Overview of Multiple Sequence Alignment Algorithms

Yu He
04/13/2016

Adapted from the multiple sequence alignment presentations by Mingchao Xie and Julie Thompson

Last update: 08/09/2021

1

Multiple sequence alignments

Multiple Sequence Alignment (MSA) can be seen as a generalization of a **Pairwise Sequence Alignment** (PSA). Instead of aligning just two sequences, three or more sequences are aligned simultaneously.

MSA is used for:

- Detection of conserved domains in a group of genes or proteins
- Construction of a phylogenetic tree
- Prediction of a protein structure (*e.g.*, AlphaFold, RoseTTAFold)
- Determination of a consensus sequence (*e.g.,* transposons)

2

Multiple sequence alignments

Example: part of an alignment of globin from 7 sequences

Symbol	Meaning
*	Fully conserved
:	Conservation between groups of amino acids with strongly similar properties
.	Conservation between groups of amino acids with weakly similar properties
	Not conserved

3

Alignment algorithms

Three types of algorithms:

1. Progressive: ***Clustal W***
2. Iterative: ***MUSCLE*** (multiple sequence alignment by log-expectation)
3. Hidden Markov models: ***HMMER***

***Clustal Omega*:** Iterative progressive alignment using hidden Markov models

4

Step 1 : Pairwise alignment of all sequences

Example : Alignment of 7 globins (Hbb_human, Hbb_horse, Hba_human, Hba_horse, Myg_phyca, Glb5_petma and Lgb2_lupla)

The alignment can be obtained with:

- global or local methods
- heuristic methods

Adapted from Julie Thompson, IGBMC

5

Step 2 : Distance matrix construction

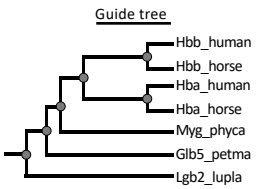
Distance between two sequences = 1 - (No. identical residues / No. aligned residues)

		1	2	3	4	5	6	7
Hbb_human	1	-						
Hbb_horse	2	.17	-					
Hba_human	3	.59	.60	-				
Hba_horse	4	.59	.59	.13	-			
Myg_phyca	5	.77	.77	.75	.75	-		
Glb5_petma	6	.81	.82	.73	.74	.80	-	
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-

Adapted from Julie Thompson, IGBMC

6

Step 3 : Guide tree construction



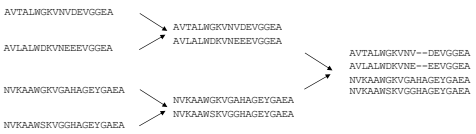
Hbb_human	1	-						
Hbb_horse	2	.17	-					
Hba_human	3	.59	.60	-				
Hba_horse	4	.59	.59	.13	-			
Myg_phyca	5	.77	.77	.75	.75	-		
Glb5_petma	6	.81	.82	.73	.74	.80	-	
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-
		1	2	3	4	5	6	7

UPGMA clustering method:
- Join the two closest sequences, create consensus
- Recalculate distances and join the two closest sequences or nodes
- Step 2 is repeated until all sequences are joined

Adapted from Julie Thompson, IGBMC

7

Step 4 : Progressive alignment



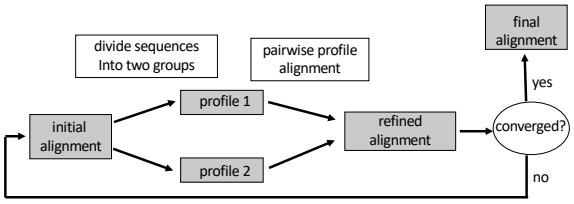
The sequences are aligned progressively (global or local algorithm):
- alignment of 2 sequences, create profile (consensus)
- alignment of 1 sequence and a profile (group of sequences)
- alignment of 2 profiles (groups of sequences)

Adapted from Julie Thompson, IGBMC

8

Iterative alignment

Iterative alignment refines an initial progressive multiple alignment by iteratively dividing the alignment into two profiles and realigning them.

