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 exercise you will need:

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


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 SwissProt first (as indicated in Figure 2.1). If no significant hits (see 8b below) are found using SwissProt, record that fact in your notebook and then repeat the search using the NR database. If significant hits are then found using the NR database you can proceed using those results.

4. Leave the other settings in Default.
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 2.4 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.

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.
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.

![Figure 2.4](image)

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](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 $1 \times 10^{-3}$. 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 $4 \times 10^{-173}$. 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 $1 \times 10^{-3}$ 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 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).

---

**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 $4 \times 10^{-173}$ 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).
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 *Streptomyces 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.

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

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.
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.

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

<table>
<thead>
<tr>
<th>Range 1: 10 to 517 GenPept Graphics</th>
<th>Score</th>
<th>Expect</th>
<th>Method</th>
<th>Identities</th>
<th>Positives</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>517 bits(1331) 8e-178 Compositional matrix adjust.</td>
<td>286/513(56%)</td>
<td>356/513(69%)</td>
<td>25/513(4%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Query 11**
AIWQEMVHLQGAGIAPRDIGVLRLAVLLEGSTALLAVYDNKDAVEGLHREDVSTA ++W+ A+ L G+ +RL +GL+TALLAV D KD +E +RE ++ A SWHERSALQDD--+VTVQRGQRAPVLRTPZLDDGALLAVTDNLKDVIEQKVRERPLTRA 70

**Sbct 10**
LAELERDIIALSVDPMVSA----AOEFAAPPASDDEFPATGEPFLSTATVDG-- L+I+ LAV+VDP E + P+ E + G + T + DG 123

**Sbct 12**
LSEAYGSPREVAVTVFSDISIQIVLTSPRTGKSHGSGVSVPVRE---TGSVLTGDD 124

**Query 124**
--AVEKHEGSSRFAVESVAPATTASLTA---TNSSPFGV---------YDYSALNHXYTF +++ S T + PC R + S LN XY F 70

**Sbct 125**
GLHLDERGSGLEISDPSDDPSDLFTFEGVDRGFGTQFRTPTRIENSLNPKYIF 184

**Query 171**
DTFTCLSGSNFPAHAATAVAAEAPARAYNLFIVYGGSLGKTHLHEAIGHYARTLDVSDRVR +TFV+C+SNRPAAAVAAAEAPAYL+NPFLYGGSLGKTHLHEAIGHYAR T+V+V

**Sbct 185**
ETTVGASNFGPAHAATAVAAEAPAYL+NPFLYGGSLGKTHLHEAIGHYARQNLQYPVYYQVQV 244

**Query 231**
KYVNFQSTFQAFNAVAGAQAOFQORFYVDVAVLIDIQOFKQGMOKMFTEFHFHTFTNLH +YVNFQSTFQAFNAVAGAQAOFQORFYVDVAVLIDIQOFKQGMOKMFTEFHFHTFTNLH 290

**Sbct 245**
KYVNFQSTFQAFNAVAGAQAOFQORFYVDVAVLIDIQOFKQGMOKMFTEFHFHTFTNLH 304

**Query 291**
NSEEKQIVTSIQPPKLSGCWRAMSFRPWGKLDVQCPPDLEDTRAILRKRAAADKLKDIP N+KQ+VTSTDPPK+LGCERSRPWGKLDVQCPPDLEDTRAILR+KA ++L+P

**Sbct 305**
NASKQVKVTSIQPPKLSGCWRAMSFRPWGKLDVQCPPDLEDTRAILRKAAGKELYPQVQYQV 364

**Query 351**
DDVLHLIASKISSNIRELEGALTTRVFAAPLSSGLSPDELAYRVTMLDVMPGDSQQTPT DDV IASKISSNIRELEGALTTRVFAAPLSSGLSPDELAYRVTMLDVMPGDSQQTPTA 410

**Sbct 365**
DDVLHLIASKISSNIRELEGALTTRVFAAPLSSGLSPDELAYRVTMLDVMPGDSQQTPTA 424

**Query 411**
MILEETAGYFVUSVIRQGASSRRNLTRAQIMLYCRLEDLSLPIKGEFGRHDHTTV I+TA YF +E+E+ G SRS R ARQIMLYCRLEDLSLPIKGEFGRHDHTTV 470

**Sbct 425**
AIMQGQTVSTVTDLRSQTRQIMLYCRLEDLSLPIKGEFGRHDHTTV 484

**Query 471**
MHEAKKIQQQLGEQDRVYDEVSeLTSTIRKKA 503

**Sbct 485**
MHEAKKIQQQLGEQDRVYDEVSeLTSTIRKKA 517

---

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

**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 NCBI BLAST search.

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.

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

Figure 2.14. The COG description page. The COG name and number are indicated along with the description of the COG.
Figure 2.15. The CDD notebook page. Spaces are available for data from up to 2 COG hits.

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)

1. On 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 *Kyttococcus* 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*.

---

**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.)
2. 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.

3. 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).
4. 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.

5. 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.

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

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

C. 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.
6. 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).
7. Select all 11 of the FASTA formatted sequences and copy them.
8. Got to EBI’s T-Coffee server at [http://www.ebi.ac.uk/Tools/msa/tcoffee/](http://www.ebi.ac.uk/Tools/msa/tcoffee/) (Figure 2.21), paste all 11 sequences into the input window and click submit.

![Figure 2.21. The T-COFFEE start page.](image)
9. A results page with CLUSTAL FORMAT for T-Coffee is seen as shown in Figure 2.22.

A. You will see repeating blocks of 50 amino acid stretches 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

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.
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).
10. 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.

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

![Figure 2.25. The WebLogo start page.](image)

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

![Figure 2.26. The WebLogo creation page. The multiple alignment generated from T-Coffee is shown pasted into the sequence data box.](image)
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.

1. 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.

2. 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.

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.
Figure 2.29. 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 2.30. Full WebLogo for the sequence shown partially in figure 2.29. 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.