

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

```

LGB2_LUPLU  -----GALTESDAALVKSWEFFNANIPKHTHRFFILVLEIAPAAKDLISFLKGTSE
MYG_PHYCA  -----VLSGEENQVLVHWIAKVEADVAAGHGQDILIRLFKSPFETLEKDFDFKHLKLT
GLB5_PETMA  PIVDTGSAVPLSAAEKTITRSAWAPVYSTVETSQVDILVKFFSTPAAQEFPPKFKGLTT
HBA_HUMAN  -----VLSPADKTNVKAAWKMGVGHAGVEYGAEALERMFLSPPTTKTYFPHFDL---
HBB_HUMAN  -----VLSAADKTNVKAAWKMGVGHAGVEYGAEALERMFLSPPTTKTYFPHFDL---
HBB_HORSE  -----VHLTPPEKSAVTALWGMVW--DEVGGEALGRLLVVPWTQRFESFGDLST
HBB_HORSE  -----VQLSGEKAAYIALWDMVNF--EEVGGEALGRLLVVPWTQRFESFGDLST
  
```

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

```

Hbb_human 1 VHLTPPEKSAVTALWGMVW--DEVGGEALGRLLVVPWTQRFESFGDLST ...
Hbb_horse 2 VQLSGEKAAYIALWDMVNF--EEVGGEALGRLLVVPWTQRFESFGDLST ...
Hba_human 3 LSPADKTNVKAAWKMGVGHAGVEYGAEALERMFLSPPTTKTYFPHFDL ...
Hba_horse 4 LSPADKTNVKAAWKMGVGHAGVEYGAEALERMFLSPPTTKTYFPHFDL ...
Myg_phyca 5 PIVDTGSAVPLSAAEKTITRSAWAPVYSTVETSQVDILVKFFSTPAAQEFPPKFKGLTT ...
Glb5_petma 6 PIVDTGSAVPLSAAEKTITRSAWAPVYSTVETSQVDILVKFFSTPAAQEFPPKFKGLTT ...
Lgb2_lupla 7 -----GALTESDAALVKSWEFFNANIPKHTHRFFILVLEIAPAAKDLISFLKGTSE
  
```

Adapted from Julie Thompson, IGBMC

5

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

6

Step 3 : Guide tree construction

Guide tree

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, IGBC

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, IGBC

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, IGBC

9

Clustal Omega

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

Multiple Sequence Alignment

Tools > Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, **Pairwise Sequence Alignment** tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium to large alignments.

EMBOSS Cons

EMBOSS Cons creates a consensus sequence from a protein or nucleotide multiple alignment.

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

10

Clustal Omega: setting up

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of few sequences please instead use our [pairwise sequence alignment tools](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of **PROTEIN** sequences in any supported format

Paste sequences into this box (you can also upload a file)

OR, upload a file:

STEP 2 - Set your parameters

OUTPUT FORMAT: ClustalW with character counts

The default settings will fulfil the needs of most cases.

[More options...]

STEP 3 - Submit your job

* Be notified by email (Tick this box if you want to be notified by email when the results are available)

