

Multiple Sequence Alignment: Background

The purpose of this tutorial is to describe several commonly encountered Multiple Sequence Alignment (MSA) format types, namely the (1) **Clustal**, (2) **FASTA**, and (3) **Phylip** MSA formats.

MSAs are alignments of three or more DNA, RNA or protein sequences.

Usually these sequences come from different organisms but sometimes they can be duplicated gene families from the same organism. MSAs have many uses in Bioinformatics. One major use of MSAs is to determine which parts of a sequence (nucleotides or amino acids) have evolved the most or the least over time.

Sequence positions (nucleotide or amino acid) positions that have mutated frequently are called “variable” positions. Ones that mutate rarely or at all, are called “conservative” positions. Positions or regions of the sequence alignment that are very conservative, especially compared to the rest of the sequence, may indicate functionally important regions of the sequence.

Multiple Sequence Alignment: Clustal Format

The clustal MSA alignment format is generated by the Clustal Multiple sequence aligner. The following steps show an alignment of three DNA sequences (FASTA file in the box) using Clustal Omega program.

```
>NucSeq1
ATGAACGACGAAACACAATTTACAAATAAGGCCAACGAAATTATCCGTTTGGCCCAGAAATTGGCTCAGG
ATCACAGACATGCTCAGTTACAACCAATTCACCTACTTGTCTGCATTTGTTGAGCCAAACGAGGATGGTTC
>NucSeq2
ATGGCTGATTATCCTTTTACTGACAAAGCCGCAAAGACATTGTCTGATGCGTACTCAATTGCACAATCTT
ATGGTCATTCACAATTAACCCCTATTACATTGCTGCTGCTCTTTTGTCCGACAGTGACAGTAACGGTAC
>NucSeq3
ATGAACGACGAAACGAAGTTTACGAACAAAGCTCTCGATATCATCACCATTGCACAGAAACTAGCACAGG
ACCACCAGCATTGACGCTGGTGCCTCTACACGTGCTTGCAGCGTTTCGTAGAGACACCTGCTGATGGTAG
```

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HI

Step 1: Data pasted into input window.

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

```
>NucSeq1
ATGAACGACGAAACACAATTTACAAATAAGGCCAACGAAATTATCCGTTTGGCCCAGAAATTGGCTCAGG
ATCACAGACATGCTCAGTTACAACCAATTCACCTACTTGTCTGCATTTGTTGAGCCAAACGAGGATGGTTC
>NucSeq2
ATGGCTGATTATCCTTTTACTGACAAAGCCGCAAAGACATTGTCTGATGCGTACTCAATTGCACAATCTT
ATGGTCATTCACAATTAACCCCTATTACATTGCTGCTGCTCTTTTGTCCGACAGTGACAGTAACGGTAC
>NucSeq3
```

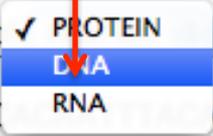
Or, upload a file: No file chosen

Step 2. Because these are DNA sequences, select “DNA” in the pop-down menu.

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and H

STEP 1 - Enter your input sequences

Enter or paste a set of  sequences in any supported format:

```
>NucSeq1
ATGAACGACGAAACAAATAAGGCCAACGAAATTATCCGTTTGGCCCAGAAATTGGCTCAGG
ATCACAGACATGCTCAGTTACAACCAATTCACCTTACTTGCTGCATTTGTTGAGCCAAACGAGGATGGTTC
>NucSeq2
ATGGCTGATTATCCTTTTACTGACAAAGCCGCAAAGACATTGTCTGATGCGTACTCAATTGCACAATCTT
ATGGTCATTCACAATTAACCCCTATTCACATTGCTGCTGCTCTTTTGTCCGACAGTGACAGTAACGGTAC
>NucSeq3
```

Step 3. Scroll Down and click the Submit button.

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)*



Step 4. Interpreting the output: After waiting a bit, the output will appear in the browse.

Clustal Omega

[Input form](#) | [Web services](#) | [Help & Documentation](#)

[Tools](#) > [Multiple Sequence Alignment](#) > Clustal Omega

Links to data, output, phylogeny.

Results for job clustalo-l20140211-234917-0285-23328113-es

[Alignments](#) | [Result Summary](#) | [Phylogenetic Tree](#) | [Submission Details](#)

Download to a text file.

[Download Alignment File](#) | [Show Colors](#) | [Send to ClustalW2_Phylogeny](#)

The sequence alignment.

