Practicum Assignment 2 *Due Friday, Nov. 10 by 4:00pm*

Submit your assignment by email to comp-bio@cs.cmu.edu.

In this assignment, you will construct and analyze multiple sequence alignments of the pylS and pylB enzyme families. These two proteins are components in the machinery for incorporating the non-standard amino acid pyrrolysine into proteins. These include many methanogenic Archaea, various Firmicute species, and a few species from other bacterial phyla.

To do this assignment, you will need

* The Jalview multiple alignment and visualization software.
* The Practicum 2 worksheet.
* *pylB.fasta,* *pylS.fasta:* These files contain collections of amino acid sequences in fasta format.

These two FASTA files and the worksheet can be downloaded from the syllabus page.

***What to hand in:***

* The completed Practicum 2 worksheet.
* Screen shots of your PylS Neighbor Joining trees with differences circled in the Muscle tree.
* A fasta file with your MAFFT alignment of PylS.
* A screen shot of your PylB alignment with the DEMMA sequence.
* A fasta file with your trimmed MAFFT alignment of PylB in FASTA format.

## **Section 1: The PylS gene family**

1. **Rename sequences to provide easy identification of sequence origin**
2. **Setting Jalview preferences**
3. **Align the PylS sequences**
4. **Comparing the output of two alignment programs**
5. *Comparison of Neighbor Joining (NJ) trees*:
   1. **Save a screen shot of the two trees. In the MUSCLE alignment NJ tree, draw a circle around nodes that are placed differently in the MUSCLE and MAFFT trees. You can do this in a program like Paint or Power Point. Or, print out the image, mark it up with a pen and hand in a hardcopy version.**
6. *Comparison of Functional Features*:
7. **What are the first and last amino acids in the Pyrrolysyl-tRNA ligase, N-terminal (IPR023878) domain?**
8. **What are the first and last amino acids in the Pyrrolysyl-tRNA ligase, C-terminal (IPR023877) domain?**
9. **In the table in the worksheet, for each alignment, record the column number along the top track for each of the following amino acids:**

* **The last amino acid in in the Pyrrolysyl-tRNA ligase, N-terminal (IPR023878) domain in METBA.**
* **The first amino acid in in the Pyrrolysyl-tRNA ligase, C-terminal (IPR023877) domain in METBA.**
* **DESAC Trp(5); i.e., the 5th residue, in the** *Desulfotomaculum acetoxidans* **sequence which is a tryptophan (Y).**
* **DESAC Phe(27)**
* **DESAC Trp (37)**

|  |  |  |
| --- | --- | --- |
|  | Column in Muscle alignment | Column in MAFFT alignment |
| METBAPylS N-term domain (end) |  |  |
| METBA PylS C-term domain (start) |  |  |
| DESAC Trp(5) |  |  |
| DESAC Phe(27) |  |  |
| DESAC Trp(37) |  |  |

1. **In the MUSCLE alignment, how well is the DESAC sequence aligned, relative to other PylS sequences? Explain your reasoning.**
2. **In the MAFFT alignment, how well is the DESAC sequence aligned, relative to other PylS sequences? Explain your reasoning.**
3. *Compare the placement of gaps in your alignments*:
4. **What is your assessment of the quality of the MUSCLE alignment with respect to the placement of gaps? Explain your reasoning.**
5. **What is your assessment of the quality of the MAFFT alignment with respect to the placement of gaps? Explain your reasoning.**
6. *Overall comparison:* In this section, we have used several strategies for the comparison and assessment of alignments. Considering all of these comparisons, answer the following question.
7. **Overall, based on your answers in this section, do you prefer one alignment to the other? Explain your reasoning.**
8. **Save your MAFFT alignment in FASTA format as “<your name>-PylS-MafftL.fasta”**

**Section 2: The PylB gene family**

1. **Assess the phylogenetic distribution of your data set.**
2. *Open Jalview.*
3. *Open the PylB sequences :*
4. *Sort your sequences*:
5. *Remove redundant sequences:* .
   1. **How many sequences are there?**
   2. **How many sequences would you remove if you set the threshold to 90?**
   3. **Considering the first three characters of the five letter short names, how many genera would be completely removed from the alignment at the threshold of 90? What are they?**
   4. **How many sequences would you remove if you set the threshold to 65?**

