Guide for Bioinformatics Project Module 3

Structure-Based Evidence and Multiple Sequence Alignment

In this module we will revisit some topics we started to look at while performing our BLAST search and looking at the CDD database in the first module. In the first part of this module we will investigate predicted functions based on the structure of your protein. This will involve looking for protein domains using additional structural recognition programs beyond the COG hits we have previously identified. This will help you to hypothesize about the probable function of your gene's product in the cell based on its structural elements and similarity to structures of known proteins or protein domains by searching in several different databases. We will also look at characteristics of the protein sequence that may not form a domain but could have secondary structure or enzymatic activity in order to help predict function. Next we will analyze what changes have occurred or not occurred in the evolution of your protein through different species. This analysis will allow us to determine what parts of the protein may be key to its function, as those will have been conserved over time.

Structure-Based Evidence

TIGRFAM

TIGRFAM is a database of protein families that features curated multiple sequence alignments, Hidden Markov Models (HMMs), and associated information designed to support the automated functional identification of proteins by sequence homology. In contrast to Pfams (which you will use in the next section), TIGRFAMs are often constructed from full-length protein sequences or well-conserved and functionally understood domains. Therefore, TIGRFAM results are very useful for predicting the name and/or function of the gene product.

Go to TIGRFAM at http://tigrblast.tigr.org/web-hmm.

Paste the protein sequence in FASTA format into the search box and click "Start HMM Search". *Protein sequence in FASTA format can be downloaded from the Protein Tab on SGD from the Predicted Sequence line or be copied from your Module 1 Worksheet.*

After searching the TIGRFAM database, raw text results will show which TIGRFAMs have been found to be matches. The name of the TIGRFAM ('**Description'**' column) may be cut off. If this is the case, identify the TIGRFAM number (e.g. **TIGR#####**) by the code found in the '**Model**' column and then go to the following page:

http://cmr.jcvi.org/tigr-scripts/CMR/shared/MakeFrontPages.cgi?page=text_search&crumbs=searches&type=hmm

Search the database with the full number to find the entire TIGRFAM name.

For each significant TIGRFAM hit, report the Number, Name, Score, and E-value in the Module 3 Worksheet. If a GO (Gene Ontology) number or an EC (Enzyme Commission) number is shown in the full description of the TIGRFAM entry, you should record these as well since they may prove very useful when attempting to predict the function of the gene product. ***Only record TIGRFAM hits (those starting with the prefix "TIGR") at this point. In the next section, you will search for Pfam hits.***

Pfam

Pfam is a database of protein families, each represented by Hidden Markov Models (HMMs) generated from manually-curated multiple-sequence alignments of common protein families and domains. [Recall that domains are "modules" in proteins that usually have conserved tertiary structure and function]. Pfam can be used to identify likely

protein domains within an amino acid query sequence. For each domain identified, Pfam can provide a great deal of information (if it is available) pertaining to the domain function, sequence conservation, and critical residues. A protein sequence may contain multiple Pfam domains, but these will never overlap.

Navigate to the Pfam Search page at http://pfam.sanger.ac.uk/search. Enter your protein's amino acid sequence and click "Submit". You will be directed to a new page that at the top under Sequence Search Results shows a graphic of your protein with the domains found and lower down the page has a Significant Pfam-A Matches section. By using the hyperlinks under FAMILY and CLAN and the 'Show' button under Show/hid alignment you can learn important information about these Pfam protein domains.

Predicted Domains

Note the E-value for the hit (match between the query and a database sequence). Why might it be useful to examine a hit even though it has a relatively high E-value?

For each Pfam **record the Pfam Name (Description), Number, Score, and E-value and Predicted Active Sites in the Module 3 Worksheet** *(if not all of this information is found here you will find it when looking at the Domain Summary section below).*

[Note: Is the whole Pfam hit (listed as an HMM) covered by the alignment between your protein and the HMM? If not, the text in either the 'From' or 'To' field under "HMM" will be highlighted in ***red****. If a large portion of the Pfam domain is missing due to truncation, it is possible that the domain in your protein may not fold in the same way or perform the same function as it does in other family members.]*

In some cases, there might be more than one Pfam hit for the query sequence. In this case, be sure to record the relevant information as above, and consider why the different domains might exist in the same protein. This could help in determining the identity or function of the gene product.