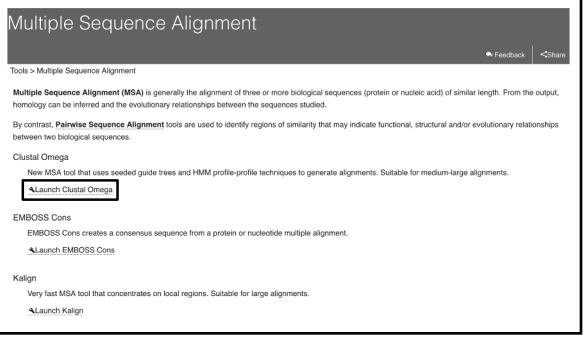


Adapted from Julie Thompson, IGBMC

9

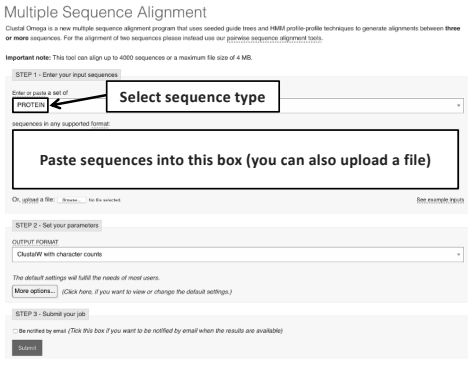
Clustal Omega

Navigate to https://www.ebi.ac.uk/Tools/msa/



10

Clustal Omega: setting up



11

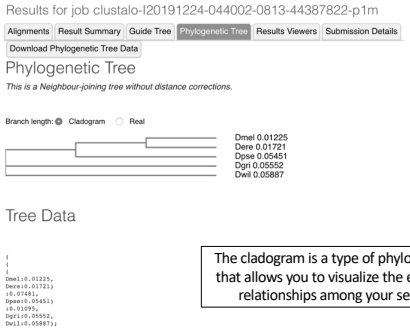
Drosophila eyeless protein sequences

12

Clustal Omega results — alignments

13

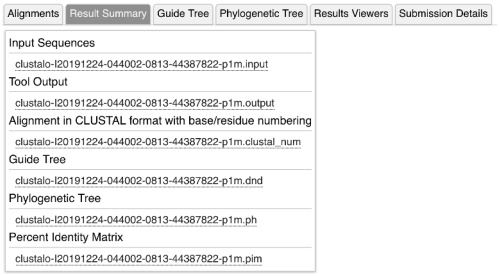
Clustal Omega results — phylogenetic tree



The cladogram is a type of phylogenetic tree that allows you to visualize the evolutionary relationships among your sequences

14

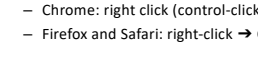
Clustal Omega results — result summary



15

Use *Jalview Desktop* to visualize the alignment

- Download *Jalview Desktop*:
  - https://www.jalview.org/getdown/release/
- Copy the link to the *Clustal Omega* alignment

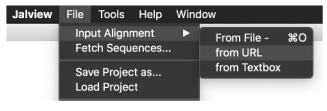


  - Chrome: right click (control-click on macOS) ➔ **Copy Link Address**
  - Firefox and Safari: right-click ➔ **Copy Link**

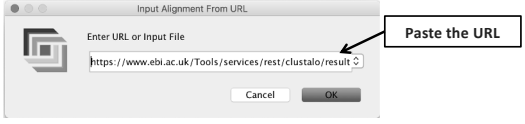
16

Open the *Clustal Omega* alignment in *Jalview Desktop*

- Select File ➔ Input Alignment ➔ from URL

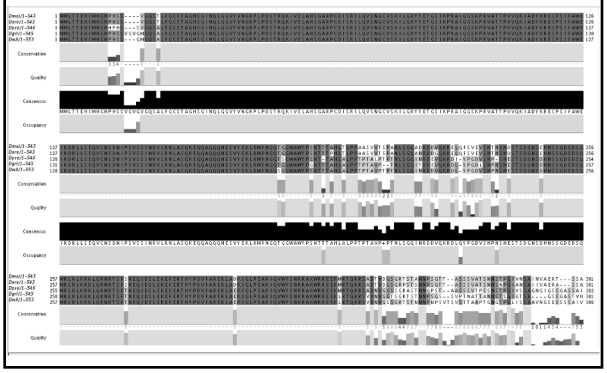


- Paste the URL into the textbox ➔ click "OK"



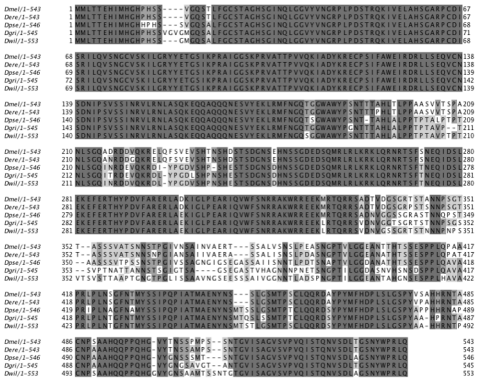
17

Use *Jalview Desktop* to color the *Clustal Omega* alignment by percent identity



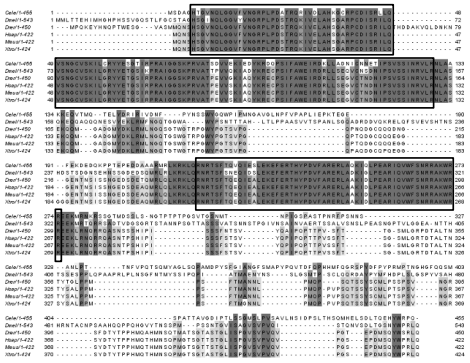
18

Alignment for the *Drosophila* eyeless protein



19

Alignment for the eyeless protein in a broader range of species



20

Conclusions

- *Clustal Omega* uses a modified iterative progressive alignment method and can align over 10,000 sequences quickly and accurately
- *Clustal Omega* is very useful for finding evidence of conserved function in DNA and protein sequences
  - But remember that sequence similarity does not always imply conserved function!
- *Clustal Omega* can be used to find promoters and other cis-regulatory elements

21