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: 12/21/2017

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
- Determination of a consensus sequence (*e.g.,* transposons)

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

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

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

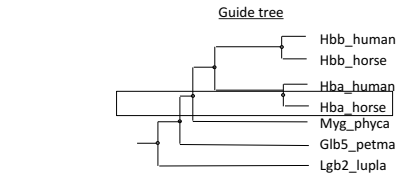
Step 2 : Distance matrix construction

Distance between two sequences = 1 - $\frac{\text{No. identical residues}}{\text{No. aligned residues}}$

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

Adapted from Julie Thompson, IGBMC

Step 3 : Guide tree construction

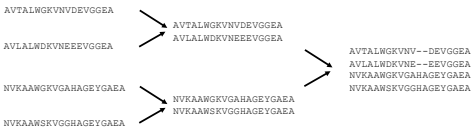


Hbb_human	1	-						
Hbb_horse	2	.17	-					
Hba_human	3	.59	.60	-				
Hba_horse	4	.59	.59	.13	-			
Myg_phyca	5	.77	.77	.75	.75	-		
Glb5_petma	6	.81	.82	.73	.74	.80	-	
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-
		1	2	3	4	5	6	7

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

Step 4 : Progressive alignment

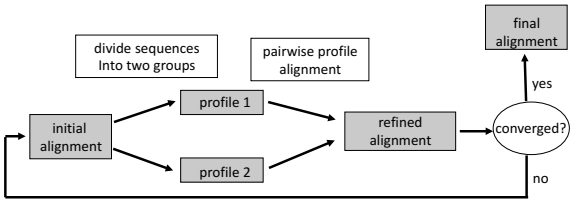


The sequences are aligned progressively (global or local algorithm):
- alignment of 2 sequences, create profile (consensus)
- alignment of 1 sequence and a profile (group of sequences)
- alignment of 2 profiles (groups of sequences)

Adapted from Julie Thompson, IGBMC

Iterative alignment

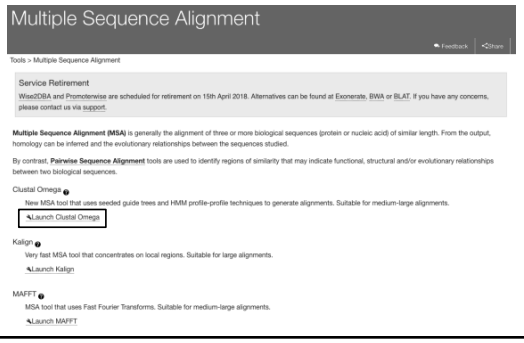
Iterative alignment refines an initial progressive multiple alignment by iteratively dividing the alignment into two profiles and realigning them.



