png file on a hard drive or flash drive, click the edit icon in the notebook and then click on the browse button as shown by the arrow.



Alignment of the top hit and the query sequence

Upload Image:

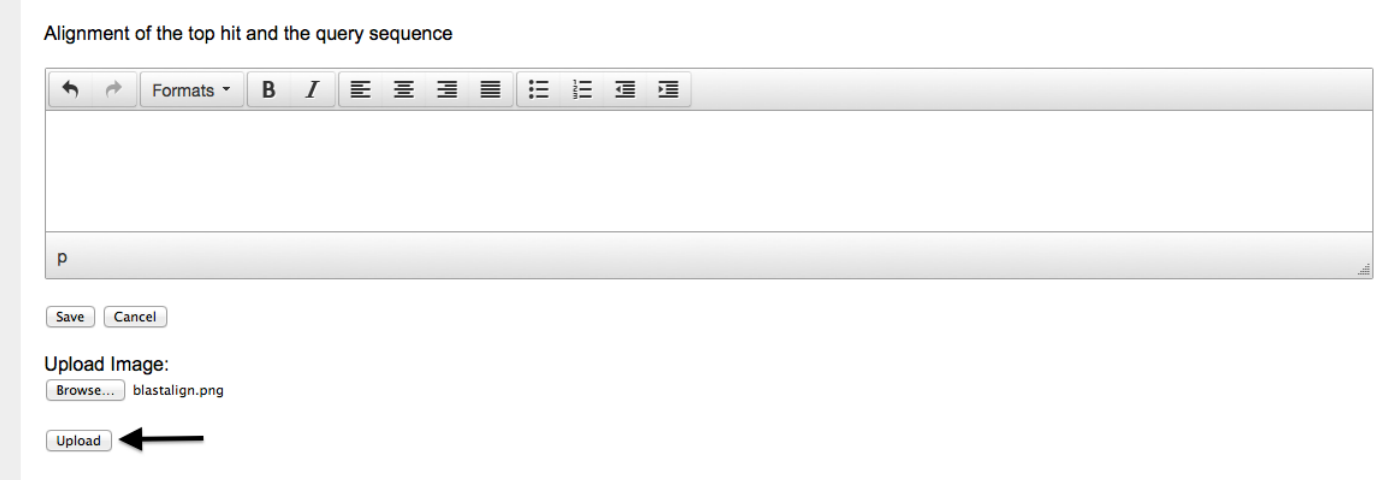
No file selected.

Gene product name (second hit

File Upload

Name	blastalign.png
Kind	Portable Network...
Size	160 KB
Created	Today, 2:25 PM
Modified	Today, 2:25 PM
Last opened	Today, 2:25 PM
Dimensions	647 × 494

Figure 2.9. Navigate to the saved image file of the alignment and select or “open” it. This figure shows windows on a Mac for doing so.

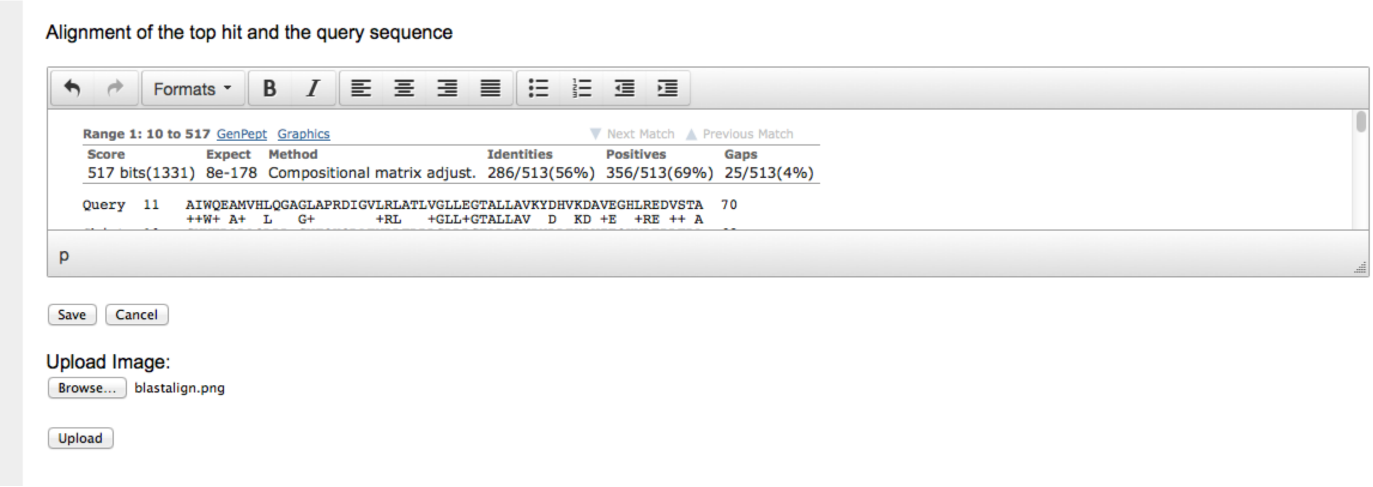


Alignment of the top hit and the query sequence

Upload Image:

blastalign.png

Figure 2.10. After selecting the file you should see the file name appear next to the Browse button. Click the upload button to add it to the notebook.



Alignment of the top hit and the query sequence

Range 1: 10 to 517 GenPept Graphics ▼ Next Match ▲ Previous Match

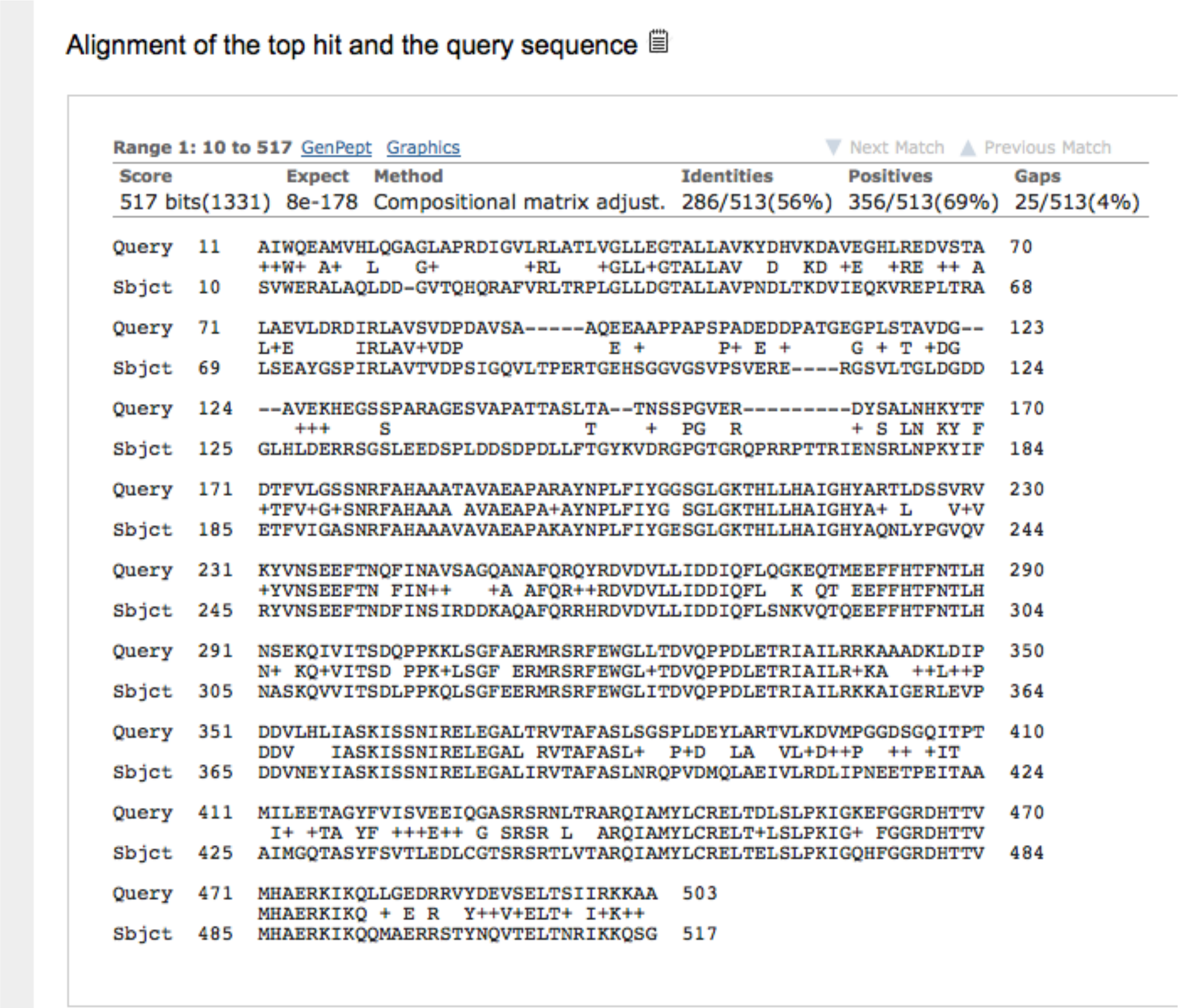
Score	Expect	Method	Identities	Positives	Gaps
517 bits(1331)	8e-178	Compositional matrix adjust.	286/513(56%)	356/513(69%)	25/513(4%)

```
Query  11   AIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEGHLREDVSTA  70
            ++W+ A+  L   G+         +RL    +GLL+GTALLAV   D   KD +E   +RE ++ A
```

Upload Image:

blastalign.png

Figure 2.11. After uploading the image you will see part of it appear in the notebook page as shown in this figure. Click Save to permanently add it to your notebook.