Pairwise Alignment

To examine the alignment between the query sequence and the Pfam HMM, go back to the starting page and under the Significant Pfam-A Matches section, click the "Show" button under "Show/hide alignment". **Record this alignment in the Module 3 Worksheet.** [*Note: You may need to save this file and print it out separately if you are having trouble inserting it into the worksheet. In this case please print the document and just put a note in this section of the worksheet that says see page __ for alignment*].

The first row in the alignment shows the consensus sequence for the HMM (shown in various blue colors); in this sequence, capital letters correspond to highly conserved amino acids in the alignment used to generate the HMM. The bottom row in green is the actual sequence of your protein domain that has been aligned to the consensus sequence in blue. The row directly below the blue consensus sequence lists residues in your query sequence that are identical to those in the consensus by using the actual letter of the amino acid and indicates those that are not identical but are similar in nature by using a + sign. This similarity is based on amino acid characteristics that we will look at in more detail in the Multiple Sequence Alignment portion of this module. **Identify any relatively large regions with fully identically sequences or relatively large regions that are lacking any conservation in the Module 3 Worksheet.**

Be sure to comment on details such as the alignment and E-value and how this information contributes to your annotation.

Guide for Bioinformatics Project Module 3

Domain Summary

Now look at the **Domain Summary** page for the domain by clicking on the hyperlinked family name under FAMILY. This links to a new page with the full Family Name indicated at the top of the page; the Pfam number will often be listed at then end of this name in parentheses in the following format: (PF#####). **Copy the Pfam number (formatted as PF#####) from the top of the page and the Clan name and number (if available) from the bottom of the page into the Module 3 Worksheet.**

Read any text on this page carefully. Since each Pfam domain has been manually curated, this information can be extremely useful in predicting the function of a query gene product containing a match to the domain. If a GO or EC number is shown, record that number in the Module 3 Worksheet as it will aid in predicting the function of the gene product.

HMM Logo

On the left-hand menu of the Domain Summary page, click “HMM Logo”. This visualization tool will help to highlight what may be the most important residues in this domain based on those that have been unchanged throughout evolution. To do this HMM Logos provide the researcher with a quick overview of the features of a Pfam HMM while conserving as much information as possible. The larger a letter is in an HMM Logo, the more conserved this residue is in the protein family; the smaller a letter or presence of multiple letters indicates little to no conservation of amino acids at that position. Colors correspond to different amino acid types (e.g. neutral, acidic, etc.). Letters are sorted vertically in descending order depending on their probability of occurring at a given position in a sequence that contains the domain.

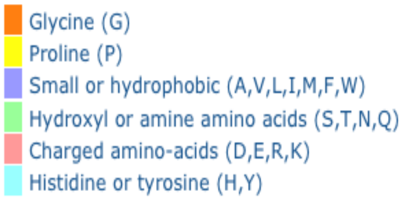
Save the HMM logo as a .PNG file. Go back to the Module 3 Worksheet and upload the file. [*Note: You may need to save this file and print it out separately if you are having trouble inserting it into the worksheet. In this case please print the document and just put a note in this section of the worksheet that says see page # __ for alignment*].

Curated Alignment

While we have identified potential functional domains and the important residues within those domains we have yet to explore if any of the residues are involved in formation of enzyme active sites. To find the active site residues, we can look at the Curated Alignment (HMM logos do not identify active site residues).

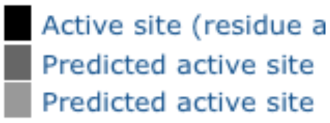
On the left-hand menu, click “Alignments”. Under View options find the HTML line in the table and scroll over the first checkmark in that row (chose the checkmark under Seed if possible, but if this is an X, select any checkmark in the HTML row). By scrolling over the checkmark a hyperlink called View is uncovered, click here; a new page will open displaying your sequence alignments.