11

Drosophila eyeless protein sequences

```

>Dmel
MMLTTEHHMHHGHPHSSVGGSTLFGCTAGHSGINQLGQVYVNGRPLPDRTRQKVELAHSGARPCDSIRLQVNSGVSKLGRYETGSIKPRAIGSKPRVATTPVQKIA
DYKRECFSAWEIRDRLLEQVCSNDSPVSSINRVLNLSAQEQDAQDQNESVEYKLRMFGQSGWAWYPSNTTTHALPPTTAVPTNLGQNRDDQK
RELQSVESVHTNSDSDGNSHNSGDEDSQMLRKLKLNRTSPNEQDLEKEFERTHPDVFARELAKGLPEARLQVWFSNRRAKWRREKMRTRBSA
DTVDSGRTSTANPSSGTTASSVATSNSTGIVNSAINVAERTSSALVNSLPEASNGPTVLGGEAMTHYTSSESPLOPAAPRLNPSGFTNTMSSIPQIATMAENYS
SLGSMTFCLQGRDAPYFMHFDLSLSPVAAHHRNTCPNFAAHHQPPHQDQVYTNSSMPSNTGVSAGVSVQVQISTQVSDLSTGNYWPLQL
>Dgr1
MMLTTEHHMHHGHPHSSVGGMGSALFGCTAGHSGINQLGQVYVNGRPLPDRTRQKVELAHSGARPCDSIRLQVNSGVSKLGRYETGSIKPRAIGSKPRVATTPVQKIA
DYKRECFSAWEIRDRLLEQVCSNDSPVSSINRVLNLSAQEQDAQDQNESVEYKLRMFGQSGWAWYPSNTTTHALPPTTAVPTNLGQNRDDQK
RDLYPGDVSHFNSHETSQDSDGNSHNSGDEDSQMLRKLKLNRTSPNEQDLEKEFERTHPDVFARELAKGLPEARLQVWFSNRRAKWRREKMRTRBSA
DVTGDSGRTSTANPSSGTTASSVATSNSTGIVNSAINVAERTSSALVNSLPEASNGPTVLGGEAMTHYTSSESPLOPAAPRLNPSGFTNTMSSIPQIATMAENYS
LGSMTFCLQGRDAPYFMHFDLSLSPVAAHHRNTCPNFAAHHQPPHQDQVYTNSSMPSNTGVSAGVSVQVQISTQVSDLSTGNYWPLQL
>Dpe
MMLTTEHHMHHGHPHSSVGGMGSALFGCTAGHSGINQLGQVYVNGRPLPDRTRQKVELAHSGARPCDSIRLQVNSGVSKLGRYETGSIKPRAIGSKPRVATTPVQKIA
DYKRECFSAWEIRDRLLEQVCSNDSPVSSINRVLNLSAQEQDAQDQNESVEYKLRMFGQSGWAWYPSNTTTHALPPTTAVPTNLGQNRDDQK
RELQSVESVHTNSDSDGNSHNSGDEDSQMLRKLKLNRTSPNEQDLEKEFERTHPDVFARELAKGLPEARLQVWFSNRRAKWRREKMRTRBSA
DVTGDSGRTSTANPSSGTTASSVATSNSTGIVNSAINVAERTSSALVNSLPEASNGPTVLGGEAMTHYTSSESPLOPAAPRLNPSGFTNTMSSIPQIATMAENYS
LGSMTFCLQGRDAPYFMHFDLSLSPVAAHHRNTCPNFAAHHQPPHQDQVYTNSSMPSNTGVSAGVSVQVQISTQVSDLSTGNYWPLQL
>Dpse
MMLTTEHHMHHGHPHSSVGGMGSALFGCTAGHSGINQLGQVYVNGRPLPDRTRQKVELAHSGARPCDSIRLQVNSGVSKLGRYETGSIKPRAIGSKPRVATTPVQKIA
ADYKRECFSAWEIRDRLLEQVCSNDSPVSSINRVLNLSAQEQDAQDQNESVEYKLRMFGQSGWAWYPSNTTTHALPPTTAVPTNLGQNRDDQK
RDLYPGDVSHFNSHETSQDSDGNSHNSGDEDSQMLRKLKLNRTSPNEQDLEKEFERTHPDVFARELAKGLPEARLQVWFSNRRAKWRREKMRTRBSA
DVTGDSGRTSTANPSSGTTASSVATSNSTGIVNSAINVAERTSSALVNSLPEASNGPTVLGGEAMTHYTSSESPLOPAAPRLNPSGFTNTMSSIPQIATMAENYS
LGSMTFCLQGRDAPYFMHFDLSLSPVAAHHRNTCPNFAAHHQPPHQDQVYTNSSMPSNTGVSAGVSVQVQISTQVSDLSTGNYWPLQL
    
```

12

Clustal Omega results — alignments

Results for job clustalo-I20191224-044002-0813-44387822-p1m

Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Results Viewers | Submission Details