Alignment of the top hit and the query sequence

Range 1: 10 to 517 GenPept Graphics
▼ Next Match ▲ Previous Match

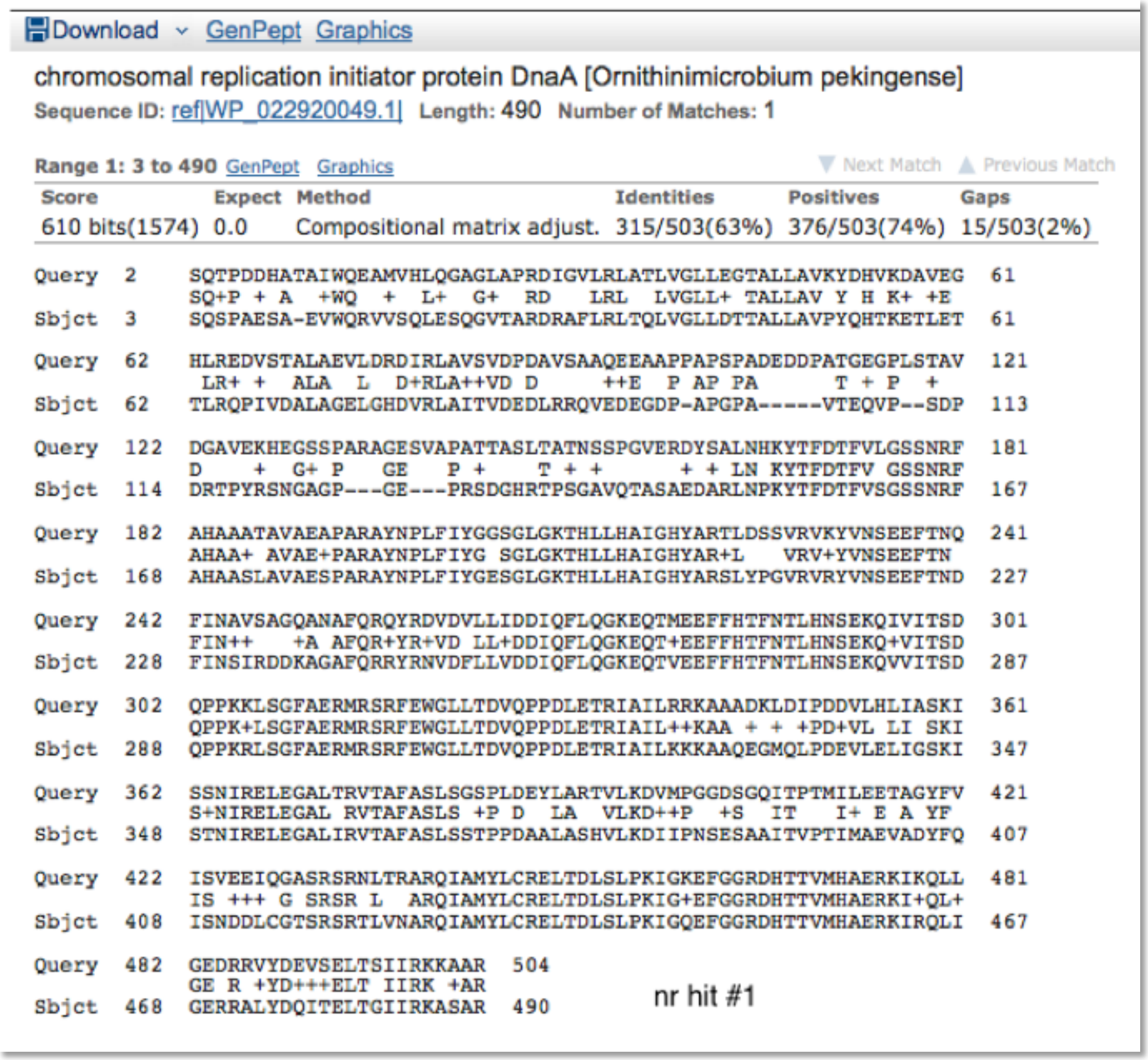
Score	Expect	Method	Identities	Positives	Gaps
517 bits(1331)	8e-178	Compositional matrix adjust.	286/513(56%)	356/513(69%)	25/513(4%)

```
Query  11   AIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEGHLREDVSTA  70
            ++W+ A+  L   G+       +RL   +GLL+GTALLAV  D  KD +E  +RE ++ A
Sbjct  10   SVWERALAQLDD-GVTQHQRAFVRLTRPLGLLDGTALLAVPNDLTKDVIEQKVREPLTRA  68

Query  71   LAEVLDRDIRLAVSVDPDAVSA-----AQEEAAPPAPSPADEDDPATGEGPLSTAVDG--  123
            L+E     IRLAV+VDP            E +      P+ E +     G + T +DG
Sbjct  69   LSEAYGSPIRLAVTVDPSIGQVLTPERTGEHSGGVGSVPSVERE----RGSVLTGLDGDD  124

Query  124  --AVEKHEGSSPARAGESVAPATTASLTA--TNSSPGVER---------DYSALNHKYTF  170
               +++    S                T    +  PG  R         + S LN KY F
Sbjct  125  GLHLDERRSGSLEEDSPLDDSDPDLLFTGYKVDRGPGTGRQPRRPTTRIENSRLNPKYIF  184

Query  171  DTFVLGSSNRFAHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRV  230
            +TFV+G+SNRFAHAAA AVAEAPA+AYNPLFIYG SGLGKTHLLHAIGHYA+ L   V+V
Sbjct  185  ETFVIGASNRFAHAAAVAVAEAPAKAYNPLFIYGESGLGKTHLLHAIGHYAQNLYPGVQV  244

Query  231  KYVNSEEFTNQFINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLH  290
            +YVNSEEFTN FIN++   +A AFQR++RDVDVLLIDDIQFL  K QT EEFFHTFNTLH
Sbjct  245  RYVNSEEFTNDFINSIRDDKAQAFQRRHRDVDVLLIDDIQFLSNKVQTQEEFFHTFNTLH  304

Query  291  NSEKQIVITSDQPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIP  350
            N+ KQ+VITSD PPK+LSGF ERMRSRFEWGL+TDVQPPDLETRIAILR+KA  ++L++P
Sbjct  305  NASKQVVITSDLPPKQLSGFEERMRSRFEWGLITDVQPPDLETRIAILRKKAIGERLEVP  364

Query  351  DDVLHLIASKISSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMPGGDSGQITPT  410
            DDV   IASKISSNIRELEGAL RVTAFASL+  P+D  LA  VL+D++P  ++ +IT
Sbjct  365  DDVNEYIASKISSNIRELEGALIRVTAFASLNRQPVDMQLAEIVLRDLIPNEETPEITAA  424

Query  411  MILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTV  470
             I+ +TA YF +++E++ G SRSR L  ARQIAMYLCRELT+LSLPKIG+ FGGRDHTTV
Sbjct  425  AIMGQTASYFSVTLEDLCGTSRSRTLVTARQIAMYLCRELTELSLPKIGQHFGGRDHTTV  484

Query  471  MHAERKIKQLLGEDRRVYDEVSELTSIIRKKAA  503
            MHAERKIKQ + E R  Y++V+ELT+ I+K++
Sbjct  485  MHAERKIKQQMAERRSTYNQVTELTNRIKKQSG  517
```

Figure 2.12. The appearance of the notebook page after you have saved the image file that was uploaded to the notebook.

13. The results obtained for the nr database search will be slightly different in the way they are displayed in terms of the pairwise alignments. The first thing to keep in mind is that the first hit may, in fact, be an exact match to the protein under investigation. If you see that the top hit has 100% query coverage and 100% identity to the query, it is likely that the first hit is your protein. Thus, in the nr database you often DO NOT include the first hit in the list as the top hit in your notebook, but rather skip to the second hit in the list as your "top" hit. Secondly, as is shown in figure 2.12b, the organism name

appears in the text above the alignment, making it unnecessary to open the full Genbank record to find that information.



Download ⌄ GenPept Graphics

chromosomal replication initiator protein DnaA [Ornithinimicrobium pekingense]

Sequence ID: ref|WP_022920049.1| Length: 490 Number of Matches: 1

Range 1: 3 to 490 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
610 bits(1574)	0.0	Compositional matrix adjust.	315/503(63%)	376/503(74%)	15/503(2%)

```
Query  2     SQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEG  61
             SQ+P + A  +WQ  +  L+  G+  RD   LRL  LVGLL+ TALLAV Y H K+ +E
Sbjct  3     SQSPAESA-EVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET  61

Query  62    HLREDVSTALAEVLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAV  121
              LR+ +  ALA  L  D+RLA++VD D     ++E  P AP PA      T + P  +
Sbjct  62    TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP  113

Query  122   DGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDTFVLGSSNRF  181
             D    +  G+ P   GE   P +    T + +      + + LN KYTFDTFV GSSNRF
Sbjct  114   DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDTFVSGSSNRF  167

Query  182   AHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTNQ  241
             AHAA+ AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+L   VRV+YVNSEEFTN
Sbjct  168   AHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTND  227

Query  242   FINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSD  301
             FIN++   +A AFQR+YR+VD LL+DDIQFLQGKEQT+EEFFHTFNTLHNSEKQ+VITSD
Sbjct  228   FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSD  287

Query  302   QPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI  361
             QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAIL++KAA + + +PD+VL LI SKI
Sbjct  288   QPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAILKKKAAQEGMQLPDEVLELIGSKI  347

Query  362   SSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMPGGDSGQITPTMILEETAGYFV  421
             S+NIRELEGAL RVTAFASLS +P D  LA  VLKD++P  +S  IT   I+ E A YF
Sbjct  348   STNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMAEVADYFQ  407

Query  422   ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQLL  481
             IS +++ G SRSR L  ARQIAMYLCRELTDLSLPKIG+EFGGRDHTTVMHAERKI+QL+
Sbjct  408   ISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLI  467

Query  482   GEDRRVYDEVSELTSIIRKKAAR  504
             GE R +YD+++ELT IIRK +AR
Sbjct  468   GERRALYDQITELTGIIRKASAR  490
```

nr hit #1

Figure 2.12b. An nr database pairwise alignment for Ksed_00010. Note the name of the organism, *Ornithinimicrobium pekingense* is part of the text description.

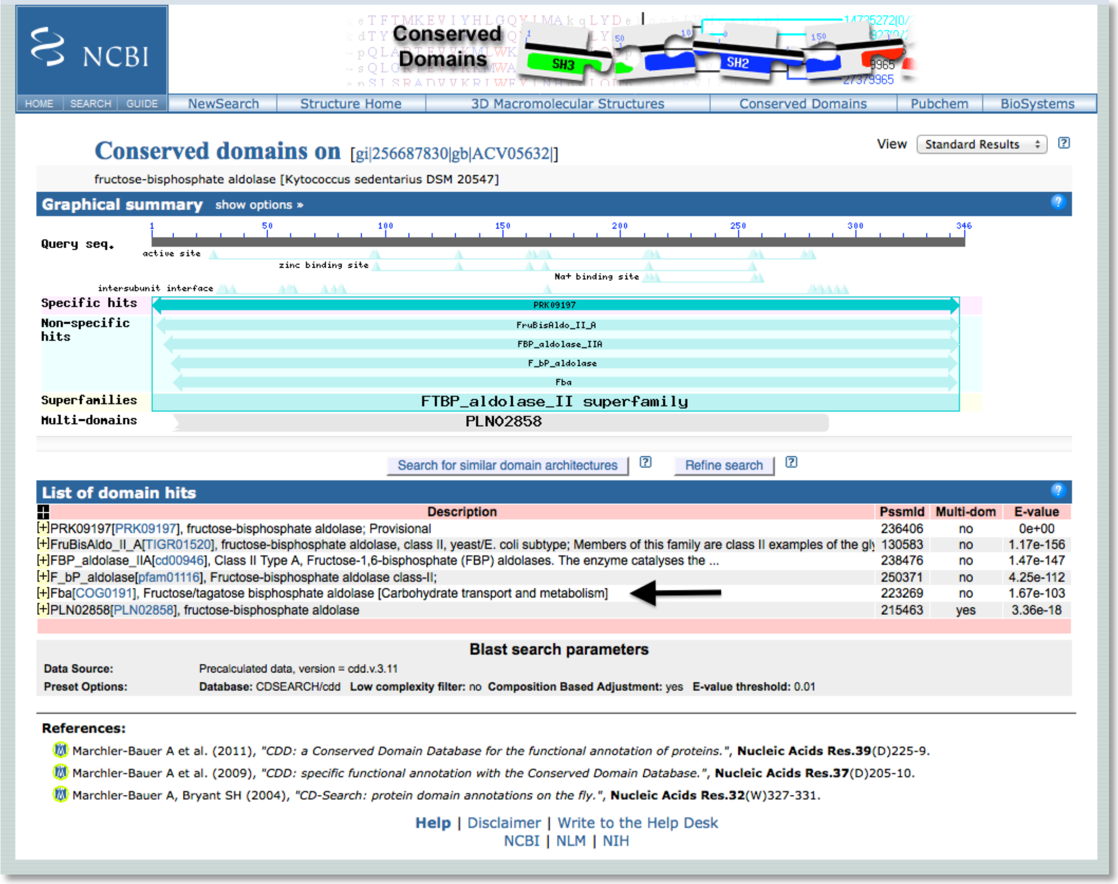
Conserved Domain Database Search (CDD):

Background: Domains in proteins refer to parts of the protein that have a particular structure or function. You can think of them as building blocks that can be put together in different ways in different proteins. If you find a particular building block in your protein that has been correlated with a structure or function in other, well annotated or curated proteins, it is strong evidence that your protein will have that structure or function. You can read more about conserved domains here: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#CDWhat.

A. The CDD search is automatically run in parallel with **ANY** NCBI BLAST search. It will be identical with either a Swiss-Prot or nr database search and thus does not need to be run in duplicate.

B. After performing a BLAST search, a graphical representation of any putative conserved domains (e.g. superfamilies, COGS, Pfams, etc), if any have been found, will be seen at the top of the Results page as shown in Figure 2.4 (arrow, CDD Results).

C. Click on this graphic to view the CDD search results page and it will take you to a page similar to the one illustrated in Figure 2.13.



Conserved domains on [gi|256687830|gb|ACV05632|]

fructose-bisphosphate aldolase [Kytococcus sedentarius DSM 20547]

List of domain hits

Description	PssmId	Multi-dom	E-value
[+]PRK09197[PRK09197], fructose-bisphosphate aldolase; Provisional	236406	no	0e+00
[+]FruBisAldo_II_A[TIGR01520], fructose-bisphosphate aldolase, class II, yeast/E. coli subtype; Members of this family are class II examples of the gl	130583	no	1.17e-156
[+]FBP_aldolase_IIA[cd00946], Class II Type A, Fructose-1,6-bisphosphate (FBP) aldolases. The enzyme catalyses the ...	238476	no	1.47e-147
[+]F_bP_aldolase[pfam01116], Fructose-bisphosphate aldolase class-II;	250371	no	4.25e-112
[+]Fba[COG0191], Fructose/tagatose bisphosphate aldolase [Carbohydrate transport and metabolism]	223269	no	1.67e-103
[+]PLN02858[PLN02858], fructose-bisphosphate aldolase	215463	yes	3.36e-18

Blast search parameters

Data Source: Precalculated data, version = cdd.v.3.11

Preset Options: **Database:** CDSEARCH/cdd **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01

References:

Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.

Marchler-Bauer A et al. (2009), *"CDD: specific functional annotation with the Conserved Domain Database."*, **Nucleic Acids Res.37**(D)205-10.

Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

Help | Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

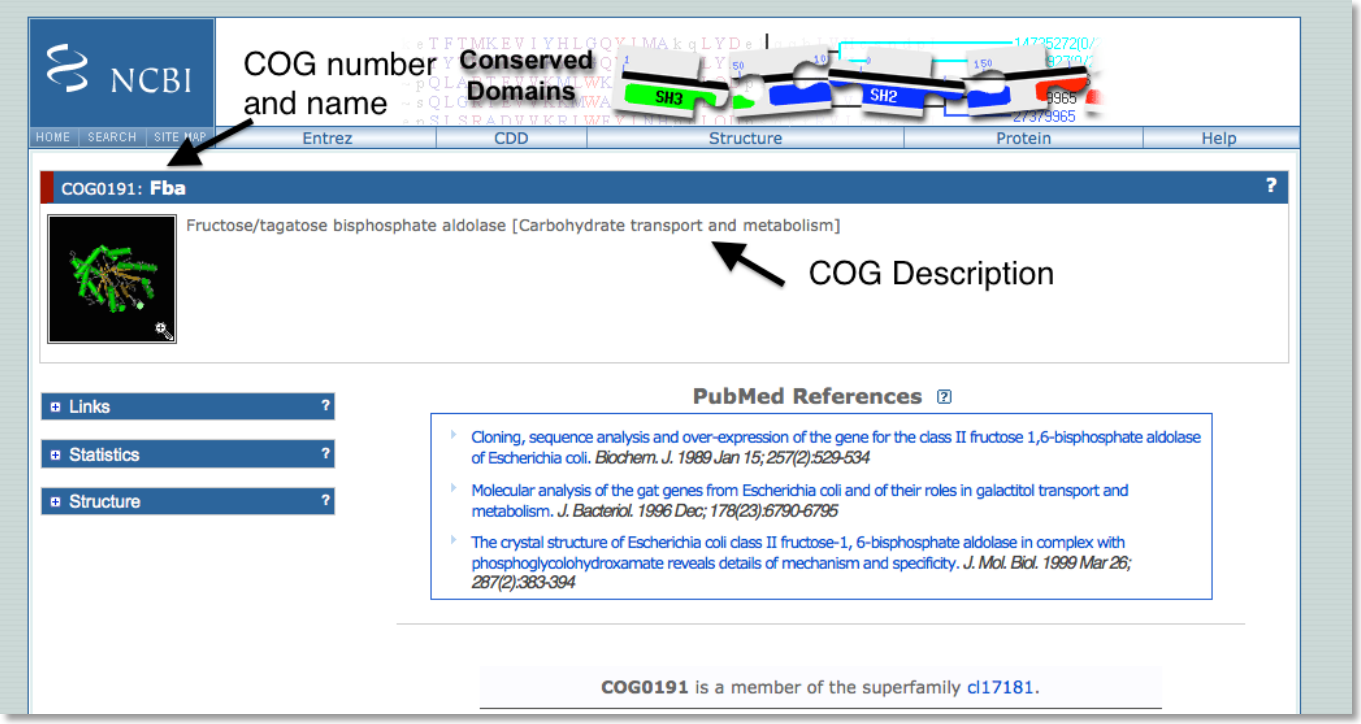
Figure 2.13. The Conserved Domain Database search results page. The arrow points to the domain that is a COG.

1. You may notice a number of different types of results in your CDD search. We are only interested in COG hits at this point in the annotation, so scan through the results to find one that has COG in its name. In the example shown in figure 2.13, there is one COG result indicated by the arrow.

2. COG is an acronym for Clusters of Orthologous Groups. COGs represent one attempt to characterize protein domains. Orthologs are proteins that are believed to be derived from a common ancestor during evolution. Such proteins will likely have a similar domain structure. If a newly discovered protein has domains in common with characterized proteins, it is good evidence that the newly discovered protein is an ortholog as well. PFAM and TIGRFam databases will be searched in the Structure Based Evidence module, and represent another way to identify protein domains. Further information about COGs can be found at the following link: http://www.ncbi.nlm.nih.gov/books/NBK21090/

3. Click on the hyperlink of the top most COG in your list of CDD results. When you do so, you will see a page similar to the one in Figure 2.14.

4. The COG Number, Name, and E-value of any significant COG hits in the lab notebook (Figures 2.15 and 2.16.



COG0191: Fba

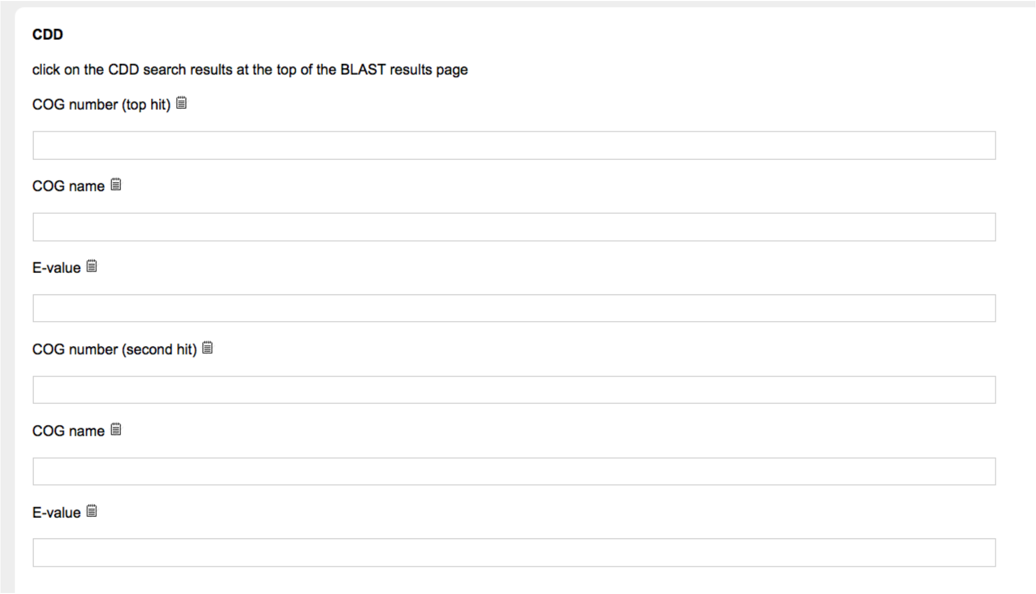
Fructose/tagatose bisphosphate aldolase [Carbohydrate transport and metabolism]

PubMed References

- Cloning, sequence analysis and over-expression of the gene for the class II fructose 1,6-bisphosphate aldolase of Escherichia coli. *Biochem. J.* 1989 Jan 15;257(2):529-534
- Molecular analysis of the gat genes from Escherichia coli and of their roles in galactitol transport and metabolism. *J. Bacteriol.* 1996 Dec;178(23):6790-6795
- The crystal structure of Escherichia coli class II fructose-1, 6-bisphosphate aldolase in complex with phosphoglycolohydroxamate reveals details of mechanism and specificity. *J. Mol. Biol.* 1999 Mar 26;287(2):383-394

COG0191 is a member of the superfamily cl17181.

Figure 2.14. The COG description page. The COG name and number are indicated along with the description of the COG.



CDD

click on the CDD search results at the top of the BLAST results page

COG number (top hit)

COG name

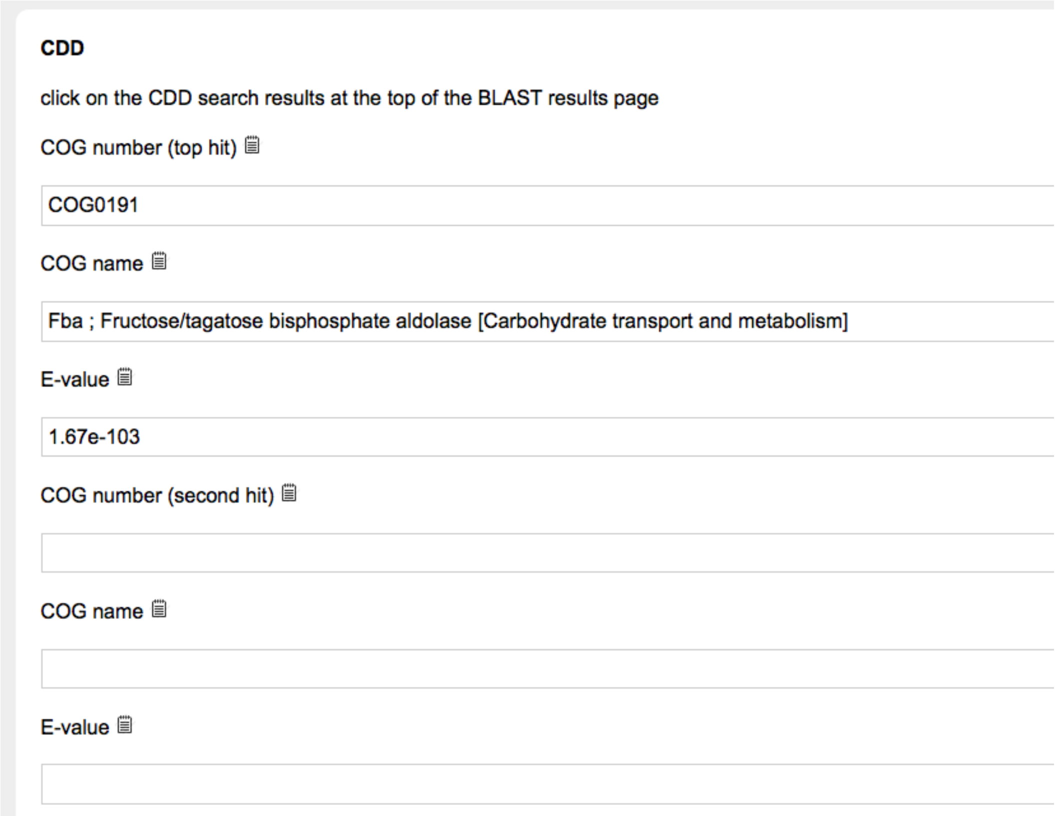
E-value

COG number (second hit)

COG name

E-value

Figure 2.15. The CDD notebook page. Spaces are available for data from up to 2 COG hits.



CDD

click on the CDD search results at the top of the BLAST results page

COG number (top hit)

COG0191

COG name

Fba ; Fructose/tagatose bisphosphate aldolase [Carbohydrate transport and metabolism]

E-value

1.67e-103

COG number (second hit)

COG name

E-value

Figure 2.16. The CDD notebook page populated with data. The COG number, name (and description) and E-value were obtained from figures 2.13 and 2.14. Only one COG hit was obtained, so the second hit information is blank.

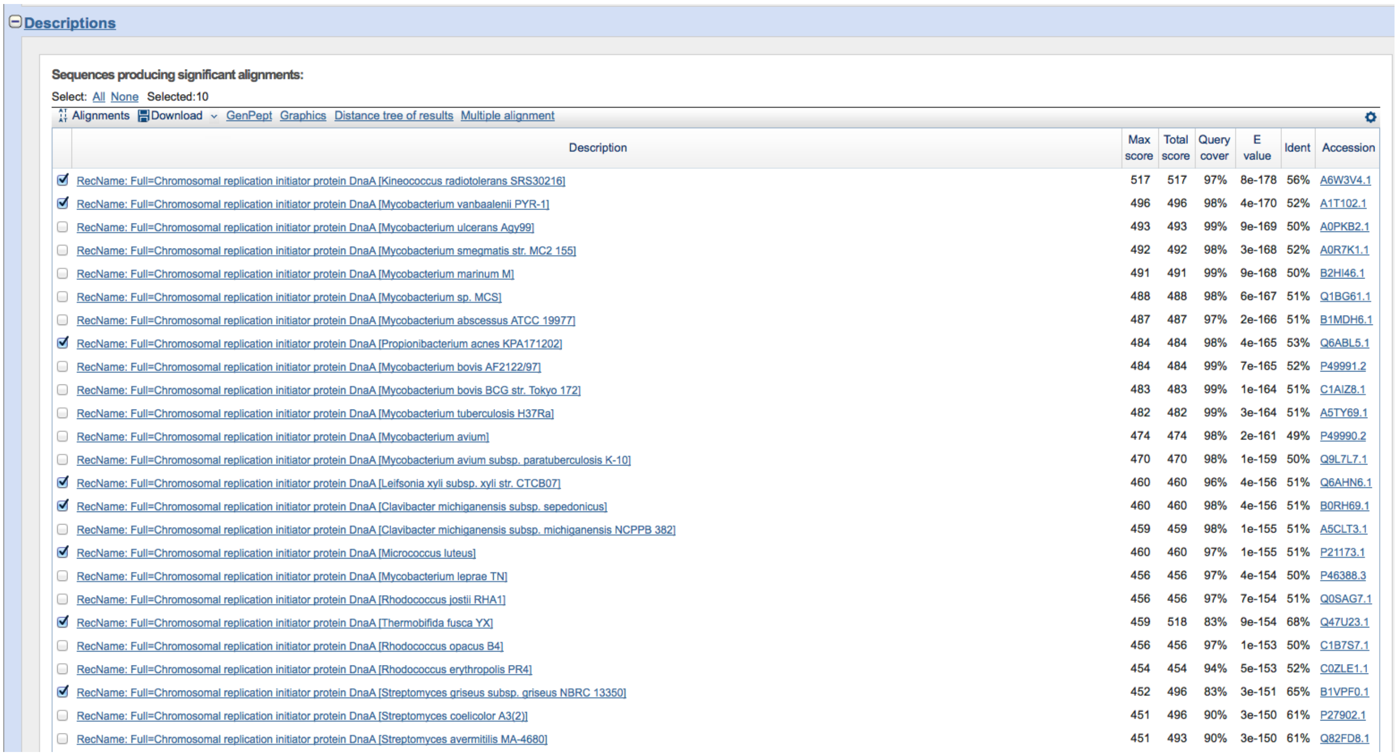
5. The results from the CDD search should be interpreted as described for the BLAST. A low E-value is taken to represent significance.
    a. Compare the CDD results to those that you obtained from BLAST. The COG should make sense based on the name you determined for your protein in BLAST
    b. Some proteins have more than one domain and thus may have more than one COG hit. In the event that you have more than one hit, fill out the information for the second hit in the boxes provided. You can add more data for a third or fourth significant COG hit to the notebook yourself should you have more than two.
    c. Some proteins (particularly hypothetical proteins) may not have a COG hit. In the event that you do not obtain a significant COG hit you should write "no significant COG hits found" in the COG number box of your notebook.
6. CDD search can also be performed manually using the link http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml. Search by entering the protein sequence in FASTA format. This procedure may not display as many hits as the above method.

