Universidad de los Andes

Departamento de Ingeniería de Sistemas y Computación

UNACLOUD Multiple Sequence Alignment Tool

UnaCloud MSA Guide

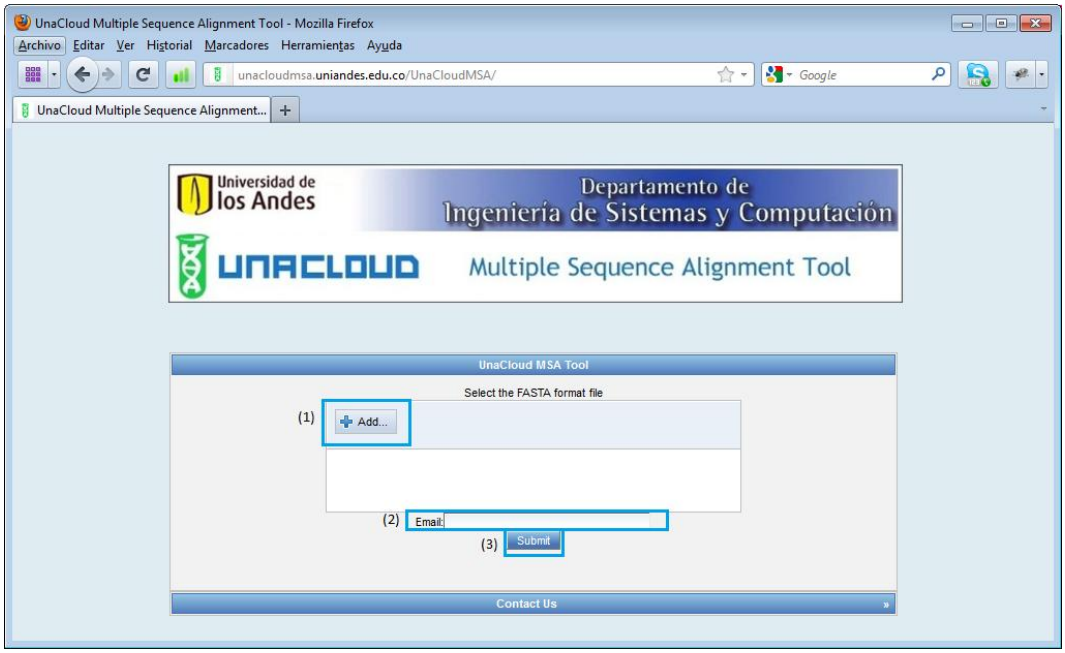
Submitting a Job

Getting started with UnaCloud MSA is very simple. In order to submit a job the user has to:

(1) Upload a FASTA formatted file indicating the sequences to be analyzed. Currently UnaCloud MSA only supports analysis for protein based sequences. The sequences in the FASTA format should not be aligned.

(2) Introduce a valid e-mail address. When the job is complete, the user will receive a notification with the link where the results are available.