Download Alignment File | Show Colors

CLUSTAL Omega multiple sequence alignment

```

DmE1  MLVTERINHGHPSE---VQGLTFCSTAGHGIINGLQVYVGRFLPSTQKQVE  54
DmEe  MLVTERINHGHPSE---VQGLTFCSTAGHGIINGLQVYVGRFLPSTQKQVE  54
DmEe  MLVTERINHGHPSE---VQGLTFCSTAGHGIINGLQVYVGRFLPSTQKQVE  57
DmE1  MLVTERINHGHPSE---VQGLTFCSTAGHGIINGLQVYVGRFLPSTQKQVE  60
DmE1  MLVTERINHGHPSE---VQGLTFCSTAGHGIINGLQVYVGRFLPSTQKQVE  57
DmE1  MLVTERINHGHPSE---VQGLTFCSTAGHGIINGLQVYVGRFLPSTQKQVE  57
DmE1  LARGAPPCDIERLQVHSCVSLGRVYETVSIFKRALDGERFFATFPVQGLADTK  114
DmEe  LARGAPPCDIERLQVHSCVSLGRVYETVSIFKRALDGERFFATFPVQGLADTK  114
DmEe  LARGAPPCDIERLQVHSCVSLGRVYETVSIFKRALDGERFFATFPVQGLADTK  117
DmE1  LARGAPPCDIERLQVHSCVSLGRVYETVSIFKRALDGERFFATFPVQGLADTK  120
DmE1  LARGAPPCDIERLQVHSCVSLGRVYETVSIFKRALDGERFFATFPVQGLADTK  117
DmE1  RECFSIFAMEIDKLLSQQVCSNDFIYVSSISBVLNKLASQFQQAQOQSEYVEKLM  176
DmEe  RECFSIFAMEIDKLLSQQVCSNDFIYVSSISBVLNKLASQFQQAQOQSEYVEKLM  176
DmEe  RECFSIFAMEIDKLLSQQVCSNDFIYVSSISBVLNKLASQFQQAQOQSEYVEKLM  177
DmE1  RECFSIFAMEIDKLLSQQVCSNDFIYVSSISBVLNKLASQFQQAQOQSEYVEKLM  180
DmE1  RECFSIFAMEIDKLLSQQVCSNDFIYVSSISBVLNKLASQFQQAQOQSEYVEKLM  177
DmE1  PROGDGMAYFHTTALALPFPFPAAYVTEPAMLQAGARQDQREKQVVEYVETNS  236
DmEe  PROGDGMAYFHTTALALPFPFPAAYVTEPAMLQAGARQDQREKQVVEYVETNS  236
DmEe  PROGDGMAYFHTTALALPFPFPAAYVTEPAMLQAGARQDQREKQVVEYVETNS  234
DmE1  PROGDGMAYFHTTALALPFPFPAAYVTEPAMLQAGARQDQREKQVVEYVETNS  237
DmE1  PROGDGMAYFHTTALALPFPFPAAYVTEPAMLQAGARQDQREKQVVEYVETNS  236
    
```

13

Clustal Omega results — phylogenetic tree

Results for job clustalo-I20191224-044002-0813-44387822-p1m

Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Results Viewers | Submission Details

Download Phylogenetic Tree Data

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: Cladogram Real

Tree Data

```

DmE1:0.01225,
DmEe:0.01721,
DmE1:0.05451,
DmE1:0.05552,
DmE1:0.05887
    
```

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

Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Results Viewers | Submission Details

Input Sequences
clustalo-I20191224-044002-0813-44387822-p1m.input

Tool Output
clustalo-I20191224-044002-0813-44387822-p1m.output

Alignment in CLUSTAL format with base/residue numbering
clustalo-I20191224-044002-0813-44387822-p1m.clustal_num