The next two sections of your annotation will allow you to directly visualize how well the hits you identified by BLAST align with your query sequence and allow you to see the level of homology along the length of the match in a very straightforward way

Tree Based Objective Function for Alignment Evaluation (T-Coffee)

7. **On the nr BLAST results page**, Scroll down to the list of best hits. Select 10-15 of the top orthologs with significant E-values by checking the box next to the selection. Orthologs are proteins that share similarity with your protein, but which are found in a different organism. You may occasionally find paralogs as well. Paralogs are proteins with similarity to your query that are found in the same organism. You may wonder why your bacterium would have variant forms of a particular protein, and we will explore reasons why this might be so in a later module. Do NOT select any paralog entries for this module

   A. Make sure the orthologs you pick are not just the first 10 in the list. Look at the species name in the column labeled "Description" and try to pick significant hits from 10 different organisms (Figure 2.17). Sometimes you will have different strains of the same organism appearing multiple times in at the top of the list or lots of members of the same genus appearing in the list. A strain represents a slight variant of an organism, but their genomes are generally very similar. In order to see where homology in the alignment is best preserved it is better to select proteins from different species for comparison. In the example shown in figure 2.18, note that the 3rd through the 7th sequences in the list were skipped over because a *Mycobacterium* sequence was chosen as the 2nd sequence in the list and the 3rd -10th hits were different species of *Mycobacterium*.



Descriptions

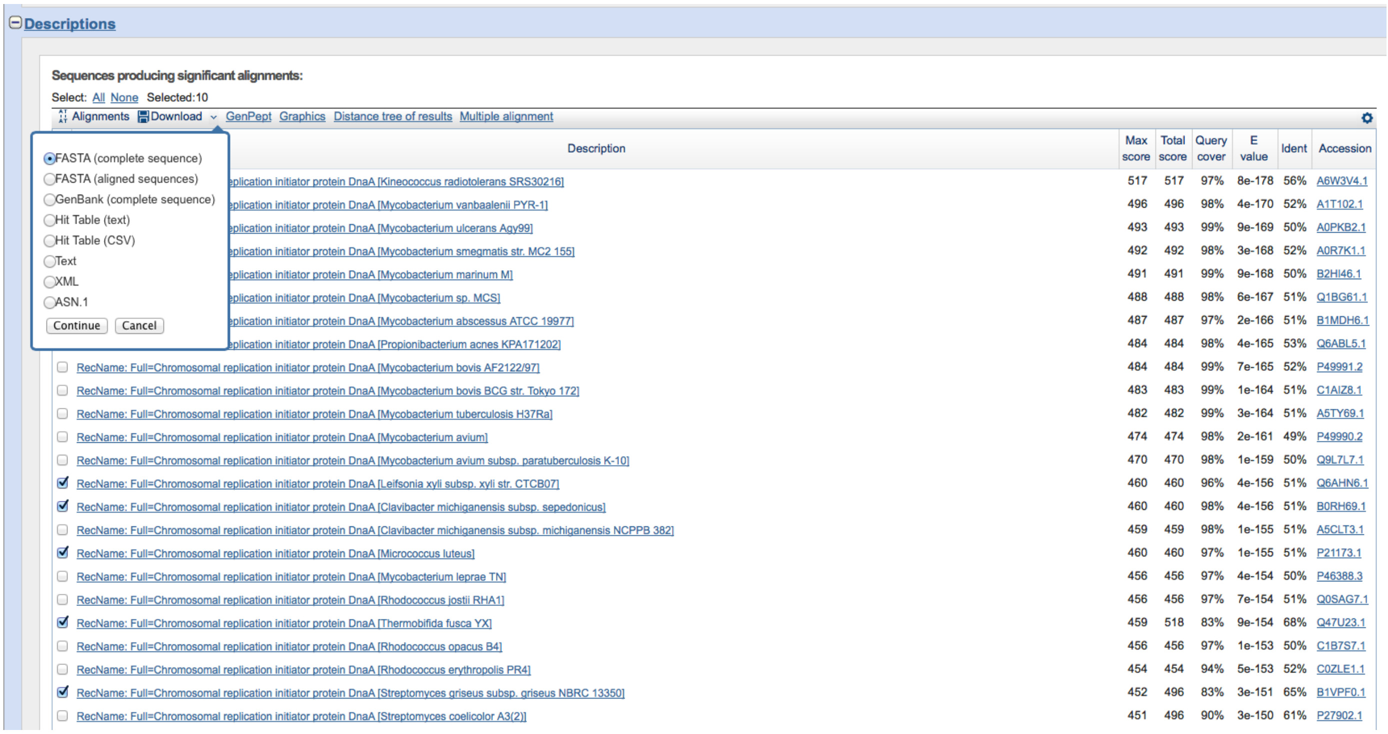
Sequences producing significant alignments:

Select: All None Selected:10

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Kineococcus radiotolerans SRS30216]	517	517	97%	8e-178	56%	A6W3V4.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium vanbaalenii PYR-1]	496	496	98%	4e-170	52%	A1T102.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium ulcerans Agy99]	493	493	99%	9e-169	50%	A0PKB2.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium smegmatis str. MC2 155]	492	492	98%	3e-168	52%	A0R7K1.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium marinum M]	491	491	99%	9e-168	50%	B2HI46.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium sp. MCS]	488	488	98%	6e-167	51%	Q1BG61.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium abscessus ATCC 19977]	487	487	97%	2e-166	51%	B1MDH6.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Propionibacterium acnes KPA171202]	484	484	98%	4e-165	53%	Q6ABL5.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium bovis AF2122/97]	484	484	99%	7e-165	52%	P49991.2
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium bovis BCG str. Tokyo 172]	483	483	99%	1e-164	51%	C1AIZ8.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium tuberculosis H37Ra]	482	482	99%	3e-164	51%	A5TY69.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium avium]	474	474	98%	2e-161	49%	P49990.2
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium avium subsp. paratuberculosis K-10]	470	470	98%	1e-159	50%	Q9L7L7.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Leifsonia xyli subsp. xyli str. CTCB07]	460	460	96%	4e-156	51%	Q6AHN6.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Clavibacter michiganensis subsp. sepedonicus]	460	460	98%	4e-156	51%	B0RH69.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Clavibacter michiganensis subsp. michiganensis NCPPB 382]	459	459	98%	1e-155	51%	A5CLT3.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Micrococcus luteus]	460	460	97%	1e-155	51%	P21173.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium leprae TN]	456	456	97%	4e-154	50%	P46388.3
□	RecName: Full=Chromosomal replication initiator protein DnaA [Rhodococcus jostii RHA1]	456	456	97%	7e-154	51%	Q0SAG7.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Thermobifida fusca YX]	459	518	83%	9e-154	68%	Q47U23.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Rhodococcus opacus B4]	456	456	97%	1e-153	50%	C1B7S7.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Rhodococcus erythropolis PR4]	454	454	94%	5e-153	52%	C0ZLE1.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Streptomyces griseus subsp. griseus NBRC 13350]	452	496	83%	3e-151	65%	B1VPF0.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Streptomyces coelicolor A3(2)]	451	496	90%	3e-150	61%	P27902.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Streptomyces avermitilis MA-4680]	451	493	90%	3e-150	61%	Q82FD8.1

Figure 2.17. The BLAST top hits for Ksed_00010. Note the mix of genus and species that are checked (only 8 are shown in this image, but you should check 10-15 as noted in the text.

8. After you have made your selections, click the Download pull down menu at the top of the page and make sure the FASTA (complete sequence) radio button is checked as shown in figure 2.18.



Descriptions

Sequences producing significant alignments:

Select: All None Selected:10

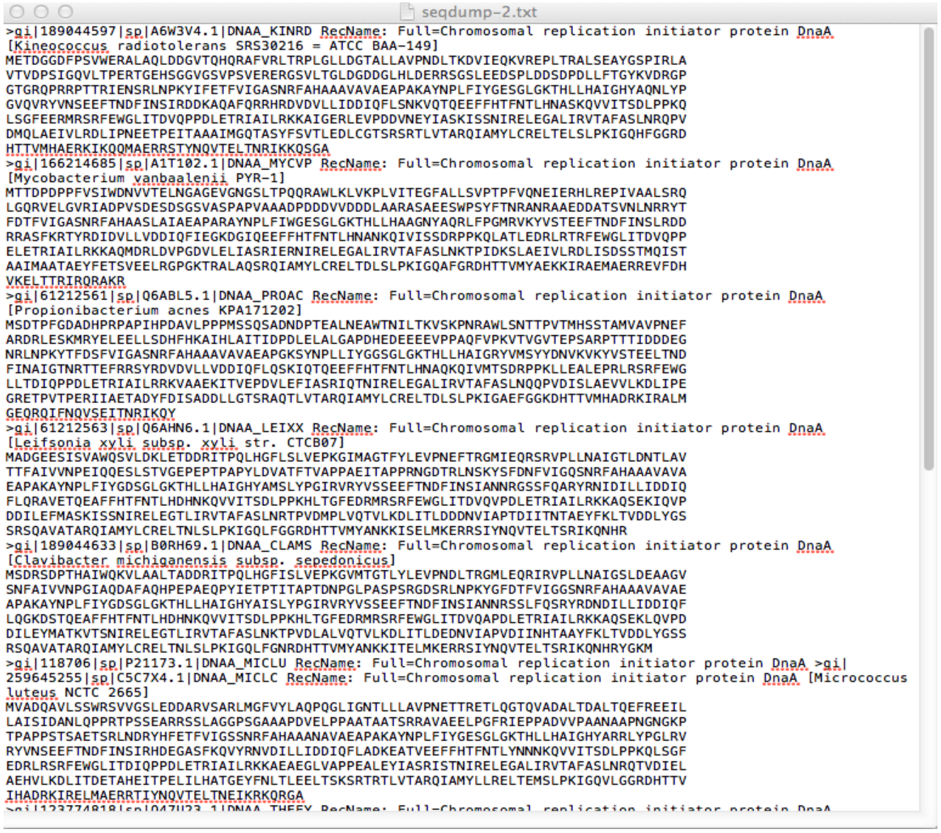
● FASTA (complete sequence)
○ FASTA (aligned sequences)
○ GenBank (complete sequence)
○ Hit Table (text)
○ Hit Table (CSV)
○ Text
○ XML
○ ASN.1

	Description	Max score	Total score	Query cover	E value	Ident	Accession
	eplication initiator protein DnaA [Kineococcus radiotolerans SRS30216]	517	517	97%	8e-178	56%	A6W3V4.1
	eplication initiator protein DnaA [Mycobacterium vanbaalenii PYR-1]	496	496	98%	4e-170	52%	A1T102.1
	eplication initiator protein DnaA [Mycobacterium ulcerans Agy99]	493	493	99%	9e-169	50%	A0PKB2.1
	eplication initiator protein DnaA [Mycobacterium smegmatis str. MC2 155]	492	492	98%	3e-168	52%	A0R7K1.1
	eplication initiator protein DnaA [Mycobacterium marinum M]	491	491	99%	9e-168	50%	B2HI46.1
	eplication initiator protein DnaA [Mycobacterium sp. MCS]	488	488	98%	6e-167	51%	Q1BG61.1
	eplication initiator protein DnaA [Mycobacterium abscessus ATCC 19977]	487	487	97%	2e-166	51%	B1MDH6.1
	eplication initiator protein DnaA [Propionibacterium acnes KPA171202]	484	484	98%	4e-165	53%	Q6ABL5.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium bovis AF2122/97]	484	484	99%	7e-165	52%	P49991.2
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium bovis BCG str. Tokyo 172]	483	483	99%	1e-164	51%	C1AIZ8.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium tuberculosis H37Ra]	482	482	99%	3e-164	51%	A5TY69.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium avium]	474	474	98%	2e-161	49%	P49990.2
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium avium subsp. paratuberculosis K-10]	470	470	98%	1e-159	50%	Q9L7L7.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Leifsonia xyli subsp. xyli str. CTCB07]	460	460	96%	4e-156	51%	Q6AHN6.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Clavibacter michiganensis subsp. sepedonicus]	460	460	98%	4e-156	51%	B0RH69.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Clavibacter michiganensis subsp. michiganensis NCPPB 382]	459	459	98%	1e-155	51%	A5CLT3.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Micrococcus luteus]	460	460	97%	1e-155	51%	P21173.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Mycobacterium leprae TN]	456	456	97%	4e-154	50%	P46388.3
□	RecName: Full=Chromosomal replication initiator protein DnaA [Rhodococcus jostii RHA1]	456	456	97%	7e-154	51%	Q0SAG7.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Thermobifida fusca YX]	459	518	83%	9e-154	68%	Q47U23.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Rhodococcus opacus B4]	456	456	97%	1e-153	50%	C1B7S7.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Rhodococcus erythropolis PR4]	454	454	94%	5e-153	52%	C0ZLE1.1
☑	RecName: Full=Chromosomal replication initiator protein DnaA [Streptomyces griseus subsp. griseus NBRC 13350]	452	496	83%	3e-151	65%	B1VPF0.1
□	RecName: Full=Chromosomal replication initiator protein DnaA [Streptomyces coelicolor A3(2)]	451	496	90%	3e-150	61%	P27902.1

Figure 2.18. Preparing to download the 10 selected from the list of BJAST hits. Note the FASTA (complete sequence) radio button is selected.

