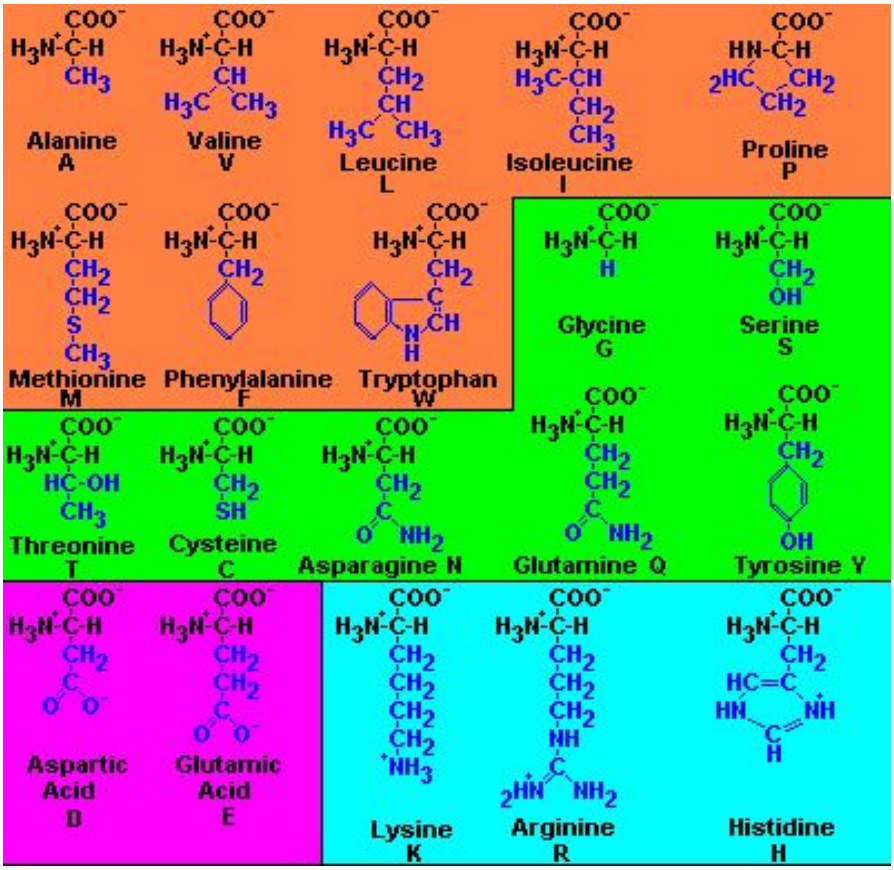
Yeast ORFan Gene Project: Module 4 Guide

Multiple Sequence Alignment

Amino Acid Properties:

Utilizing multiple sequence alignment tools requires a basic knowledge of amino acid properties. This bioinformatics module has been kept intentionally short so that you have enough time to spend time on the basic details and properly analyze your results. An amino acid characteristics chart has been provided below for use as a starting point.



http://njms2.umdnj.edu/biochweb/education/bioweb/PreK2010/AminoAcids.htm

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.

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 ability 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. 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.*

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.

Yeast ORFan Gene Project: Module 4 Guide

From your gene page on SGD, click on the Protein Tab, scroll down to the **Resources** section, and under the Homologs section, select BLASTP at NCBI. Perform the BLAST using the non-redundant database.

When your results have loaded you will scroll through the sequences [under the Alignments section] and check the first 10-30 sequences with significant E-value 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.]

The more significantly similar a group of sequences is, the easier it will be to determine which amino acid residues are the most highly conserved within that group. Limiting your selection to those sequences with the greatest similarity to yours, however, can produce misleading results since these may have functionally diverged from other homologs. In this case, some amino acid residues that are actually important for function in your protein may not appear to be well-conserved in the sequence alignment. You may want to sample a wide range of significant hits, not just the very best. Try using several different sets of sequences in the following steps.

Once you have selected 10-30 sequences, scroll back up to the top of the Description section and click **Download**, FASTA (complete sequence) should be the selected default, hit continue. 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 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 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.txt** file you saved on the desktop. Click the green SUBMIT button in Step 3.

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.

You should scroll through and 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 in the sequences chosen. At the bottom of the 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 4 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 4 Worksheet** about conserved regions of interest.

Click the “**Download Alignment File**” button; this will link you to a new page where you can copy the alignment with greater ease. Copying this alignment will be used in the next step of creating a WebLogo.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.

Yeast ORFan Gene Project: Module 4 Guide

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.* 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. Under the **Multiple Sequence Alignment** section, click Browse next to Upload Sequence Data and upload the file you just saved from your T-COFFEE alignment.

Under **Advanced Logo Options,** check **"Multiline Logo",** and click **"Create Logo".** If the letters are too thin, try clicking on the logo, the default of this program is that the mouse pointer serves as a zoom in 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 4 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?

Homology

The overview page on SGD provides icons under "comparative info" provided by the Alliance of Genome Resources. These icons indicate which organisms contain a potentially comparative gene. Many genes of unknown function are only found in S. cerevisiae (orphan genes), but some genes of unknown function have homologs in other species. Look at the summary page for your gene and indicate whether there is comparative information for other species, and if so, which ones. Click on the icons and make a brief note of the gene function.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only. (revision removed gene context and SGD gene synteny [rev 6..6.2022 removed SGD syteny viewer and added comparative info]