**Considering the first three characters of the five letter short names, how many genera would be completely removed from the alignment at the threshold of 65?**

* 1. **What are they?**

1. **Assess alignment success using functional features.**
2. *Sequence alignment*
3. *Assessment of the alignment using functional features.* Look for the CxxxCxxC motif your alignments. For each conserved cysteine (C) in the motif,
4. **Note the following information in the table below:**
   * **The amino acid position in the *Methanosarcina acetivorans* (METAC) sequence.**
   * **The percentage of sequences that have this residue. To obtain this information, mouse over the bar corresponding to this column in the consensus histogram at the bottom of the alignment window. A tool tip will pop up that shows, for each residue, the percentage of sequences in this column that have residue.**
   * **Other amino acids found in this column, if any.**

|  |  |  |  |
| --- | --- | --- | --- |
| **IN UNIPROT** |  | **In THE MAFFT ALIGNMENT** | |
| **Binding site position in METAC** | **Residue in METAC** | **Most common amino acid (residue, %)** | **Other amino acids (residue, %)** |
| 71 | C | C (100%) |  |
|  |  |  |  |
|  |  |  |  |

1. **Is the CxxxCxxC motif present in all the sequences?**
2. **Has MAFFT properly aligned the CxxxCxxC motif across the majority of sequences?**
3. *Functional features*
4. **What is the other cofactor that is required for PylB function?**
5. **How many binding sites are there for this co-factor?**
6. **Note the following information in the table below:**
   * **The binding site position in the METAC sequence as specified on the Uniprot page.**
   * **The residue at this position in the METAC sequence.**
   * **The most frequent residue at this position and the percentage of residues that have that residue.**
   * **Other amino acids found in this column, if any. For each one, give the residue and the percentage of sequences in the column in which it appears.**

|  |  |  |  |
| --- | --- | --- | --- |
| **IN UNIPROT** |  | **In THE MAFFT ALIGNMENT** | |
| **Binding site position in METAC** | **Residue in METAC** | **Most common amino acid (residue, %)** | **Other amino acids (residue, %)** |
| 77 | F | F (100%) |  |
|  |  |  |  |

Add more rows as needed.

**Based on your table, answer the following questions:**

1. **Which sites are 100% conserved?**
2. **For those sites that are not 100% conserved, what is the most common biochemical group? To see the biochemical properties of amino acid, open the Jalview Documentation (F1) and look for "Amino Acid Properties" under "Useful Information" at the bottom of the table of contents. You may also find the Zappo color scheme helpful.**
3. **What percentage of the sequences in this column possesses a residue in primary biochemical group?**
4. **A residue with divergent biochemical properties in a strongly conserved column may indicate a loss of function mutation or a functional change. Based on the patterns of conservation in your table, are there any sequences that are candidates for such a hypothesis? If so, what are they?**

**Considering all of your analyses in this section:**

1. **In one or two sentences, why is it important to verify that this motif has been properly aligned?**
2. **Manual Re-alignment**
3. **Record the phylum (and subphylum, if any) from which this sequence originated.**
4. **Is the CxxxCxxC motif that you identified in Step B.2.A present in this sequence. Is it correctly aligned with the CxxxCxxC motif in other sequences?**
5. **Find the binding sites for the other co-factor in the DEMME sequence, using the METAC sequence as a reference. Enter the residue at each binding site in the table below.**

|  |  |  |
| --- | --- | --- |
| **Binding site position in METAC** | **Residue in METAC** | **Residue in DEMME** |
|  |  |  |
|  |  |  |

1. **In the DEMMA sequence, are the binding sites conserved (i.e., in the majority biochemical group)? If not, which ones differ?**
2. **Position the DEMMA insertion in the center of the alignment window. Take screen shot of and hand it in with your assignment.**
3. **Trimming your alignment**
4. **Save your trimmed alignment in Fasta format in a file. This file should be named:**

**<your name>-PylB-MAFFT-trimmed.fasta**