(3) Click the submit button



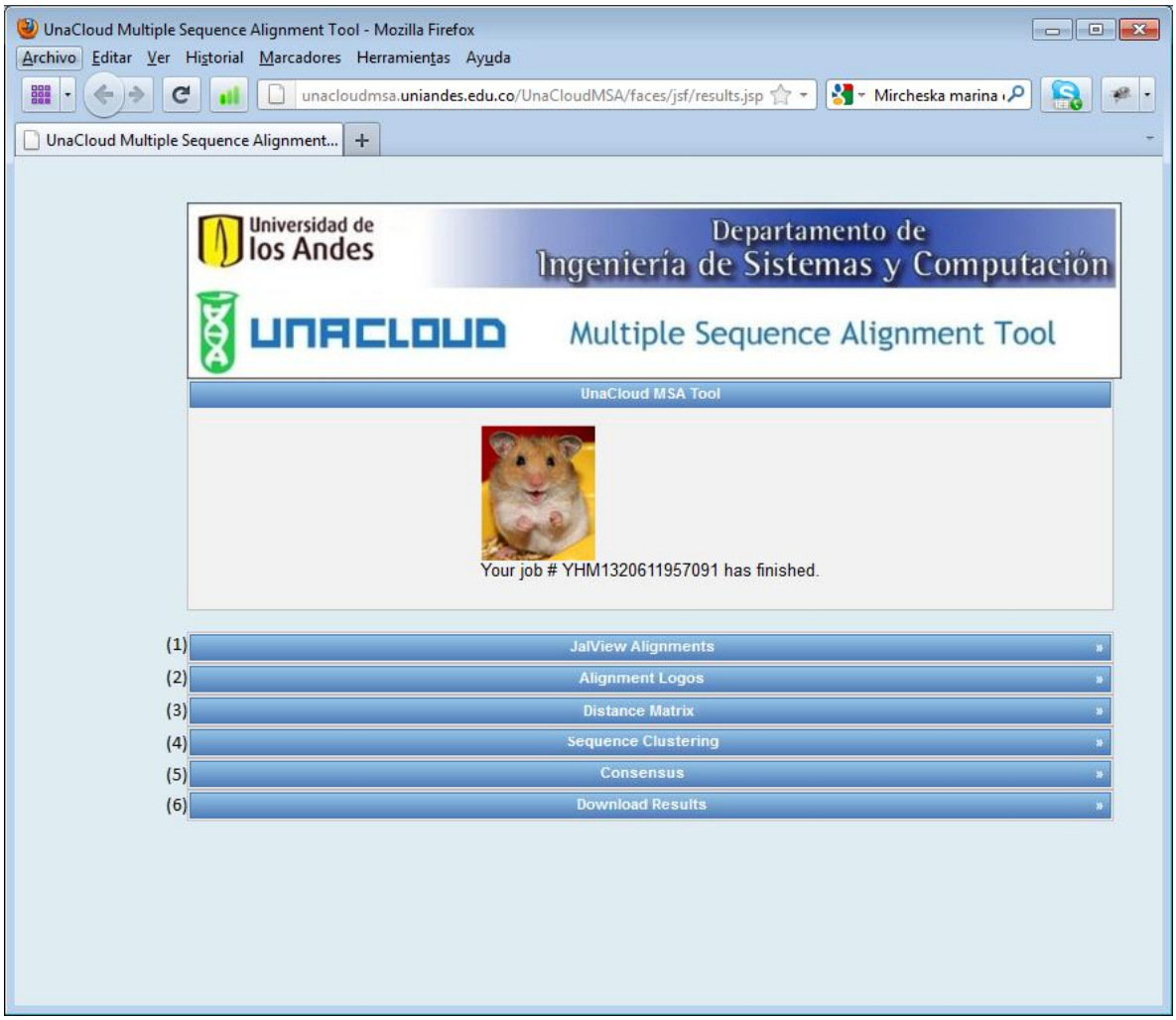
Depending on the number of sequences in the introduced file, the performed workflow has some changes. If the number of sequences is lower or equal to 100, the programs used to perform the MSA are: CLUSTAL W2 (Thompson et al., 2007), MAFFT (Katoh et al., 2009), MUSCLE (Edgar, 2004), T-Coffee (Notredame et al., 2000) and PROBCONS (Do et al., 2005). If the number of sequences is higher, the programs T-Coffee and PROBCONS are replaced by the program KAlign (Lassman et al.,2009). This separation is done because of the high computational cost of these 2 alignment applications.

Guide prepared by the UnaCloud MSA team
Contact: unacloudmsa@uniandes.edu.co

Analyzing results:

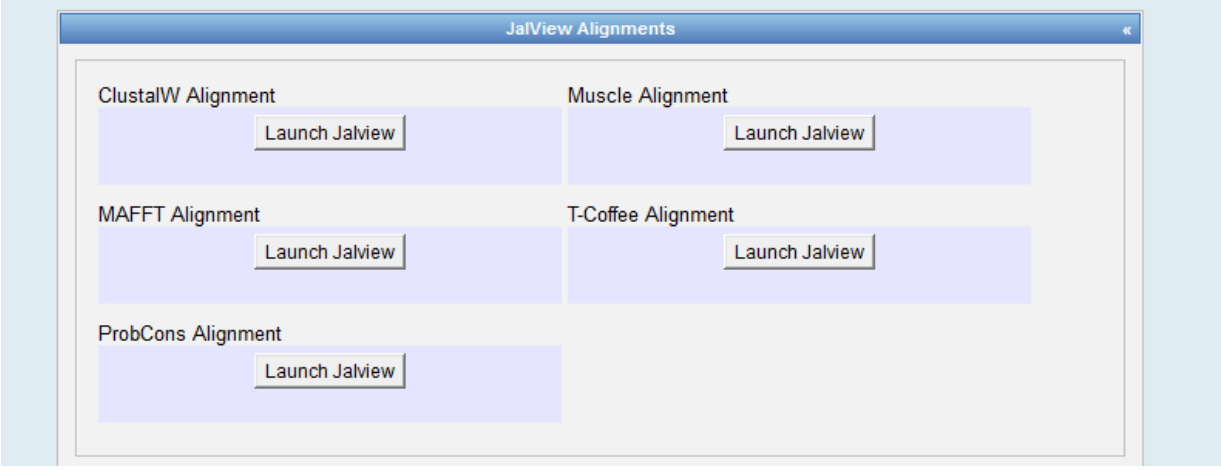
When the job is complete, the user will receive an email with the link to the result page.