9. Click the continue button shown in the pull down menu of figure 2.18 to download the sequences you selected. You will then get a list of sequences with FASTA headers separating each one as shown in figure 2.19 (if you use notepad on a PC the sequences may look different at this point).

10. Copy and paste the sequences into your notebook in the “Sequences used for alignment” section of your notebook (Figure 2.20). These are the sequences that you will used for generating a multiple alignment to see how well the amino acids match for all of the proteins you selected. Be sure to save the notebook after pasting.



```
>gi|189044597|sp|A6W3V4.1|DNAA_KINRD RecName: Full=Chromosomal replication initiator protein DnaA
[Kineococcus radiotolerans SRS30216 = ATCC BAA-149]
METDGGDFPSVWERALAQLDDGVTQHQRAFVRLTRPLGLLDGTALLAVPNDLTKDVIEQKVREPLTRALSEAYGSPIRLA
VTVDPSIGQVLTPERTGEHSGGVGSVPSVERERGSVLTGLDGDDGLHLDERRSGSLEEDSPLDDSDPDLLFTGYKVDRGP
GTGRQPRRPTTRIENSRLNPKYIFETFVIGASNRFAHAAAVAVAEAPAKAYNPLFIYGESGLGKTHLLHAIGHYAQNLYP
GVQVRYVNSEEFTNDFINSIRDDKAQAFQRRHRDVDVLLIDDIQFLSNKVQTQEEFFHTFNTLHNASKQVVITSDLPPKQ
LSGFEERMRSRFEWGLITDVQPPDLETRIAILRKKAIGERLEVPDDVNEYIASKISSNIRELEGALIRVTAFASLNRQPV
DMQLAEIVLRDLIPNEETPEITAAAIMGQTASYFSVTLEDLCGTSRSRTLVTARQIAMYLCRELTELSLPKIGQHFGGRD
HTTVMHAERKIKQQMAERRSTYNQVTELTNRIKKQSGA
>gi|166214685|sp|A1T102.1|DNAA_MYCVP RecName: Full=Chromosomal replication initiator protein DnaA
[Mycobacterium vanbaalenii PYR-1]
MTTDPDPPFVSIWDNVVTELNGAGEVGNGSLTPQQRAWLKLVKPLVITEGFALLSVPTPFVQNEIERHLREPIVAALSRQ
LGQRVELGVRIADPVSDESDSGSVASPAPVAAADPDDDVVDDDLAARASAEESWPSYFTNRANRAAEDDATSVNLNRRYT
FDTFVIGASNRFAHAASLAIAEAPARAYNPLFIWGESGLGKTHLLHAAGNYAQRLFPGMRVKYVSTEEFTNDFINSLRDD
RRASFKRTYRDIDVLLVDDIQFIEGKDGIQEEFFHTFNTLHNANKQIVISSDRPPKQLATLEDRLRTRFEWGLITDVQPP
ELETRIAILRKKAQMDRLDVPGDVLELIASRIERNIRELEGALIRVTAFASLNKTPIDKSLAEIVLRDLISDSSTMQIST
AAIMAATAEYFETSVEELRGPGKTRALAQSRQIAMYLCRELTDLSLPKIGQAFGRDHTTVMYAEKKIRAEMAERREVFDH
VKELTTRIRQRAKR
>gi|61212561|sp|Q6ABL5.1|DNAA_PROAC RecName: Full=Chromosomal replication initiator protein DnaA
[Propionibacterium acnes KPA171202]
MSDTPFGDADHPRPAPIHPDAVLPPPMSSQSADNDPTEALNEAWTNILTKVSKPNRAWLSNTTPVTMHSSTAMVAVPNEF
ARDRLESKMRYELEELLSDHFHKAIHLAITIDPDLELALGAPDHEDEEEEVPPAQFVPKVTVGVTEPSARPTTTIDDDEG
NRLNPKYTFDSFVIGASNRFAHAAAVAVAEAPGKSYNPLLIYGGSGLGKTHLLHAIGRYVMSYYDNVKVKYVSTEELTND
FINAIGTNRTTEFRRSYRDVDVLLVDDIQFLQSKIQTQEEFFHTFNTLHNAQKQIVMTSDRPPKLLEALEPRLRSRFEWG
LLTDIQPPDLETRIAILRRKVAAEKITVEPDVLEFIASRIQTNIRELEGALIRVTAFASLNQQPVDISLAEVVLKDLIPE
GRETPVTPERIIAETADYFDISADDLLGTSRAQTLVTARQIAMYLCRELTDLSLPKIGAEFGGKDHTTVMHADRKIRALM
GEQRQIFNQVSEITNRIKQY
>gi|61212563|sp|Q6AHN6.1|DNAA_LEIXX RecName: Full=Chromosomal replication initiator protein DnaA
[Leifsonia xyli subsp. xyli str. CTCB07]
MADGEESISVAWQSVLDKLETDDRITPQLHGFLSLVEPKGIMAGTFYLEVPNEFTRGMIEQRSRVPLLNAIGTLDNTLAV
TTFAIVVNPEIQQESLSTVGEPEPTPAPYLDVATFTVAPPAEITAPPRNGDTRLNSKYSFDNFVIGQSNRFAHAAAVAVA
EAPAKAYNPLFIYGDSGLGKTHLLHAIGHYAMSLYPGIRVRYVSSEEFTNDFINSIANNRGSSFQARYRNIDILLIDDIQ
FLQRAVETQEAFFHTFNTLHDHNKQVVITSDLPPKHLTGFEDRMRSRFEWGLITDVQVPDLETRIAILRKKAQSEKIQVP
DDILEFMASKISSNIRELEGTLIRVTAFASLNRTPVDMPLVQTVLKDLITLDDDNVIAPTDIITNTAEYFKLTVDDLYGS
SRSQAVATARQIAMYLCRELTNLSLPKIGQLFGGRDHTTVMYANKKISELMKERRSIYNQVTELTSRIKQNHR
>gi|189044633|sp|B0RH69.1|DNAA_CLAMS RecName: Full=Chromosomal replication initiator protein DnaA
[Clavibacter michiganensis subsp. sepedonicus]
MSDRSDPTHAIWQKVLAALTADDRITPQLHGFISLVEPKGVMTGTLYLEVPNDLTRGMLEQRIRVPLLNAIGSLDEAAGV
SNFAIVVNPGIAQDAFAQHPEPAEQPYIETPTITAPTDNPGLPASPSRGDSRLNPKYGFDTFVIGGSNRFAHAAAVAVAE
APAKAYNPLFIYGDSGLGKTHLLHAIGHYAISLYPGIRVRYVSSEEFTNDFINSIANNRSSLFQSRYRDNDILLIDDIQF
LQGKDSTQEAFFHTFNTLHDHNKQVVITSDLPPKHLTGFEDRMRSRFEWGLITDVQAPDLETRIAILRKKAQSEKLQVPD
DILEYMATKVTSNIRELEGTLIRVTAFASLNKTPVDLALVQTVLKDLITLDEDNVIAPVDIINHTAAYFKLTVDDLYGSS
RSQAVATARQIAMYLCRELTNLSLPKIGQLFGNRDHTTVMYANKKITELMKERRSIYNQVTELTSRIKQNHRYGKM
>gi|118706|sp|P21173.1|DNAA_MICLU RecName: Full=Chromosomal replication initiator protein DnaA >gi|
259645255|sp|C5C7X4.1|DNAA_MICLC RecName: Full=Chromosomal replication initiator protein DnaA [Micrococcus
luteus NCTC 2665]
MVADQAVLSSWRSVVGSLEDDARVSARLMGFVYLAQPQGLIGNTLLLAVPNETTRETLQGTQVADALTDALTQEFREEIL
LAISIDANLQPPRTPSSEARRSSLAGGPSGAAAPDVELPPAATAATSRRAVAEELPGFRIEPPADVVPAANAAPNGNGKP
TPAPPSTSAETSRLNDRYHFETFVIGSSNRFAHAAANAVAEAPAKAYNPLFIYGESGLGKTHLLHAIGHYARRLYPGLRV
RYVNSEEFTNDFINSIRHDEGASFKQVYRNVDILLIDDIQFLADKEATVEEFFHTFNTLYNNNKQVVITSDLPPKQLSGF
EDRLRSRFEWGLITDIQPPDLETRIAILRKKAEAEGLVAPPEALEYIASRISTNIRELEGALIRVTAFASLNRQTVDIEL
AEHVLKDLITDETAHEITPELILHATGEYFNLTLEELTSKSRTRTLVTARQIAMYLLRELTEMSLPKIGQVLGGRDHTTV
IHADRKIRELMAERRTIYNQVTELTNEIKRKQRGA
>gi|123774818|sp|Q47U23.1|DNAA_THEFY RecName: Full=Chromosomal replication initiator protein DnaA
```

Figure 2.19. The FASTA formatted amino acid sequence download.

11. We will want to have the gene you are working on included in the alignment, so it will need to be added to the list of BLAST hits. This can be done in one of two ways.

A. The first is to simply select the top hit in the nr database search along with the orthologs that you choose. Remember, the top hit in the nr database search is often your own sequence. If you are convinced this is the case (100% query coverage and 100% identity with an E-value of essentially 0), all you have to do is edit the FASTA header to match that of your protein sequence in the basic information section of your notebook (Module 1).

B. If the top nr hit is NOT a perfect match to your sequence (which will be a rare event, there is a second way to include your sequence in the alignment.

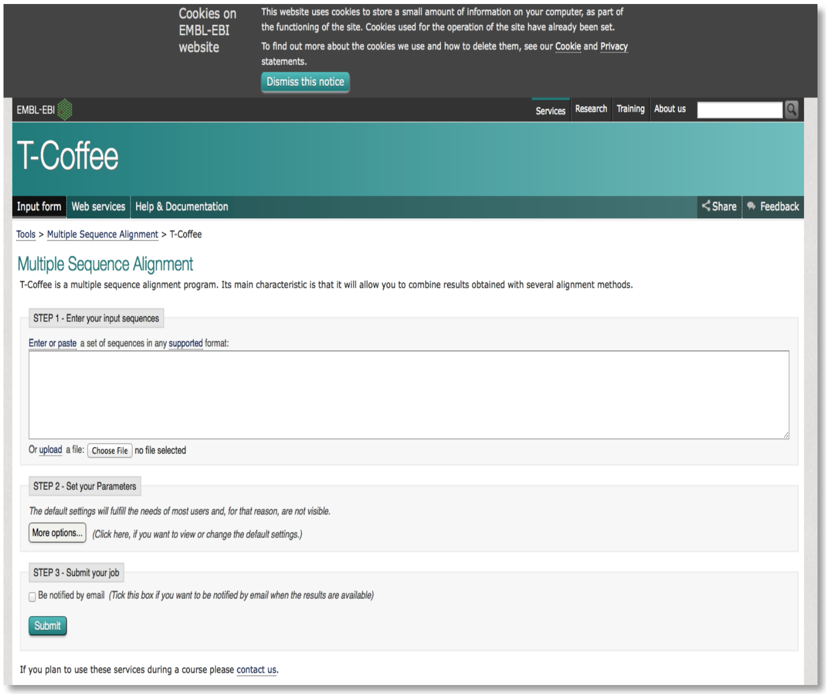
1. To do this, Open the Basic Information Module and copy the FASTA formatted amino acid sequence of the protein encoded by your gene.