The conserved residues are highlighted using the following color scheme:



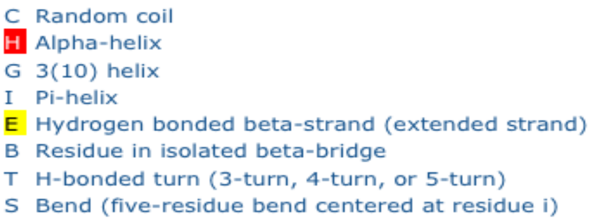
🟧 Glycine (G)
🟨 Proline (P)
🟪 Small or hydrophobic (A,V,L,I,M,F,W)
🟩 Hydroxyl or amine amino acids (S,T,N,Q)
🟥 Charged amino-acids (D,E,R,K)
🟦 Histidine or tyrosine (H,Y)

Active site residues are highlighted in black and grey as shown below:



■ Active site (residue a
▩ Predicted active site
▨ Predicted active site

If protein structure can be predicted from the sequence, then a line, highlighted in grey, will appear between the lines of text and the following symbols represent the secondary structure identified:



C Random coil
H Alpha-helix
G 3(10) helix
I Pi-helix
E Hydrogen bonded beta-strand (extended strand)
B Residue in isolated beta-bridge
T H-bonded turn (3-turn, 4-turn, or 5-turn)
S Bend (five-residue bend centered at residue i)

Guide for Bioinformatics Project Module 3

By comparing information from the pairwise alignment, the HMM logo, and the curated alignment, you should be able to identify any key functional residues for your protein.

Report the key functional or structural amino acid residues in the Module 3 Worksheet.

Protein Data Bank (PDB)

The Protein Data Bank is the single worldwide depository of information about the three-dimensional structures of large biological molecules, including proteins and nucleic acids. A variety of information associated with each structure is available through the PDB including sequence details, atomic coordinates, crystallization conditions, 3-D structure neighbors computed using various methods, derived geometric data, structure factors, 3-D images and a variety of links to other resources.

Copy your FASTA format protein sequence and go to:
http://www.rcsb.org/pdb/search/advSearch.do?st=SequenceQuery

You will land on a Sequence Search page, copy the amino acid sequence of your protein into the box labeled Sequence. The default cut-off E-value is 10; change this to 0.01.

Click the "Submit Query" button, when your new page loads if you only have one hit you will be taken directly to the page with information about that hit, however if you have multiple hit results you will need to scroll down to the section under the blue box that says Filter. Performing this search runs a BLAST search just like you did in NCBI BLAST in the Sequence-based Similarity module. In this case, however, the query sequence is searched against all of the protein sequences that have solved structures in the PDB.

Examine the quality of the alignments between the query and the BLAST hits in the Protein Data Bank. If the E-value meets the cutoff of 0.01 and a significant length (residue count) of the protein is aligned, the match is considered a good one. When two proteins are very similar in amino acid sequence and have approximately the same length, it is highly probable that they fold in a similar manner. Therefore, the structure corresponding to the PDB BLAST hit predicts how your query gene product is likely to fold.

Record the PDB Code, Name, Length, Score, Alignment length, and E-value of your top three hits that are good matches, in the Module 3 Worksheet.

For each of these top three hits – click the hyperlinked name of the hit to move to the PDB page that tells you about this structure and its function. Each page will be different based on what is known about this structure and folding but you will most likely find lots of good information on this page including, a picture of the 3D structure, a molecular description of the protein, a link to the paper that published the structure with an abstract of the work, information about ligands for this structure, etc. **Read through this whole page and make notes in your Module 3 Worksheet about relevant information that may help in explaining a possible function of your protein.** [Note: You may need to save this file and print it out separately if you are having trouble inserting it into the worksheet. In this case please print the document and just put a note in this section of the worksheet that says see page __ for alignment].

If there is a literature reference associated with the protein structure, it may be beneficial to read this (or at least attempt to). When a structure is published, the authors will frequently characterize the function of the protein and identify important residues within its amino acid sequence. If these residues are conserved your gene product, this helps to confirm its identity and function.

Don't be concerned if there are no significant hits with your query sequence since not all proteins are included in the PDB database.

Guide for Bioinformatics Project Module 3

Protein Tab on SGD

There are many databases in existence that use different algorithms to predict and detect structural regions of proteins. Therefore in this section we will check additional structure detection programs for potential domains, you may get many more results, duplicate results or no hits, all of which are informative.