Guide Tree
clustalo-I20191224-044002-0813-44387822-p1m.dnd

Phylogenetic Tree
clustalo-I20191224-044002-0813-44387822-p1m.ph

Percent Identity Matrix
clustalo-I20191224-044002-0813-44387822-p1m.pim

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
 - Alignment in CLUSTAL format with base/residue numbering
 - [clustalo-I20191224-044002-0813-44387822-p1m.clustal_num](#)
 - 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

```

Dmel1-543 1 MLLTTERIMHGRHPS---VSDITLCCCTACMSINQLCCGVVNCVFPDSTRKIVELAHSGARCD 67
Dmex1-543 1 MLLTTERIMHGRHPS---VSDITLCCCTACMSINQLCCGVVNCVFPDSTRKIVELAHSGARCD 67
Dpse1-546 1 MLLTTERIMHGRHPSVQVMDIAELCCCTACMSINQLCCGVVNCVFPDSTRKIVELAHSGARCD 66
Dgri1-545 1 MLLTTERIMHGRHPSVQVMDIAELCCCTACMSINQLCCGVVNCVFPDSTRKIVELAHSGARCD 71
Dmel2-551 1 MLLTTERIMHGRHPS---CQDIAELCCCTACMSINQLCCGVVNCVFPDSTRKIVELAHSGARCD 66

Dmel1-543 68 RLVQVNCVSKLGRVYETESKFAKICGSPFVATTPVQKADYKRECFSIFAWIRDRLLSDQVCH 138
Dmex1-543 68 RLVQVNCVSKLGRVYETESKFAKICGSPFVATTPVQKADYKRECFSIFAWIRDRLLSDQVCH 138
Dpse1-546 82 RLVQVNCVSKLGRVYETESKFAKICGSPFVATTPVQKADYKRECFSIFAWIRDRLLSDQVCH 135
Dgri1-545 72 RLVQVNCVSKLGRVYETESKFAKICGSPFVATTPVQKADYKRECFSIFAWIRDRLLSDQVCH 142
Dmel2-551 89 RLVQVNCVSKLGRVYETESKFAKICGSPFVATTPVQKADYKRECFSIFAWIRDRLLSDQVCH 138

Dmel1-543 139 SDNIFSVSSINRVLNLSQKIQDQDQDQNSVVEELMFGNCGIADWAPYVNTDPAHILPFAASVETSA 209
Dmex1-543 139 SDNIFSVSSINRVLNLSQKIQDQDQDQNSVVEELMFGNCGIADWAPYVNTDPAHILPFAASVETSA 209
Dpse1-546 140 SDNIFSVSSINRVLNLSQKIQDQDQDQNSVVEELMFGNCGIADWAPYVNTDPAHILPFAASVETSA 209
Dgri1-545 143 SDNIFSVSSINRVLNLSQKIQDQDQDQNSVVEELMFGNCGIADWAPYVNTDPAHILPFAASVETSA 211
Dmel2-551 140 SDNIFSVSSINRVLNLSQKIQDQDQDQNSVVEELMFGNCGIADWAPYVNTDPAHILPFAASVETSA 209

Dmel1-543 210 NLSQDITRILVNRDQVFGDLSFPMNITSDQNSQDSSDQMLKLRKLNQNETITRDIDSD 280
Dmex1-543 210 NLSQDITRILVNRDQVFGDLSFPMNITSDQNSQDSSDQMLKLRKLNQNETITRDIDSD 280
Dpse1-546 212 NLSQDITRILVNRDQVFGDLSFPMNITSDQNSQDSSDQMLKLRKLNQNETITRDIDSD 278
Dgri1-545 212 NLSQDITRILVNRDQVFGDLSFPMNITSDQNSQDSSDQMLKLRKLNQNETITRDIDSD 281
Dmel2-551 211 NLSQDITRILVNRDQVFGDLSFPMNITSDQNSQDSSDQMLKLRKLNQNETITRDIDSD 280

Dmel1-543 281 KEFEFTHYDFVAFERLDKICLIFARQWPNRANRREEMQTRQASITLDGSGRSTANNISG 351
Dmex1-543 281 KEFEFTHYDFVAFERLDKICLIFARQWPNRANRREEMQTRQASITLDGSGRSTANNISG 351
Dpse1-546 279 KEFEFTHYDFVAFERLDKICLIFARQWPNRANRREEMQTRQASITLDGSGRSTANNISG 349
Dgri1-545 282 KEFEFTHYDFVAFERLDKICLIFARQWPNRANRREEMQTRQASITLDGSGRSTANNISG 352
Dmel2-551 281 KEFEFTHYDFVAFERLDKICLIFARQWPNRANRREEMQTRQASITLDGSGRSTANNISG 351

Dmel1-543 352 T--ASSVATSNSTQIVNINVAERT--SSALVSIISREASGCVLCCGNTTISLETFQAR 417
Dmex1-543 352 T--ASSVATSNSTQIVNINVAERT--SSALVSIISREASGCVLCCGNTTISLETFQAR 417
Dpse1-546 350 --ASSVATFSRSTQIVNINVAERT--SSALVSIISREASGCVLCCGNTTISLETFQAR 416
Dgri1-545 353 --SVPTNATANDSLEICT--CCSCASTVMACNPNLSTSGRTILGGDQVNSNDQVDFQAV 417
Dmel2-551 352 VTSVETATTCQRIELISAVNGVETESAVYCGMTEADYRQRILOGDQVNSNDQVDFQAV 417

Dmel1-543 418 IREKLEKLEKINRYSIPQFATMAENK---LGMPTKISQDQVFFMHPDLSLCCVFAHNR 485
Dmex1-543 418 IREKLEKLEKINRYSIPQFATMAENK---LGMPTKISQDQVFFMHPDLSLCCVFAHNR 485
Dpse1-546 418 IREKLEKLEKINRYSIPQFATMAENK---LGMPTKISQDQVFFMHPDLSLCCVFAHNR 485
Dgri1-545 418 IREKLEKLEKINRYSIPQFATMAENK---LGMPTKISQDQVFFMHPDLSLCCVFAHNR 487
Dmel2-551 421 IREKLEKLEKINRYSIPQFATMAENK---LGMPTKISQDQVFFMHPDLSLCCVFAHNR 485

Dmel1-543 486 ENRANRNDQFPDHC---RHSVAMPSS--SNVCVLSQVSVFQISTQNSDQTESNWR 543
Dmex1-543 486 ENRANRNDQFPDHC---RHSVAMPSS--SNVCVLSQVSVFQISTQNSDQTESNWR 543
Dpse1-546 480 ENRANRNDQFPDHC---RHSVAMPSS--SNVCVLSQVSVFQISTQNSDQTESNWR 546
Dgri1-545 488 ENRANRNDQFPDHC---RQCAVAVT--MDCVLSQVSVFQISTQNSDQTESNWR 545
Dmel2-551 491 ENRANRNDQFPDHC---RHSVAMPSS--SNVCVLSQVSVFQISTQNSDQTESNWR 543
    
```

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