The results are divided into 6 categories. Each category is presented in its own panel. In order to view one category of the results, the user has to click the (>>) button that expands the specific panel. Each one of the categories are explained below.

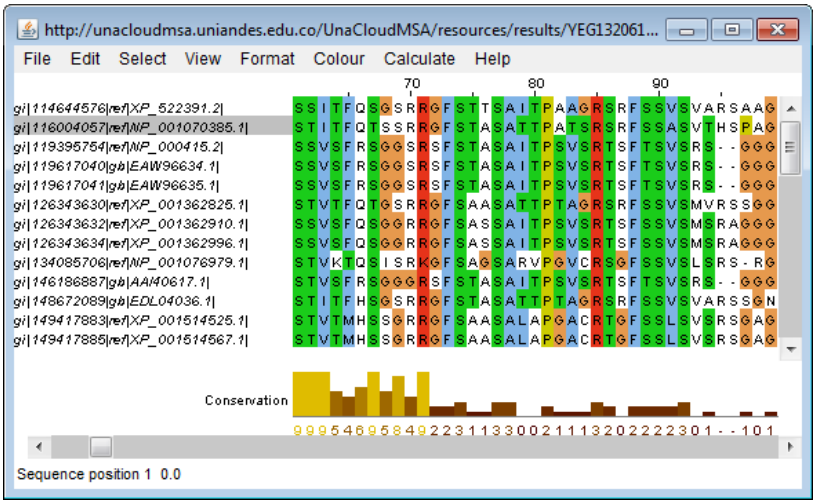


(1) Alignments and Jalview

In this category, each one of the generated MSA is presented. The presentation of the MSA is done with the visualization tool JalView (Waterhouse et al., 2009).



The user can select each one the different alignments generated and a new JalView window will appear presenting the alignment.

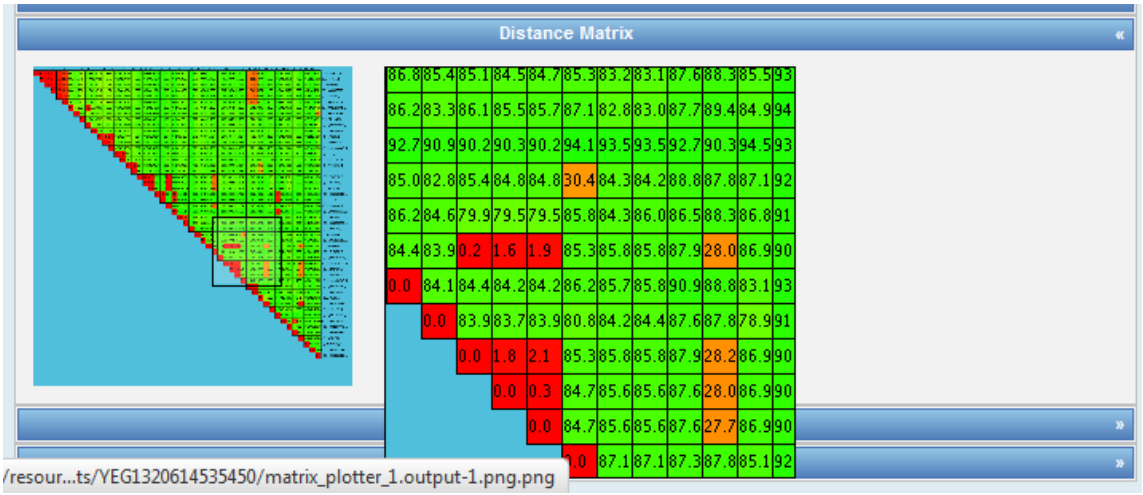


(2) Sequence Logos

In this category, the user will find, for each generated MSA, its sequence logo representation.
Sequence logos, are a very useful image representation of a MSA where each column is represented as a stack of the most frequent aminoacids in that position. The Seqlogo(Crooks et al 2004) tool was used for this visualization.
If the user wants to focus in a specific area of a logo, the application counts with a zoom utility.
The user can also click a logo; this will open the selected image.