2. Return to the T-Coffee notebook and click on the edit icon in the section you just pasted the sequences for alignment.

3. Insert the cursor in front to the first sequence FASTA header and hit return to create a space. Then paste your FASTA formatted amino acid sequence into the notebook so that it is the first sequence at the top (Figure 2.21) and hit save.

12. There will now be a total of 11 sequences in your notebook (the amino acid sequence of the protein under investigation and the 10 that were selected to perform the multiple alignment).

13. Select all 11 of the FASTA formatted sequences and copy them



This website uses cookies to store a small amount of information on your computer, as part of the functioning of the site. Cookies used for the operation of the site have already been set.

To find out more about the cookies we use and how to delete them, see our Cookie and Privacy statements.

Tools > Multiple Sequence Alignment > T-Coffee

Multiple Sequence Alignment

T-Coffee is a multiple sequence alignment program. Its main characteristic is that it will allow you to combine results obtained with several alignment methods.

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Or upload a file: no file selected

STEP 2 - Set your Parameters

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

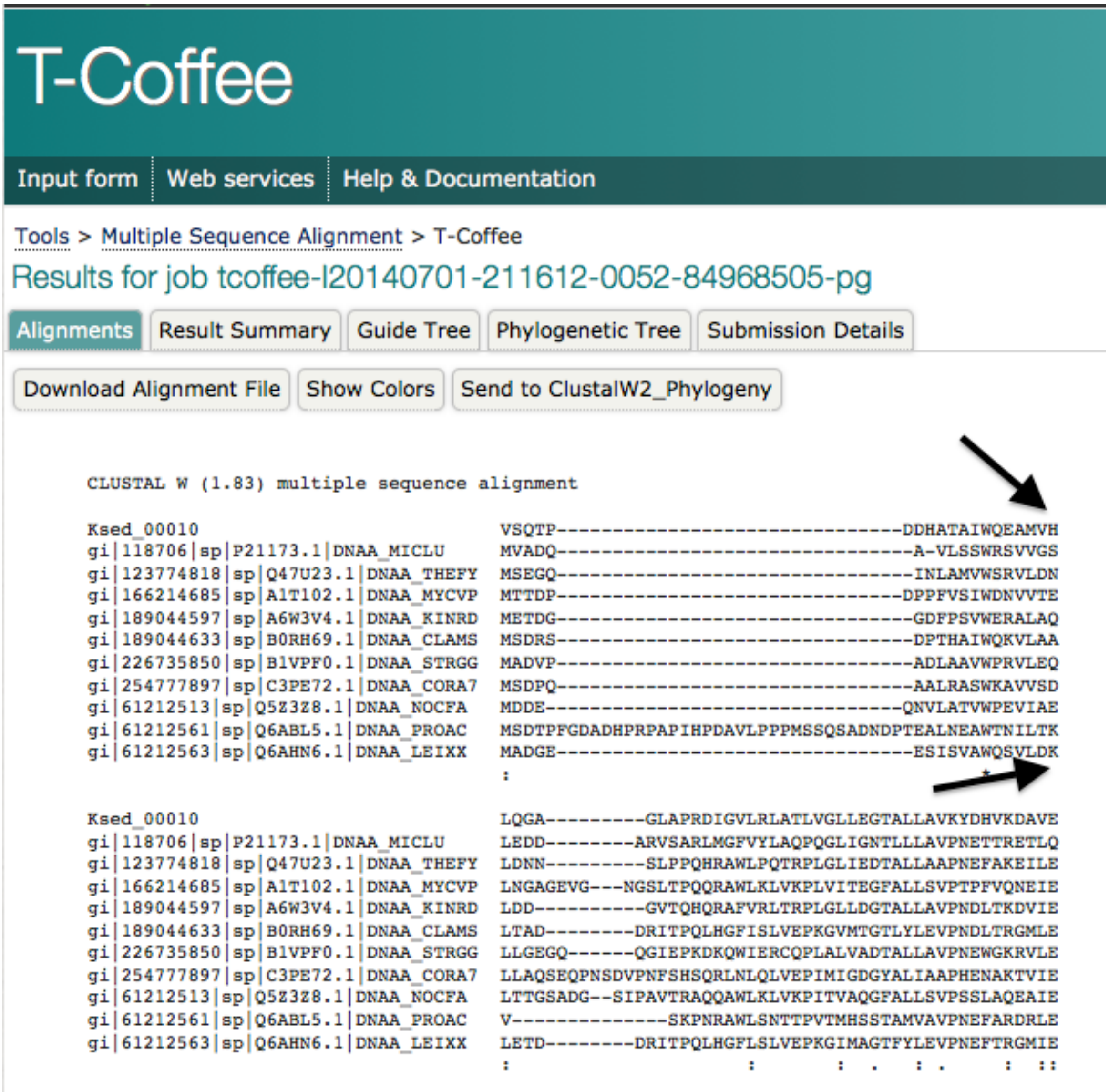
□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

If you plan to use these services during a course please contact us.

←Figure 2.21. The T-COFFEE start page.

14. Got to EBI's T-Coffee server at http://www.ebi.ac.uk/Tools/msa/tcoffee/ (Figure 2.21), paste all 11 sequences into the input window and click submit.

15. A results page with CLUSTAL FORMAT for T-Coffee is seen as show in Figure 2.22.



Results for job tcoffee-I20140701-211612-0052-84968505-pg

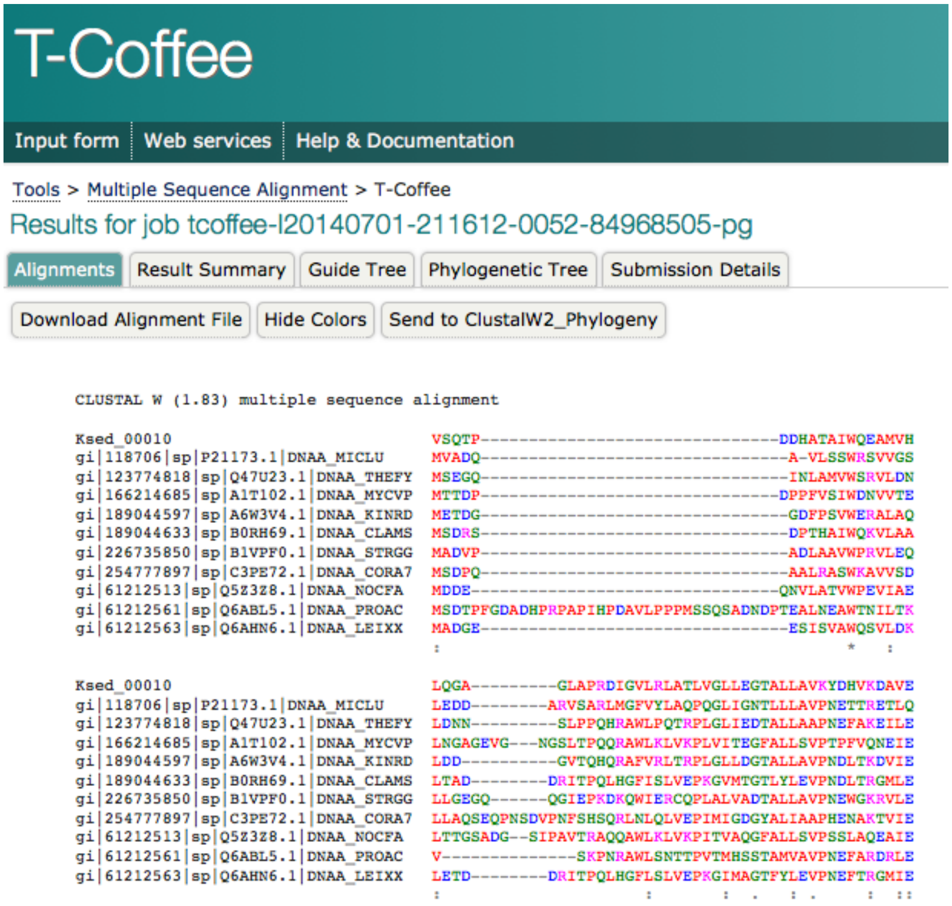
```
CLUSTAL W (1.83) multiple sequence alignment

Ksed_00010                              VSQTP-------------------------------DDHATAIWQEAMVH
gi|118706|sp|P21173.1|DNAA_MICLU        MVADQ--------------------------------A-VLSSWRSVVGS
gi|123774818|sp|Q47U23.1|DNAA_THEFY     MSEGQ--------------------------------INLAMVWSRVLDN
gi|166214685|sp|A1T102.1|DNAA_MYCVP     MTTDP-------------------------------DPPFVSIWDNVVTE
gi|189044597|sp|A6W3V4.1|DNAA_KINRD     METDG--------------------------------GDFPSVWERALAQ
gi|189044633|sp|B0RH69.1|DNAA_CLAMS     MSDRS--------------------------------DPTHAIWQKVLAA
gi|226735850|sp|B1VPF0.1|DNAA_STRGG     MADVP--------------------------------ADLAAVWPRVLEQ
gi|254777897|sp|C3PE72.1|DNAA_CORA7     MSDPQ--------------------------------AALRASWKAVVSD
gi|61212513|sp|Q5Z3Z8.1|DNAA_NOCFA      MDDE--------------------------------QNVLATVWPEVIAE
gi|61212561|sp|Q6ABL5.1|DNAA_PROAC      MSDTPFGDADHPRPAPIHPDAVLPPPMSSQSADNDPTEALNEAWTNILTK
gi|61212563|sp|Q6AHN6.1|DNAA_LEIXX      MADGE--------------------------------ESISVAWQSVLDK
                                        :                                          *

Ksed_00010                              LQGA---------GLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVE
gi|118706|sp|P21173.1|DNAA_MICLU        LEDD--------ARVSARLMGFVYLAQPQGLIGNTLLLAVPNETTRETLQ
gi|123774818|sp|Q47U23.1|DNAA_THEFY     LDNN---------SLPPQHRAWLPQTRPLGLIEDTALLAAPNEFAKEILE
gi|166214685|sp|A1T102.1|DNAA_MYCVP     LNGAGEVG---NGSLTPQQRAWLKLVKPLVITEGFALLSVPTPFVQNEIE
gi|189044597|sp|A6W3V4.1|DNAA_KINRD     LDD----------GVTQHQRAFVRLTRPLGLLDGTALLAVPNDLTKDVIE
gi|189044633|sp|B0RH69.1|DNAA_CLAMS     LTAD--------DRITPQLHGFISLVEPKGVMTGTLYLEVPNDLTRGMLE
gi|226735850|sp|B1VPF0.1|DNAA_STRGG     LLGEGQ------QGIEPKDKQWIERCQPLALVADTALLAVPNEWGKRVLE
gi|254777897|sp|C3PE72.1|DNAA_CORA7     LLAQSEQPNSDVPNFSHSQRLNLQLVEPIMIGDGYALIAAPHENAKTVIE
gi|61212513|sp|Q5Z3Z8.1|DNAA_NOCFA      LTTGSADG--SIPAVTRAQQAWLKLVKPITVAQGFALLSVPSSLAQEAIE
gi|61212561|sp|Q6ABL5.1|DNAA_PROAC      V--------------SKPNRAWLSNTTPVTMHSSTAMVAVPNEFARDRLE
gi|61212563|sp|Q6AHN6.1|DNAA_LEIXX      LETD--------DRITPQLHGFLSLVEPKGIMAGTFYLEVPNEFTRGMIE
                                        :                     :       :  .   : .     :  ::
```

Figure 2.22. The T-COFFEE results page. Arrows indicate the most 50 amino terminal amino acids of the alignment of the 11 sequences in this example. Note that Ksed_00010 is the first sequence in the alignment) See the text for additional explanation.

A. You will see repeating blocks of 50 amino acid stretches of the on the page. The arrows in figure 2.24 show the first 50 amino acid stretch of the example alignment. The batch of sequences below the first alignment include amino acids 51-100 of the alignment, the next batch 101-150 and so on.