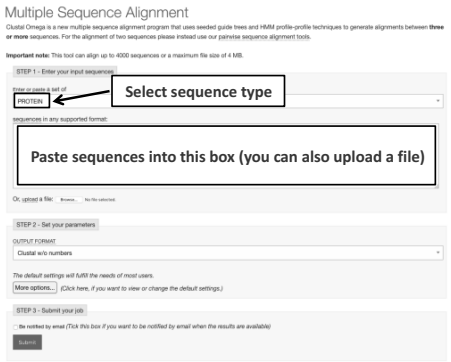
Adapted from Julie Thompson, IGBMC

Clustal Omega

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



Clustal Omega: setting up

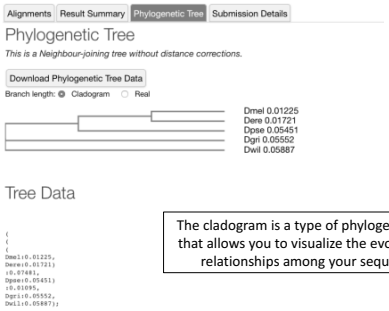


Drosophila eyeless protein sequences

>Dmel
MMLTTEHIMHGHPHSSVGQSTLFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIKPRAIGGSKPRVATTPVVQKIA
DYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSINRVLRNLASQKEQQAQQQNESVYEKLRMFNGQTGGWAWYPSNTTTAHLTPPAASVVTSPANLSGQADRDDVQK
RELQFSVEVSHTNSHDSTSDGNSEHNSSGDEDSQMRLRLKRKLQRNRTSFSNEQIDSLEKEFERTHYPDVFARERLADKIGLPEARIQVWFSNRRAKWRREEKMRTQRRSA
DTVDGSGRTSTANNPSGTTASSSVATSNNSTPGIVNSAINVAERTSSALVSNSLPEASNGPTVLGGEANTTHTSSESPPLQPAAPRLPNSGFNTMYSSIPQPIATMAENYNS
SLGSMTPSCLQQRDAYPYMFHDPLSLGSPYVSAHHRNTACNPSAAHQQPPQHGVYTNSSPMPSSNTGVISAGVSVPVQISTQNVSDLTGSNYWPRLQ
>Dgri
MMLTTEHIMHGHPHSSVGVGMGQSALFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIKPRAIGGSKPRVATTPV
VQKIADYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSINRVLRNLASQKEQQAQQQNESVYEKLRMFNGQTGGWAWYPGNTTTAHLALPPTPTAVPTNLSGQITRDEV
QKRDLYPGDLSHPNSHESTSDGNSDHNSSGDEDSQMRLRLKRKLQRNRTSFTNEQIDSLEKEFERTHYPDVFARERLAEKIGLPEARICVWFSNRRAKWRREEKLRTQRRSV
DNVGGTSGRTSTNNPSGSSVPTNATTANNSTSGIGTSAGSEGASTVHAGNNNPNETSNGPTILGGDASNVHSNSDSPPLQAVAPRLPLNTGFNTMYSSIPQPIATMAENY
NSMTQSLSSMTPTCLQQRDSYPYMFHDPLSLGSPYAAHPRNTACNPAAAHQQPPQHGVYGNGSAVGTANTGVISAGVSVPVQISTQNVSDLTGSNYWPRLQ
>Dwil
MMLTTEHIMHGHPHSSGMGQSALFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIKPRAIGGSKPRVATTPVVQK
IADYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSINRVLRNLASQKEQQAQQQNESVYEKLRMFNGQSGGWAWYPSNTTTAHLALPPTPTAVPTPTNLSGQINRDDVQK
RDLYPGDVSHPNSHESTSDGNSDHNSSGDEDSQMRLRLKRKLQRNRTSFTNEQIDSLEKEFERTHYPDVFARERLAEKIGLPEARICVWFSNRRAKWRREEKLRTQRRSVD
NVGSSGRTSTNNNPNPSVTSVSTTAAPTGNGTPGLISSAAVNGSEESSSAIVGGNNTLADSPNGPTILGGEANTAHGNSESPPLHAVAPRLPLNTGFNTMYSSIPQPIATMA
ENYNSMTSTLGSMTPSCLQQRDSYPYMFHDPLSLGSPYAAHHRNTPCNPSAAHQQPPQHGGVYGNSAAMTSSNTGTGVISAGVSVPVQISTQNVSDLAGSNYWPRLQ
>Dere
MMLTTEHIMHGHPHSSVGQSTLFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIKPRAIGGSKPRVATTPVVQKIA
DYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSINRVLRNLASQKEQQAQQQNESVYEKLRMFNGQTGGWAWYPSNTTTHLTPPAASVVTSPANLSGQANRDDGQK
RELQFSVEVSHTNSHDSTSDGNSEHNSSGDEDSQMRLRLKRKLQRNRTSFSNEQDSLEKEFERTHYPDVFARERLADKIGLPEARICVWFSNRRAKWRREEKMRTQRRSA
DTVDGSGRPSTSNNPSGTTASSSVATSNNSNPGIANSAIIVAERASSALISNSLPDASNGPTVLGGEANATHTSSESPPLQPATPRLPLNSGFNTMYSSIPQPIATMAENYNSS
LGSMTPSCLQQRDAYPYMFHDPLSLGSPYVPAHHRNTACNPAAAHQQPPQHGVYTNSSAMPSSNTGVISAGVSVPVQISTQNVSDLAGSNYWPRLQ
>Dpse
MMLTTEHIMHGHHPHSSVGQSALFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIKPRAIGGSKPRVATTPVVQKI
ADYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSINRVLRNLASQKEQQAQQQNESVYEKLRMFNGQTGGWAWYPSNTTAHLALPPTPTAPTPTNLSGQINRDEVQKRD
IYPGDVSHPSHESTSDGNSDHNSSGDEDSQMRLRLKRKLQRNRTSFTNEQIDSLEKEFERTHYPDVFARERLAEKIGLPEARICVWFSNRRAKWRREEKMRTQRRSADNVG
GSSGRASTNNQPSTAASSSVTPSSNSTPGIVSSAGNGIGSEGASSAIISNNTLPDTSNAPTVLGGDANATHTSSESPPLQAVAPRIPLNAGFNAMYSSIPQPIATMAENYNSM
TSSLGSMTPTCLQQRDSVPYMFHDPLSLGSPYAPPHHRNAPCNPAAAHQQPPQHGVYGNSSSMTSNTGVISAGVSVPVQISTQNVSDLAGSNYWPRLQ

