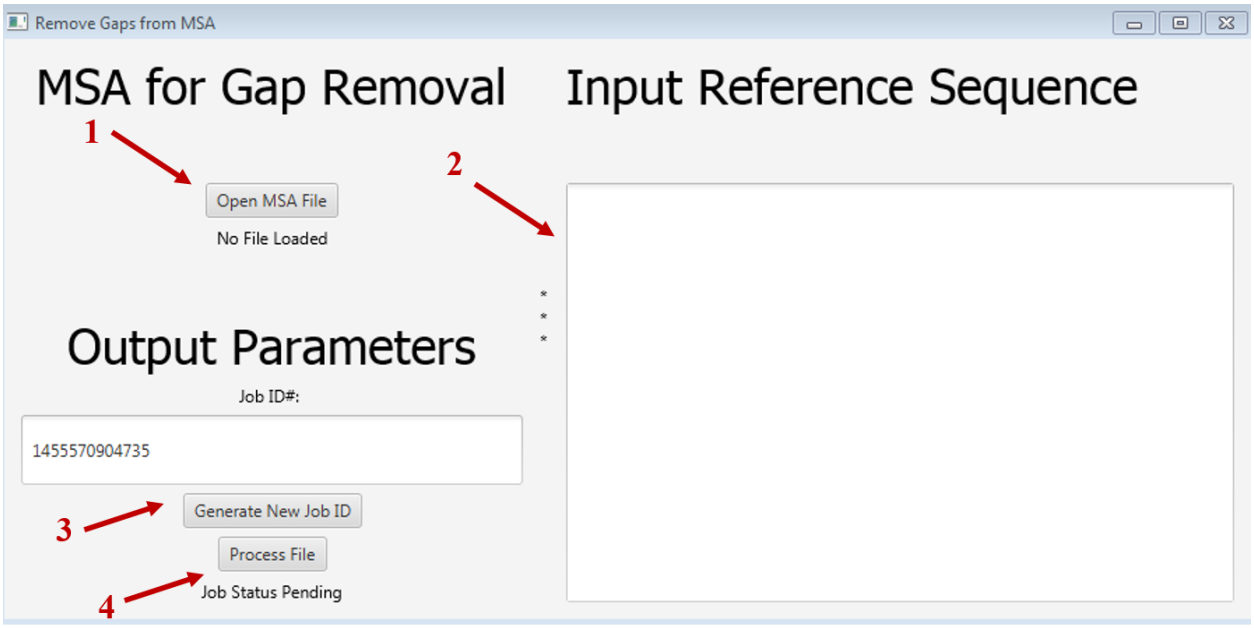
MSA Gap Remover Tool
User Manual
-Professor Camenares, Kingsborough Community College

Purpose:
This tool will take a multiple sequence alignment (MSA) and match the alignment to the full length of one of the sequences in the set. In other words, it will take a single sequence from the collection, provided or indicated by the user and identify all positions in the MSA in which that sequence has a gap, '-'. The program will then remove all characters found at those positions: all the gaps will be removed from the reference sequence, and the same positions (gap or not) will be removed from all other sequences in the collection.

Analysis of an MSA may produce information about one or more residues in the sequence; for example, some residue pairs may display co-evolution, by evidence of higher mutual information scores. However, the position of these residues is only important for the sequences obtained from a model organism; it is in this organism (such as *Escherichia coli)* any bio-informatics predictions will be tested.

Input:
The program is designed to utilize FASTA formatted MSAs obtained from the Clustal Omega program found at http://www.ebi.ac.uk/Tools/msa/clustalo/. In particular, the header for each FASTA sequence must have four "|" characters, with the fourth character separating the header from the actual amino acid or nucleotide sequence. Each sequence must be separated by starting with a ">" character. For an example, see the sample input file hisS_44_MSA

Interface:



1. **Open MSA File Button**: This button will open a .txt or .fa file with an MSA in a FASTA format; please see the Input section above for more information.

2. **Input Reference Sequence**: Here, you should copy and paste one entire FASTA sequence from the MSA you have loaded. Multiple FASTA sequences can be pasted, but only the first will be used as a reference sequence.

3. **Job ID:** This textfield, with a timestamp generated number, will be used to name the directory in which the output files will be placed. You may rename the folder by changing this value, or generate a new ID (so that previous work is not overwritten) by hitting the 'Generate New ID' button. Note that a new ID is automatically generated each time the program is loaded anew.

4. **Process Button:** Selecting this button will process the inputs; an error message will be displayed if the one or more inputs are missing.

Additional Notes:
Located alongside the executable JAR file are two sample inputs: hisS_44_MSA.txt and hisS_44_Ref_Test.txt, which represent an MSA for histidine tRNA synthetase (with several thousand genes) and a single FASTA sequence pulled from that MSA, respectively. You can load the first file into the program (using feature #1, above) and then copy and paste the contents of the second file into feature #2, above.

If you run the sample files, you should obtain the same result as the file hisS_44_MSA_no_gaps.txt. In this example, the 4th sequence in the MSA was used as a reference, and you can see that sequence in the resulting file has no gaps.