B. **The following symbols may be seen In the row below the last sequence of each section of the multiple alignment:**

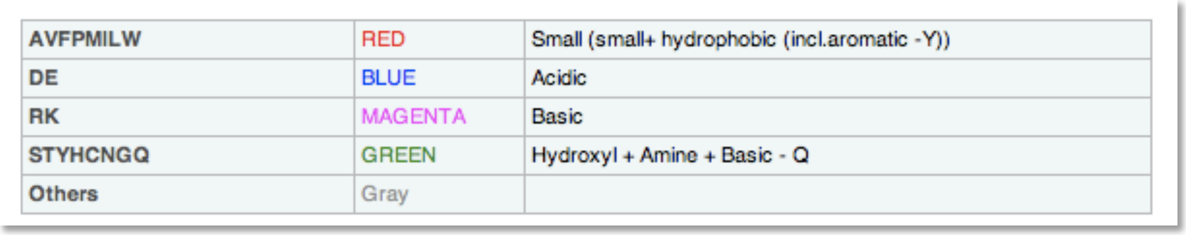
1. * → the residues or nucleotides in that column are identical in all sequences
2. : → conserved amino substitutions have been observed, according to the color data discussed below
3. . → semi-conserved amino acid substitutions are observed



```
CLUSTAL W (1.83) multiple sequence alignment

Ksed_00010                           VSQTP-------------------------------DDHATAIWQEAMVH
gi|118706|sp|P21173.1|DNAA_MICLU     MVADQ--------------------------------A-VLSSWRSVVGS
gi|123774818|sp|Q47U23.1|DNAA_THEFY  MSEGQ--------------------------------INLAMVWSRVLDN
gi|166214685|sp|A1T102.1|DNAA_MYCVP  MTTDP-------------------------------DPPFVSIWDNVVTE
gi|189044597|sp|A6W3V4.1|DNAA_KINRD  METDG--------------------------------GDFPSVWERALAQ
gi|189044633|sp|B0RH69.1|DNAA_CLAMS  MSDRS--------------------------------DPTHAIWQKVLAA
gi|226735850|sp|B1VPF0.1|DNAA_STRGG  MADVP--------------------------------ADLAAVWPRVLEQ
gi|254777897|sp|C3PE72.1|DNAA_CORA7  MSDPQ--------------------------------AALRASWKAVVSD
gi|61212513|sp|Q5Z3Z8.1|DNAA_NOCFA   MDDE--------------------------------QNVLATVWPEVIAE
gi|61212561|sp|Q6ABL5.1|DNAA_PROAC   MSDTPFGDADHPRPAPIHPDAVLPPPMSSQSADNDPTEALNEAWTNILTK
gi|61212563|sp|Q6AHN6.1|DNAA_LEIXX   MADGE--------------------------------ESISVAWQSVLDK
                                     :                                          *   :

Ksed_00010                           LQGA---------GLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVE
gi|118706|sp|P21173.1|DNAA_MICLU     LEDD--------ARVSARLMGFVYLAQPQGLIGNTLLLAVPNETTRETLQ
gi|123774818|sp|Q47U23.1|DNAA_THEFY  LDNN---------SLPPQHRAWLPQTRPLGLIEDTALLAAPNEFAKEILE
gi|166214685|sp|A1T102.1|DNAA_MYCVP  LNGAGEVG---NGSLTPQQRAWLKLVKPLVITEGFALLSVPTPFVQNEIE
gi|189044597|sp|A6W3V4.1|DNAA_KINRD  LDD----------GVTQHQRAFVRLTRPLGLLDGTALLAVPNDLTKDVIE
gi|189044633|sp|B0RH69.1|DNAA_CLAMS  LTAD--------DRITPQLHGFISLVEPKGVMTGTLYLEVPNDLTRGMLE
gi|226735850|sp|B1VPF0.1|DNAA_STRGG  LLGEGQ------QGIEPKDKQWIERCQPLALVADTALLAVPNEWGKRVLE
gi|254777897|sp|C3PE72.1|DNAA_CORA7  LLAQSEQPNSDVPNFSHSQRLNLQLVEPIMIGDGYALIAAPHENAKTVIE
gi|61212513|sp|Q5Z3Z8.1|DNAA_NOCFA   LTTGSADG--SIPAVTRAQQAWLKLVKPITVAQGFALLSVPSSLAQEAIE
gi|61212561|sp|Q6ABL5.1|DNAA_PROAC   V--------------SKPNRAWLSNTTPVTMHSSTAMVAVPNEFARDRLE
gi|61212563|sp|Q6AHN6.1|DNAA_LEIXX   LETD--------DRITPQLHGFLSLVEPKGIMAGTFYLEVPNEFTRGMIE
                                     :                     :       :  .   : .     :  ::
```

←Figure 2.23. The T-COFFEE output with the show colors option selected.

C. Click on the Show Colors tab on the multiple alignment window (Figure 2.23). The colors give information about the amino acid at the given position (The Key is shown in Figure 2.24). The colorized view gives you a quick way to see what amino acids (or type of amino acid) are common to all sequences (as indicated by the same color being present throughout the alignment segment) and what regions of the multiple alignment are not so well conserved (different colors in multiple rows of the aligned sequences).



AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + Amine + Basic - Q
Others	Gray	

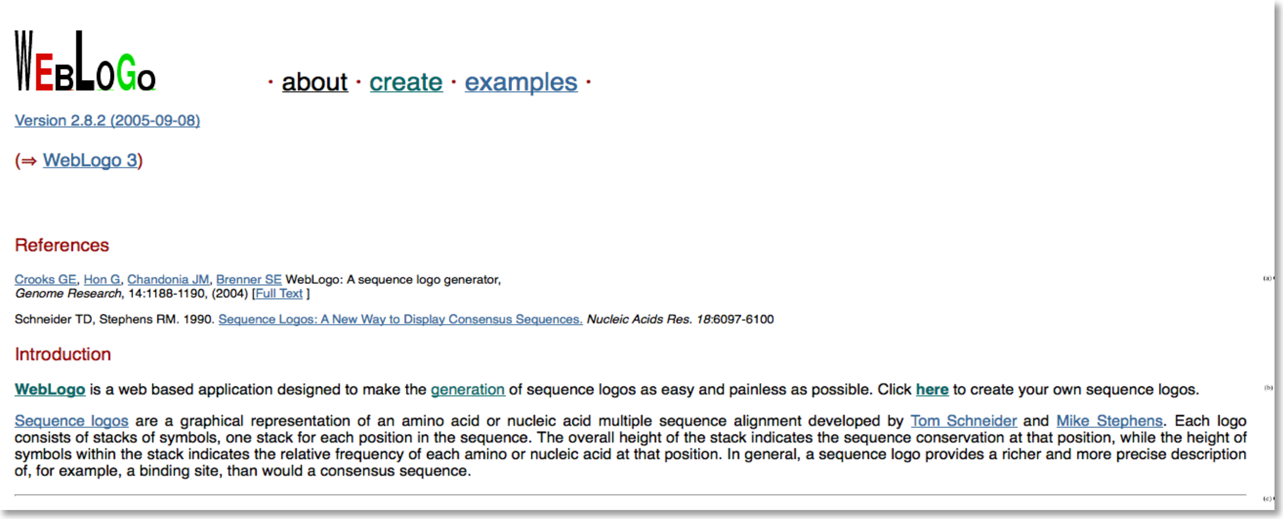
Figure 2.24. The key for the meaning of colors in the T-COFFEE alignment when the “show colors” option is selected.

16. Study the output to see the level of similarity of amino acid sequences in the alignment. Copy and paste or the multi-sequence alignment into the designated section of your lab notebook.

WEBLOGO

1. WEBLOGO will use the multiple alignment you constructed in T-Coffee above and allow you to present the alignment in a way that is easier to interpret. Each section of the multiple alignment will be reduced to essentially a single line in which the most common amino acids in each sequence used in the alignments are represented in a graphical format.

   i. Go to the Weblogo site at http://weblogo.berkeley.edu/ (Figure 2.25).



WEBLOGO

· about · create · examples ·

Version 2.8.2 (2005-09-08)

(⇒ WebLogo 3)

References

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, (2004) [Full Text]

Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res. 18*:6097-6100

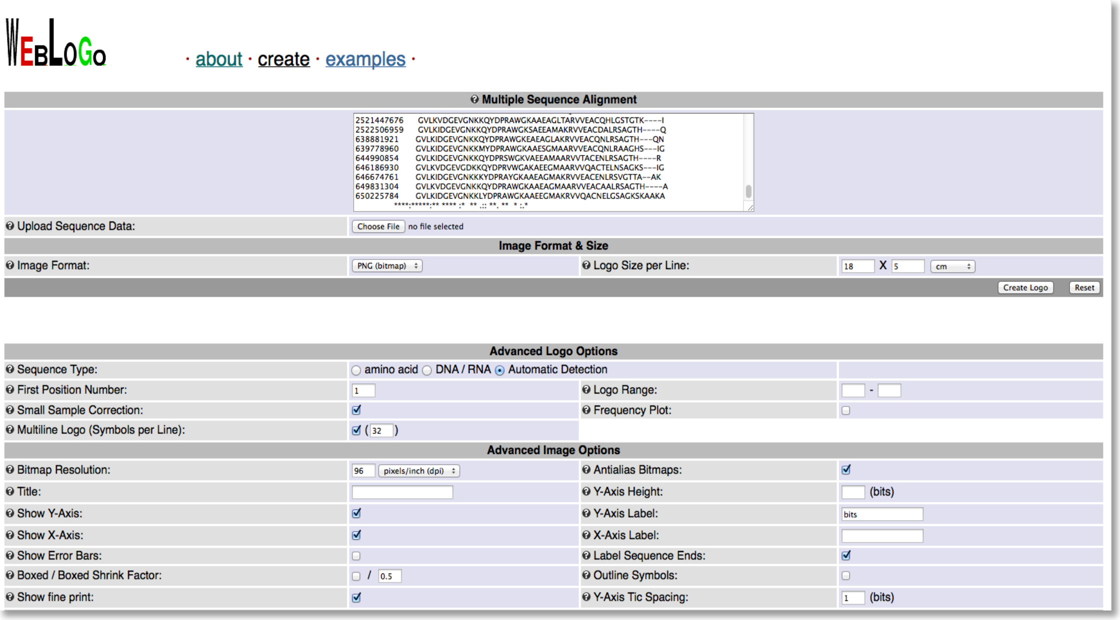
Introduction

WebLogo is a web based application designed to make the generation of sequence logos as easy and painless as possible. Click **here** to create your own sequence logos.

Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Tom Schneider and Mike Stephens. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

Figure 2.25. The WebLogo start page.

2. Click the Create hyperlink at the top of the page to go to the logo creation page (Figure 2.26).



WEBLOGO

· about · create · examples ·

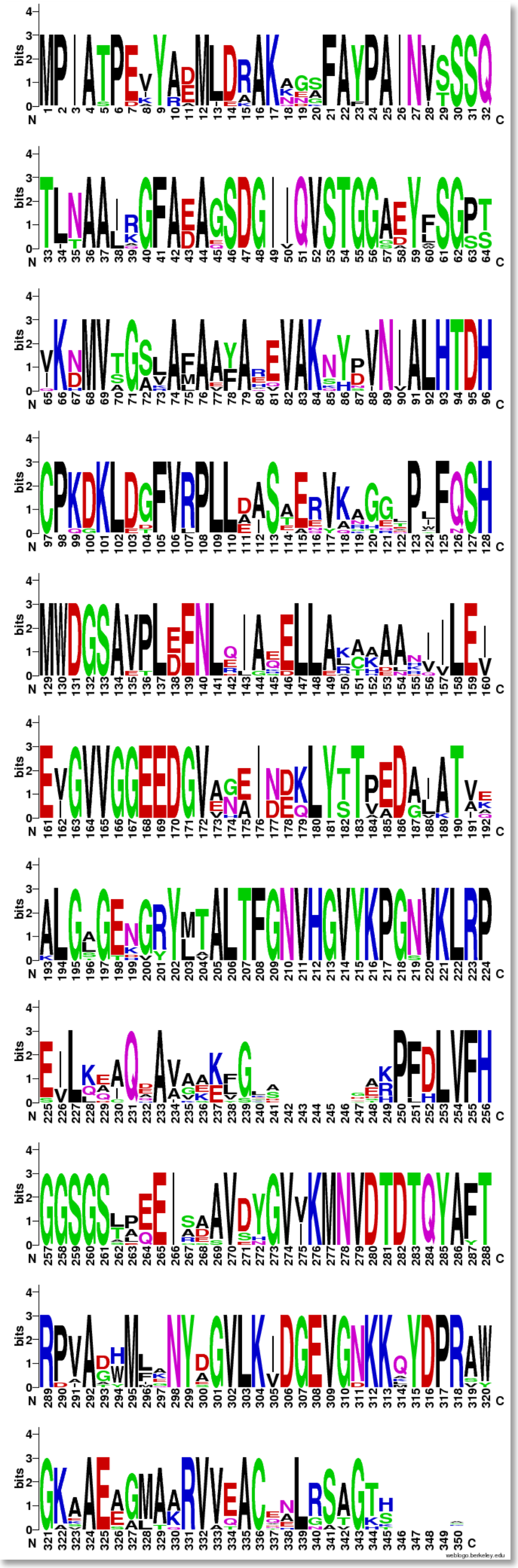
Multiple Sequence Alignment

```
2521447676    GVLKVDGEVGNKKQYDPRAWGKAAEAGLTARVVEACQHLGSTGTK----I
2522506959    GVLKIDGEVGNKKQYDPRAWGKSAEEAMAKRVVEACDALRSAGTH----Q
638881921    GVLKIDGEVGNKKQYDPRAWGKEAEAGLAKRVVEACQNLRSAGTH---QN
639778960    GVLKIDGEVGNKKMYDPRAWGKAAESGMAARVVEACQNLRAAGHS---IG
644990854    GVLKIDGEVGNKKQYDPRSWGKVAEEAMAARVVTACENLRSAGTH----R
646186930    GVLKVDGEVGDKKQYDPRVWGAKAEEGMAARVVQACTELNSAGKS---IG
646674761    GVLKIDGEVGNKKKYDPRAYGKAAEAGMAKRVVEACENLRSVGTTA--AK
649831304    GVLKVDGEVGNKKQYDPRAWGKAAEAGMAARVVEACAALRSAGTH----A
650225784    GVLKIDGEVGNKKLYDPRAWGKAAEEGMAKRVVQACNELGSAGKSKAAKA
             ****:*****:** **** :*  **  .:: **. **  * :.*
```