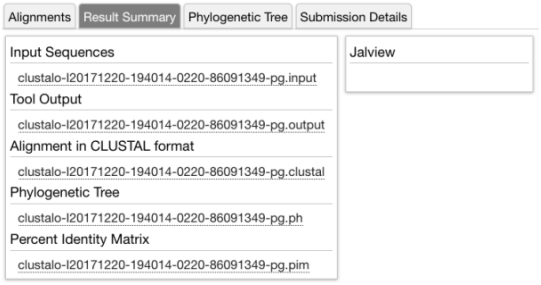
Clustal Omega results — alignments

Clustal Omega results — phylogenetic tree

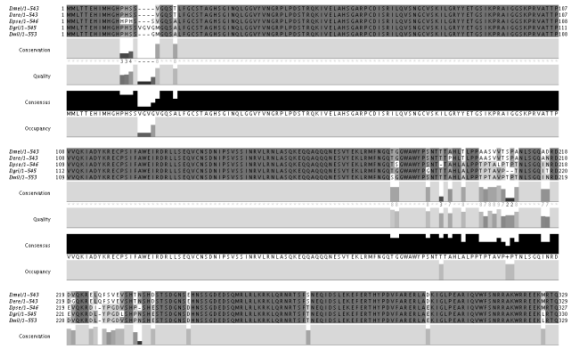


The cladogram is a type of phylogenetic tree that allows you to visualize the evolutionary relationships among your sequences

Clustal Omega results — result summary

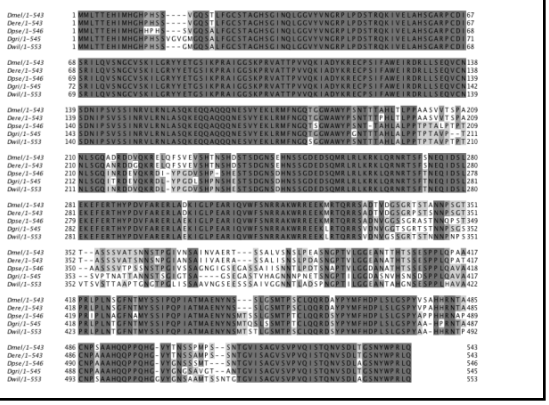


Use Jalview to visualize the alignment

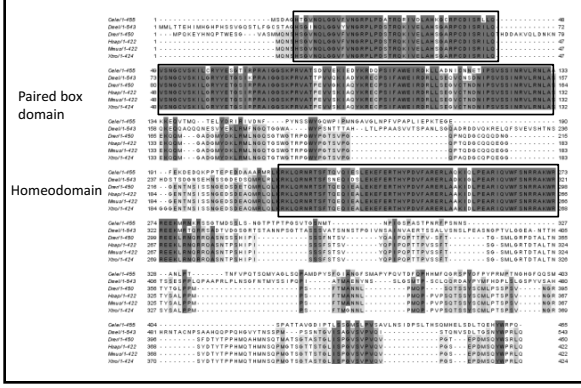


Download Jalview at http://www.jalview.org/download

Alignment for the *Drosophila* eyeless protein



Alignment for the eyeless protein in a broader range of species



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