The Saccharomyces cerevisiae Genome Database (SGD) curates information from six protein structural detection algorithms and displays this data under the Protein Tab. Navigate to SGD and type your gene name into the search box. When you are linked to your gene homepage, click on the Protein Tab and scroll down to the protein diagram graphic under the Predicted Sequence section. This graphic represents all of the structural motifs found in your protein and displays them colorimetrically based on what search algorithm identified each one. The six detection algorithms that were utilized are PFAM, TIGRFAM, SUPERFAMILY, SMART, GENE3D and PANTHER.

Click anywhere on the upper part of this graphic and you will be redirected to a new page that displays your protein, the predicted structural motifs and color key legend that tells you the database in which it was identified. **Copy this graphic into your Module 3 Worksheet.** For the 4 algorithms we have not yet checked (SUPERFAMILY, SMART, GENE3D and PANTHER) – if the graphic represents a hit from that database click on that rectangular representation in the graphic (*note that if there is more than one domain in the same color, you need to click on all of them and record that information*), read about the algorithm below and follow the instructions to interpret the data and to learn more about this domain so as to relate it to your current hypothesis of your proteins function.

SUPERFAMILY

"SUPERFAMILY is a database of structural and functional annotation for all proteins and genomes. The SUPERFAMILY annotation is based on a collection of Hidden Markov Models (HMMs), which represent structural protein domains at the SCOP (Structural Classification of Proteins) superfamily level. A superfamily groups together domains, which have an evolutionary relationship. The annotation is produced by scanning protein sequences from over 2,414 completely sequenced genomes against the hidden Markov models." "Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The **SCOP** database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification."

If you have a SUPERFAMILY identified domain – click on that colored rectangle and you should be directed to a new page that should indicate you are on the Structural Classification Tab. Directly under the Structural Classification Tab, in larger print should be displayed the name of the Superfamily that has been identified. Below this is the SCOP Classification that breaks down the Class and Fold types of this protein domain. Below this are listings of families of domains that belong to this Superfamily. This information on family members may already be familiar having seen in it previous hits or it may come up again later so read through this list and make note of its contents. **Copy this information on Superfamily, Class, Fold and Families into your Module 3 Worksheet.**

SMART

"**SMART** (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signaling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database as well as search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa."

If you have a SMART identified domain – click on that colored rectangle and you should be directed to a new page that loads the information about this protein domain. This should include the acronym of the domain name, what the acronym stands for, a SMART accession number, a simplistic cartoon drawing of the domain and sentences or paragraphs giving you more information about this domain region. **Make note of the information provided here in your Module 3 Worksheet.**

GENE3D

"Gene3D takes CATH domains (from PDB structures) and assigns them to the millions of protein sequences (using Hidden Markov Models) with no PDB structures. Assigning a CATH superfamily to a region of a protein sequence gives information on the gross **3D structure** of that region of the protein. CATH superfamilies have a limited set of functions and so the domain assignment provides some **functional insights**. Furthermore most proteins have several different domains in a specific order, and so looking for proteins with a similar **domain organization** provides further functional insights."

If you have a GENE3D identified domain click on that colored rectangle and you should be directed to a new page with the Summary information about the domain found in your protein. **At the top of the page will be the CATH Superfamily identifier number and name of this superfamily (if there is one); make note of this information in your Module 3 Worksheet.**

Next click on the Classificaiton/Domains Hyperlink under the Superfamily Links section to be taken to a new page that displays the CATH Classificaiton. **Record these Levels (hover over the colored circle to find out what it stands for), CATH Codes and Descriptions in the Module 3 Worksheet.**

PANTHER

"The PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System is a unique resource that classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence. Proteins are classified by expert biologists according to: 1) Gene families and subfamilies, including annotated phylogenetic trees 2) Gene Ontology classes: molecular function, biological process, cellular component 3) PANTHER Protein Classes 4) Pathways, including diagrams."

