

Overview of Multiple Sequence Alignment Algorithms

Yu He
04/13/2016
Adapted from Mingchao Xie and Julie Thompson
multiple sequence alignment ppt online

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

```

LGB2_LUPLU  -----GALTESQAALVKSWEFFIANPKHTRFFILVLEIAPAADLISFLKGTSE
MYG_PHYCA  -----VLSGEQWLIVHWIAVHADVAAGHGDILIRLFKSPETLEKDFRFKHLKT
GLB5_PETMA  PIVDTGVSVAFLSAAEKTKTISAWAPVSTYETSGVDILVKFFTSPPAAQEFPPKFKGLTT
HBA_HUMAN  -----VLSPADKTNVKAAWGIVHAGGEYGAEALERMFLEFPPTTKTYVPHFDL---
HBA_HORSE  -----VLSPADKTNVKAAWGIVHAGGEYGAEALERMFLEFPPTTKTYVPHFDL---
HBB_HUMAN  -----VHLTPBEKSAVIALWGHVIV--DEVGGEALGRLLVYPWTQRFEFESFGOLST
HBB_HORSE  -----VQLSGEKAAVIALWDHIVE--EEVVGGEALGRLLVYPWTQRFEFESFGOLSN
    
```

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

```

Hbb_human 1 VLTPEEKSAVTALMGK...VYRGTQDAGELLVYVPTQDFPFQDLS...
Hbb_horse 2 VQLSGEKAVALMGK...VYRGTQDAGELLVYVPTQDFPFQDLS...
Hbb_human 1 LTPPEEKSAVTALMGK...VYRGTQDAGELLVYVPTQDFPFQDLS...
Hba_human 3 LSPADKTVKAWGIVHAG...GEYGAEALERMFLEFPPTTKTYVPHFDL...
Hba_human 3 LSPADKTVKAWGIVHAG...GEYGAEALERMFLEFPPTTKTYVPHFDL...
Hbb_horse 2 LSGEKAVALMGK...VYRGTQDAGELLVYVPTQDFPFQDLS...
    
```

Adapted from Julie Thompson, IGBMC

Step 2 : Distance matrix construction

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

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

Guide tree

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

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 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: <http://www.ebi.ac.uk/Tools/msa/>

Multiple Sequence Alignment

Clustal Omega

- MUSCLE** - Accurate MSA tool, especially good with proteins. Suitable for medium alignments. Suitable for medium-large alignments.
- MAFFT** - Very fast MSA tool that concentrates on local regions. Suitable for large alignments.
- T-Coffee** - Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.
- WVA/FRANK** - The FRANK has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help guide treebuilding and clustering. Try it out at <http://www.ebi.ac.uk/Tools/msa/frank/>

Clustal Omega: setting up

Drosophila eyeless protein sequences

```

>Dmel
MMLTTEHMHGHPHSSVGQSTLFCGCTAGHSGINQLGGVYNGRPLDSTRQKVELHSGARPCDSIRLQVSNVCYKILGRYETGSKPRAIGGSKPRVATTPVQKIA
DYKRCPSFAWEIRDRLSEQVCSNDSPVSNIRVLRNLASQEQDAQQDNESYEKLRMFGQTSQWAWYPSNTTTHALPPTTAVPTTFLNSGQINRDQVQK
RDLPGDVSHPNHESTSDGNSHNSGDEDSQMRRLRKLQNRNFSNEQDLEKFERFHPDVFARELADKGLPEARQVWFSNRRAKWRREKMRTOBRSIA
DTVGGSRISTANNPSTGTAASSVATSNKTPGVGNAINVAERLTSALVSNLQVAGSNPGLVGGAAATNTTSSPPPLDAPRPLFSGGNTNYSYSPQATMAEYNS
SUGMSTFCLQQRDAPYPMFHDRLSLGSPVPAHHRNTACNPAAHQQPQHQVYTSRMPSSNTGVSAGVSPVQISTQVNSDLTGSYVWRRLQ
>Dgr1
MMLTTEHMHGHPHSSVGQSTLFCGCTAGHSGINQLGGVYNGRPLDSTRQKVELHSGARPCDSIRLQVSNVCYKILGRYETGSKPRAIGGSKPRVATTPVQK
ADYKRCPSFAWEIRDRLSEQVCSNDSPVSNIRVLRNLASQEQDAQQDNESYEKLRMFGQTSQWAWYPSNTTTHALPPTTAVPTTFLNSGQINRDQVQK
RDLPGDVSHPNHESTSDGNSHNSGDEDSQMRRLRKLQNRNFSNEQDLEKFERFHPDVFARELADKGLPEARQVWFSNRRAKWRREKMRTOBRSIA
DTVGGSRISTANNPSTGTAASSVATSNKTPGVGNAINVAERLTSALVSNLQVAGSNPGLVGGAAATNTTSSPPPLDAPRPLFSGGNTNYSYSPQATMAEYNS
SUGMSTFCLQQRDAPYPMFHDRLSLGSPVPAHHRNTACNPAAHQQPQHQVYTSRMPSSNTGVSAGVSPVQISTQVNSDLTGSYVWRRLQ
>Dmel
MMLTTEHMHGHPHSSVGQSTLFCGCTAGHSGINQLGGVYNGRPLDSTRQKVELHSGARPCDSIRLQVSNVCYKILGRYETGSKPRAIGGSKPRVATTPVQK
ADYKRCPSFAWEIRDRLSEQVCSNDSPVSNIRVLRNLASQEQDAQQDNESYEKLRMFGQTSQWAWYPSNTTTHALPPTTAVPTTFLNSGQINRDQVQK
RDLPGDVSHPNHESTSDGNSHNSGDEDSQMRRLRKLQNRNFSNEQDLEKFERFHPDVFARELADKGLPEARQVWFSNRRAKWRREKMRTOBRSIA
DTVGGSRISTANNPSTGTAASSVATSNKTPGVGNAINVAERLTSALVSNLQVAGSNPGLVGGAAATNTTSSPPPLDAPRPLFSGGNTNYSYSPQATMAEYNS
SUGMSTFCLQQRDAPYPMFHDRLSLGSPVPAHHRNTACNPAAHQQPQHQVYTSRMPSSNTGVSAGVSPVQISTQVNSDLTGSYVWRRLQ
>Dgr1
MMLTTEHMHGHPHSSVGQSTLFCGCTAGHSGINQLGGVYNGRPLDSTRQKVELHSGARPCDSIRLQVSNVCYKILGRYETGSKPRAIGGSKPRVATTPVQK
ADYKRCPSFAWEIRDRLSEQVCSNDSPVSNIRVLRNLASQEQDAQQDNESYEKLRMFGQTSQWAWYPSNTTTHALPPTTAVPTTFLNSGQINRDQVQK
RDLPGDVSHPNHESTSDGNSHNSGDEDSQMRRLRKLQNRNFSNEQDLEKFERFHPDVFARELADKGLPEARQVWFSNRRAKWRREKMRTOBRSIA
DTVGGSRISTANNPSTGTAASSVATSNKTPGVGNAINVAERLTSALVSNLQVAGSNPGLVGGAAATNTTSSPPPLDAPRPLFSGGNTNYSYSPQATMAEYNS
SUGMSTFCLQQRDAPYPMFHDRLSLGSPVPAHHRNTACNPAAHQQPQHQVYTSRMPSSNTGVSAGVSPVQISTQVNSDLTGSYVWRRLQ
    
```


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.