(3) Distance Matrix

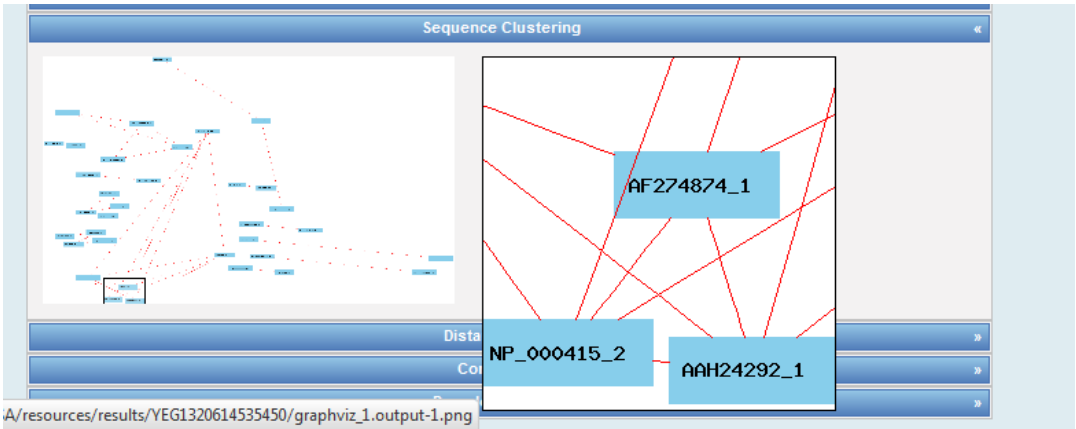
When performing a MSA, the quality of the results heavily depends on the quality of the input sequences. If the input sequences have a high level of similarity, the resulting alignment could be biased towards these groups of similar sequences. The `dist` tool from the EMBOSS suite performs a pairwise comparison between the sequences. Since the number of entries in the matrix grows very fast as the number of sequences increases, we implemented the application `matrix_plotter`. `matrix_plotter` generates an image of the matrix with a color scale, where values closer to 0 appear in red and values closer to 100 appear in green. With this image it becomes easier to identify groups of sequences that are highly similar



In the image the user can identify some red entries. These entries correspond to sequences that have a very high level of similarity and should be treated carefully in the following analysis.

(4) Sequence Clustering

In this category the user will find a graph representation of the clustering of the sequences.
In this graph, the vertices are sequences, and the length of the arcs represents the distance between the sequences. If two sequences are very close in the graph, it means that there is a high level of similarity between them and the user have to decide what to do with this problematic sequences. With this visual representation, the user can identify not only couples, but groups of highly similar sequences.



In the image the user can identify a cluster. The sequences with identifiers AF274874_1, NP_000415_2 and AAH24292_1 belong to the same cluster and have a very low distance between them. This indicated that this group can degrade the quality of the alignment.

(5) Alignment Comparison and Consensus

A consensus sequence takes the most frequent symbol in each column of the alignment. It summarizes the complete alignment in a single sequence. For each one of the generated alignment, the application presents its consensus. The user can view each one of the generated consensus.

The second part of this category is the comparison of the alignments. Since each program employs a different strategy, the resulting alignments can differ. Using the consensus of each alignment, we developed the tool `alignment_comparator` that renders an image indicating the consistent and the ambiguous areas of the alignments. A consist area of an alignment, is an area that each of the programs aligns it in the same way. An ambiguous area is an area that different programs align in different ways. The consistent areas are depicted with green color (4 or 5 alignments agreed), the regular areas are painted with yellow and orange color (3 and 2 alignments) and the ambiguous areas are painted with red.

(6)Results Download

The results are kept available online for a limited amount of time. Because of this, the user has the option to download the results to its own computer. This download consists of a compressed zip file that contains all the results mentioned above. It also includes the original sequence file that the user uploaded.

References:

Crooks, G. et al (2004) WebLogo: A sequence logo generator. Genome Research, 14, 1188-1190.

Do, C.B. et al. (2005) Probabilistic Consistency-based Multiple Sequence Alignment. Genome Research, 15, 330-340

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32, 1792-1797.

Katoh, K. et al. (2009) Multiple Alignment of DNA Sequences with MAFFT. Methods in Molecular Biology, 537, 39-64

Lassman, T. et al. (2009) Kalign: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res, 37, 858-865

Notredame, C. et al. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol., 302, 205-217

Thompson, J.D. et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947-2948

Waterhouse, A.M., et al. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25, 1189-1191.