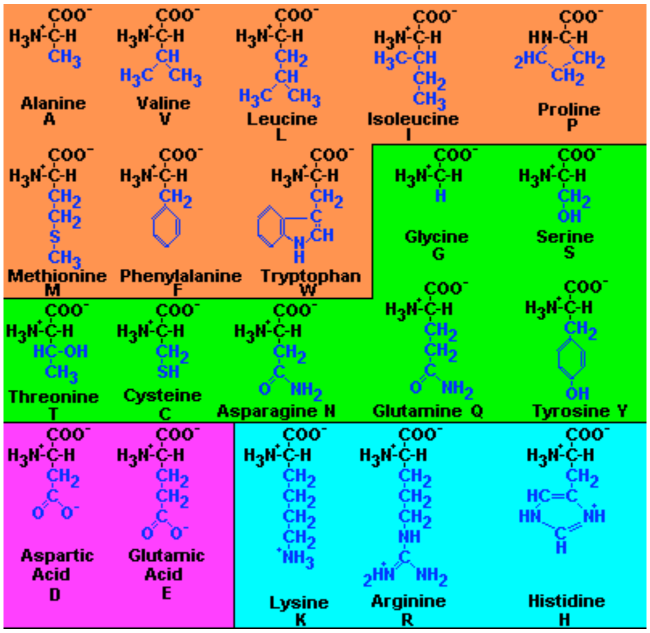
If you have a PANTHER identified domain click on that colored rectangle and you should be directed to a new page with PANTHER FAMILTY INFORMATION. **Record the Family Name in your Module 3 Worksheet. If there is a hyperlinked Subfamilies Number, click this to link to the Subfamily member and record their names in your worksheet. Also record any Gene Ontology (GO) information presented on the main page.** [***Gene Ontologies, or grouping, categorize genes based on molecular function, biological process or cellular component, these groups are then applicable to all members of the group***].

Guide for Bioinformatics Project Module 3

Multiple Sequence Alignment

Amino Acid Properties:

Utilizing multiple sequence alignment tools requires a basic knowledge of amino acid properties. Take time again now to review these basic properties so that you will be prepared to properly analyze your results. An amino acid characteristics chart has been provided on the left for your use as a starting point.



Amino acids in **orange** have hydrophobic side chain R groups. Amino acids in **green** are considered to be hydrophilic because they have electronegative groups on the side chain except tyrosine which because of the phenyl ring side chain is also hydrophobic in character. Two amino acids in **pink**, Glu and Asp, have two carboxylic acids in the side chain, are hydrophilic and contribute one negative charge to a polypeptide chain at neutral pH. The basic amino acids in light blue are also very hydrophilic and are positively charged at neutral pH. It should be clear from this that amino acid side chains which contribute to overall charge on a protein are either acidic or basic at neutral pH.

The structure of amino acids shown here are by Dr. Robert J. Huskey (retired) University of Virginia. http://njms2.umdnj.edu/biochweb/education/bioweb/PreK2010/AminoAcids.htm

Tree-based Consistency Objective Function For Alignment Evaluation (T-COFFEE)

As mutations accumulate in a gene over time, the amino acid sequence will begin to undergo modifications. If a mutation leads to loss of protein function or inability to fold correctly, the fitness of the organism may be decreased rendering it less likely that the mutated copy of the gene will be passed on. Consequently, amino acids that are conserved among modern protein sequences are likely to be those that are important for the function or structure of the protein. One way to measure conservation is by aligning a large number of similar protein sequences. T-COFFEE is a computer program that creates such multiple sequence alignments. [Note: This is similar to the HMM Logo program within PDB, however it looks at the sequence of the whole protein, not just the identified protein domain.] Multiple alignment programs are used to analyze a set of related sequences identified by other programs such as BLAST or IMG Orthologs. Once you have obtained a set of related amino acid sequences, you can use T-COFFEE to create a multiple alignment of the original query with the other sequences in the group.

From your gene homepage on SGD, click on the Protein Tab, scroll down to the **External Links** section, and under the Homologs Selection, select BLASTP (NCBI). Perform the BLAST using the non-redundant database, changing the Max Target Sequences to 500 under General Parameters section.

When your results have loaded, scroll through the sequences [under the Alignments section] and check (click the box to the left of the description section) the first 20-30 sequences with significant E-values that are not from *S. cerevisiae*. [*Note that sequences longer than 3000 amino acids cannot be analyzed. Check with your instructor if this appears to be a problem with your sequence*.] You will want to sample a wide range of significant hits, not just the very best or those only from Saccharomyces species. Try using several different sets of sequences in the following steps.