Upload Sequence Data:	Choose File no file selected		
Image Format & Size			
Image Format:	PNG (bitmap)	Logo Size per Line:	18 X 5 cm
			Create Logo Reset
Advanced Logo Options			
Sequence Type:	○ amino acid ○ DNA / RNA ● Automatic Detection		
First Position Number:	1	Logo Range:	-
Small Sample Correction:	☑	Frequency Plot:	□
Multiline Logo (Symbols per Line):	☑ (32)		
Advanced Image Options			
Bitmap Resolution:	96 pixels/inch (dpi)	Antialias Bitmaps:	☑
Title:		Y-Axis Height:	(bits)
Show Y-Axis:	☑	Y-Axis Label:	bits
Show X-Axis:	☑	X-Axis Label:	
Show Error Bars:	□	Label Sequence Ends:	☑
Boxed / Boxed Shrink Factor:	□ / 0.5	Outline Symbols:	□
Show fine print:	☑	Y-Axis Tic Spacing:	1 (bits)

Figure 2.26. The WebLogo creation page. The multiple alignment generated from T-Coffee is shown pasted into the sequence data box.

3. Copy the T-Coffee alignment obtained above into the box labeled Multiple Sequence Alignment (Figure 2.26).
4. Do not copy the header that begins with "CLUSTAL"
5. Check "Multiline Logo" (default set as 32).
6. Click "Create Logo."
7. Save this logo as a PNG file (or do a screen capture or other method of your choice) screen (Figure 2.27, shown to the right →).
8. Upload the image on to your lab notebook.
9. Comment on any well or poorly conserved regions in your lab notebook.
    a. The relative lengths/sizes of the letters at each position in the alignment indicate the frequency of the amino acid in the alignment. Therefore the taller the letter the more often it appeared at that position in the sequences entered for alignment.
    b. The relative height of the stacks at each position indicates the sequence conservation at that position. For instance a position that is extremely variable and not consistent, whether it wobbles between two different letters or many, will be a shorter stack. On the other hand, a position that is highly conserved between the sequences will be taller in comparison.
    c. The relative widths of the stacks indicate the proportion of valid readings of nucleic bases or amino acids at that position. The more gaps in the sequence at a specific position means a thinner stack.
    d. Amino acids are colored according to their chemical properties: polar amino acids (G,S,T,Y,C,Q,N) are green, basic (K,R,H) blue, acidic (D,E) red and hydrophobic (A,V,L,I,P,W,F,M) amino acids are black.
    e. In the example given in figure 2.27 it would be difficult to comment on specific regions of homology as a good portion of the logo shows tall and wide single letters. This is the characteristic of a well-conserved protein among various species, as indicated

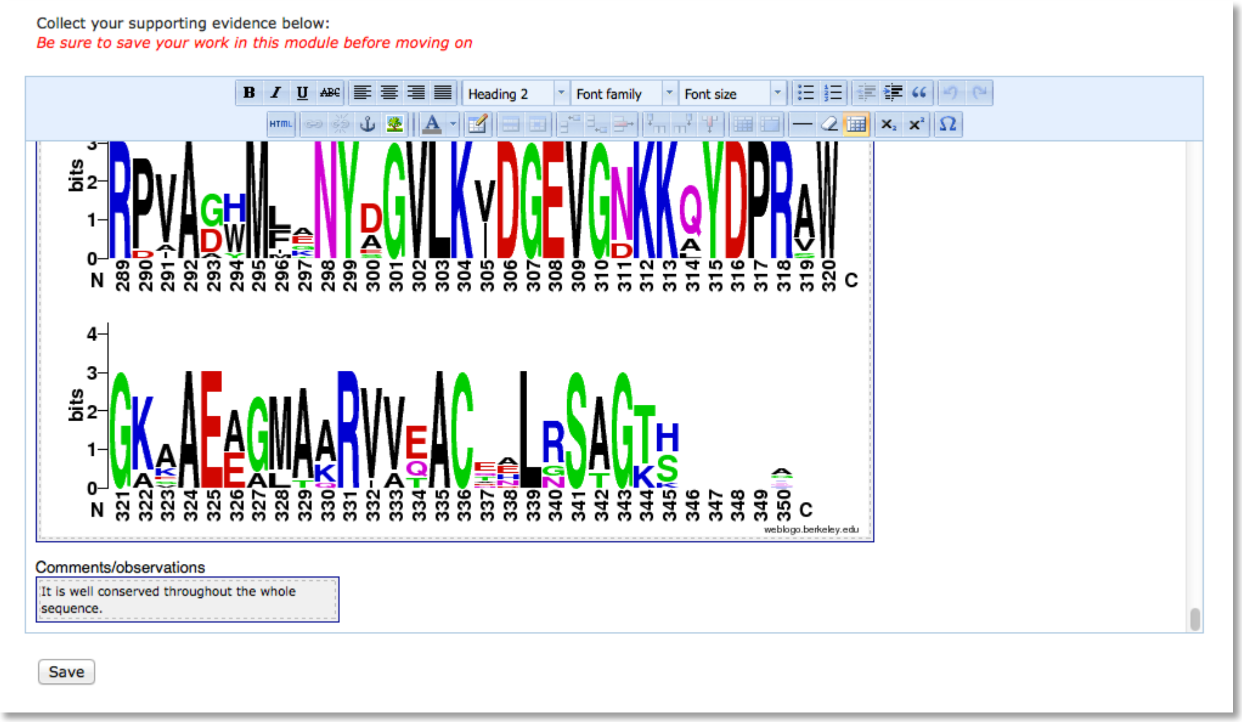


in the example notebook page in figure 2.28.

f. Not all genes that are annotated are likely to have such high conservation. An example of a less well-conserved gene is shown in Figures 2.30-2.31 in T-COFFEE and WebLogo.

   i. Figure 2.29 shows a portion of a T-COFFEE alignment from such a gene. Note the large number of gaps that are present in the alignment.

   ii. Figure 2.30 shows the WebLogo generated from the alignment. Note the amino end of the alignment (lower number residues in the alignment) has very few areas where the letter stacks are significant. On the other hand from the middle toward the carboxy end (highest residue numbers) of the alignment the letter stacks become more pronounced. Thus the proteins in this alignment show homology from the middle to the carboxy terminus only. You may also find even smaller areas of homology in your alignment. If you do you can comment on the positions of significant individual amino acids in your alignment by the noting the position (N = position number) where the alignments occur.



Collect your supporting evidence below:

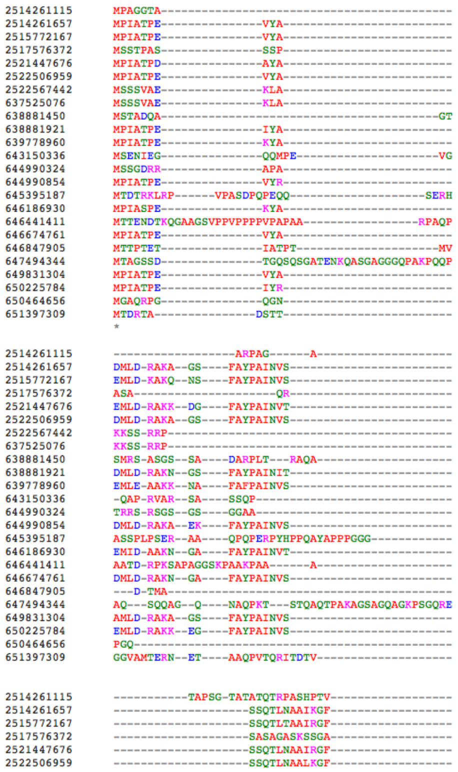
Be sure to save your work in this module before moving on

Comments/observations

It is well conserved throughout the whole sequence.

Save

Figure 2.29. An example notebook page with an alignment comment filled in. Since the logo for this protein shows tall single letter stacks at positions throughout the alignment, the comment has indicated the proteins in the alignment are conserved throughout.



```
2514261115     MPAGGTA-------------------------------------------
2514261657     MPIATPE---------------VYA-------------------------
2515772167     MPIATPE---------------VYA-------------------------
2517576372     MSSTPAS---------------SSP-------------------------
2521447676     MPIATPD---------------AYA-------------------------
2522506959     MPIATPE---------------VYA-------------------------
2522567442     MSSSVAE---------------KLA-------------------------
637525076      MSSSVAE---------------KLA-------------------------
638881450      MSTADQA-----------------------------------------GT
638881921      MPIATPE---------------IYA-------------------------
639778960      MPIATPE---------------KYA-------------------------
643150336      MSENIEG---------------QQMPE---------------------VG
644990324      MSSGDRR---------------APA-------------------------
644990854      MPIATPE---------------VYR-------------------------
645395187      MTDTRKLRP------VPASDPQPEQQ--------------------SERH
646186930      MPIASPE---------------KYA-------------------------
646441411      MTTENDTKQGAAGSVPPVPPPPVPAPAA-----------------RPAQP
646674761      MPIATPE---------------VYA-------------------------
646847905      MTTPTET---------------IATPT---------------------MV
647494344      MTAGSSD---------------TGQSQSGATENKQASGAGGGQPAKPQQP
649831304      MPIATPE---------------VYA-------------------------
650225784      MPIATPE---------------IYR-------------------------
650464656      MGAQRPG---------------QGN-------------------------
651397309      MTDRTA---------------DSTT-------------------------
               *

2514261115     ------------------ARPAG------A--------------------
2514261657     DMLD-RAKA--GS----FAYPAINVS------------------------
2515772167     EMLD-KAKQ--NS----FAYPAINVS------------------------
2517576372     ASA---------------------QR------------------------
2521447676     EMLD-RAKK--DG----FAYPAINVT------------------------
2522506959     DMLD-RAKA--GS----FAYPAINVS------------------------
2522567442     KKSS-RRP------------------------------------------
637525076      KKSS-RRP------------------------------------------
638881450      SMRS-ASGS--SA----DARPLT---RAQA--------------------
638881921      DMLD-RAKN--GS----FAYPAINIT------------------------
639778960      EMLE-AAKK--NA----FAFPAINVS------------------------
643150336      -QAP-RVAR--SA----SSQP-----------------------------
644990324      TRRS-RSGS--GS----GGAA-----------------------------
644990854      DMLD-RAKA--EK----FAYPAINVS------------------------
645395187      ASSPLPSER--AA----QPQPERPYHPPQAYAPPPGGG------------
646186930      EMID-AAKN--GA----FAYPAINVT------------------------
646441411      AATD-RPKSAPAGGSKPAAKPAA------A--------------------
646674761      DMLD-RAKN--GA----FAYPAINVS------------------------
646847905      ---D-TMA------------------------------------------
647494344      AQ---SQQAG--Q----NAQPKT---STQAQTPAKAGSAGQAGKPSGQRE
649831304      AMLD-RAKA--GS----FAYPAINVS------------------------
650225784      EMLD-RAKK--EG----FAYPAINVS------------------------
650464656      PGQ-----------------------------------------------
651397309      GGVAMTERN--ET----AAQPVTQRITDTV--------------------

2514261115     -----------TAPSG-TATATQTRPASHPTV------------------
2514261657     --------------------SSQTLNAAIKGF------------------
2515772167     --------------------SSQTLTAAIRGF------------------
2517576372     --------------------SASAGASKSSGA------------------
2521447676     --------------------SSQTLNAAIRGF------------------
2522506959     --------------------SSQTLNAALKGF------------------
```

← Figure 2.30. A T-COFFEE alignment from sequences conserved in more limited regions of the alignment compared to the example shown previously. Note the large numbers of gaps (----) in this region of the alignment and the limited number of amino acid matches in the first 100 positions of the alignment.

Figure 2.31. Full WebLogo for the sequence shown partially in figure 2.30. The amino terminal portion of the alignment shows negligible conservation, while the central to carboxy terminal regions of the alignment demonstrate much more conservation. Of particular note is a stretch of well conserved polar amino acids (GGSGS) near the carboxy terminus.