Once you have selected 20-30 sequences, scroll back up to the top of the Descriptions section and click **Download** and select **FASTA (complete sequence)**. A new window will open that will allow you to save this file, which defaults to **seqdump.txt**, save this on your desktop. **It may save it directly to the downloads folder, if so you can retrieve it directly**

from the downloads folder in the next step. [*You may want to print a copy of this file so that you know which sequences you used, you can include this in your notebook, you should be able to open the file with a text reader.*]

Navigate to EBI's T-Coffee Server at www.ebi.ac.uk/Tools/msa/tcoffee/

At the bottom of the STEP 1 box there is an option to upload a file: click the Browse button and navigate to select your seqdump.text file you saved on the desktop. Click the SUBMIT button.

You will land on a results page with a CLUSTAL FORMAT for T-COFFEE. This is the alignment of your sequences. Above the Alignment click the button that says “Show Colors”, this will color the alignment by amino acid properties to ease in your viewing and identification of important regions.

Note: You should check to see if any of the sequences in the alignment have significantly different lengths than the others. Your query sequence will be shown in the rows that have the corresponding reference numbers in the left-hand column. If the sequence being annotated is much longer or shorter at the N terminus than other sequences in the alignment, the automated gene caller may have predicted the incorrect start codon.

Briefly scan the alignment for regions that appear to be highly conserved. At the bottom of each section of alignment, highly conserved positions will be marked with a colon, and 100% conserved positions will be marked with an asterisk. Since the sequences in this alignment are grouped based on similarity to one another, see if you can spot distinct subgroups of sequences in the alignments where a position is highly conserved in that subgroup but poorly conserved outside of it. In the next section, you will build a **sequence logo** to help you identify these highly conserved regions.

You are not required to copy and paste this alignment into your Module 3 Worksheet because it will likely take up 20 pages and you will lose the colors when you copy it. However you should look over the alignment and make notes in the provided section of your Module 3 Worksheet about conserved regions of interest.

Scroll back up to the top of the page and click the “**Download Alignment File**” button; this will link you to a new page where you can save the file or copy the alignment with greater ease. This alignment will be used in the next module of creating a WebLogo.

WebLogo

WebLogo is a program designed to enable easy creation of sequence logos from multiple sequence alignment data. When comparing sequences in a simple text format, it can be very difficult to visually interpret and describe levels of conservation beyond such vague terms as “well conserved”, “partially conserved” or “poorly conserved”. Sequence logos represent the information obtained from a multiple alignment in the form of a simple graphic where the most common amino acid residue at each position in the alignment will be the tallest symbol at that position, and the overall height of a stack of symbols is proportional to the percent conservation (as seen previously with the HMM Logo). Because the logo creation program is calculating percentages, you will want to use an alignment of least 10 sequences as your input, if this is possible. If you have fewer than 10 that meet our previous criteria, use all of the homologs listed.

Navigate to WebLogo at http://weblogo.berkeley.edu/.

Click “Create” at the top of the page. Copy the T-Coffee alignment obtained earlier into the form located below **Multiple Sequence Alignment** or click the Browse button in the Upload Sequence Data section to upload a saved file. If you are copying and pasting, make sure you do not include the header (begins with the word “CLUSTAL”) since WebLogo can't read it.

Check the box next to “Multiline Logo”, and click “Create Logo”. If the letters are too thin or small to see easily, click on the logo, the default of this program is that the mouse pointer serves as a zoom tool. If this does not work, change the Symbols per Line (default 32) and Width (default 18) and click “Create Logo” again until the logo is easy to read.

Save the logo as a PNG file (right-click on the image, select View Image Info, choose Save As option in right hand corner, save as PNG file) and upload this to the Module 3 Worksheet.

Comment on any sequences that are very well conserved or very poorly conserved. You don’t need to describe this for each individual position but note any broad regions or individual amino acid residues that appear to be of particular significance, and indicate where these are located by specific position numbers (shown below each line of the logo) or by describing them in general terms (e.g. “the N-terminal third of the sequence”). Is there anything new that you can see from the logo that you didn't notice in the original text-based alignment?