CLUSTAL O(1.2.0) multiple sequence alignment

```
NucSeq3      ATGAACGACGAAACGAAGTTTACGAACAAAGCTCT--CGATATCATCACCATTGCACAGA
NucSeq1      ATGAACGACGAAACACAATTTACAAATAAGGCCAA--CGAAATTATCCGTTTGGCCAGA
NucSeq2      ---ATGGCTGATTATCCTTTTACTGACAAAGCCGAAAGACATTGCTGATGCGT--ACT
              *: *. **::. .. ***** .* **.*. .** ** .* * * :
NucSeq3      AACTAGCACAGGACCACCAGCATTCGACGCTGGTGCCCTTACACGTGCTTGCAGCGTTCG
NucSeq1      AATTGGCTCAGGATCACAGACATGCTCAGTTACAACCAATTCACTTACTTGCTGCATTTG
NucSeq2      CAATTGCACAATCTTATGGTCATTCACAATTAACCCCTATTACATTGCTGCTGCTCTTT
              .* * **::** . * . *** * ... * . **::.*:*** * ***:*** *
NucSeq3      TAGAGACACCTGCTGATGGTAG-----
NucSeq1      TTGAGCCAAACGAGGAT-----GGTTC
NucSeq2      T--GTCCGACAGTGACAGTAACGGTAC
              * . .*... * ...
```

The Clustal format is “interleaved”: these **alignments show 60 alignment positions for all sequences**, then go to the next 60 until there is no more alignment.

More information about the Clustal alignment format.

The Clustal formats begin with a line that looks like this. This is actually how some programs recognize it to be a Clustal MSA.

```
CLUSTAL O(1.2.0) multiple sequence alignment
```

```
NucSeq3      ATGAACGACGAAACGAAGTTTACGAACAAAGCTCT--CGATATCATCACCATTGCACAGA
NucSeq1      ATGAACGACGAAACACAATTTACAAATAAGGCCAA--CGAAATTATCCGTTTGGCCCAGA
NucSeq2      ---ATGGCTGATTATCCTTTTACTGACAAAGCCGCAAAGACATTGTCTGATGCGT--ACT
              *: *. **:;. .. ***** .* **.**      .** ** .**      : * * :
```

Gaps are “-” character.
The alignment put three gap characters in a row here.

A asterisk “*” indicates all the sequences have the same nucleotide.
Fully conserved.

0 dots = no conservation.
1 dot “.” = some conservation.
2 dots “:” = all pyrimidines or all purines at position.

Clustal Format: Protein Sequences

Step. 4 Wait for alignment

Step 1. Set to Protein.

Step 2. Paste in Data (or choose file).

Step 3. Submit.

STEP 1 - Enter your input sequences

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

```
>LCseedSf1
MKKLTVAISAVAASVLMAMSAQAAEIYNKDSNKL DLYGKVN AKHYFSSNDADDG
LGFKGETQINDQLTGFGQWEYEFKGNRAESQGSSKDKTRLAFAGLKF GDYGSIC
GVAYDIGAWTDVLP EFGGDTWTQTDFVMTGRTTGVATYRNNDFGLVDGLNFA/
NDRTDVT EANGDGFSTTYEYEGFVGATYAKSDRTNDQVIYGNNSLNASGQI
AAGLKYDANNIYLATTYSETQNM TVFGNNHIANKAQNFEVVAQYQDFGLRPSV/
GKDLGAWGDQDLIEYIDVGATYYFNKNMSTFVDYKINLIDKSDFTKASGVATDDIV
```

Or, upload a file: No file chosen

STEP 2 - Set your parameters

OUTPUT FORMAT

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

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)*

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Download Alignment File | Show Colors | Send to ClustalW2_Phylogeny

CLUSTAL O(1.2.0) multiple sequence alignment

```
LCseedSf1          MKKLTVAISAVAASVLMAMSAQAAEIYNKDSNKL DLYGKVN AKHYFSSNDADDGDTTYVR
PhoEseedEco2      MKMKKSTLALVVMGIVASVSVQAAEIYNKDG NKLDVYGKVKAMHYMSDNDSKDG DQSYIR
PhoEseedEco1      MKMKKSTLALVVMGIVASASVQAAEIYNKDG NKLDVYGKVKAMHYMSDNDSKDG DQSYIR
PhoEseedEco4      --MKKSTLALVVMGIVASASVQAAEIYNKDG NKLDVYGKVKAMHYMSDNDSKDG DQSYIR
PhoEseedSen1      --MNKSTLAI-VVSI IASASVHAAEVY NKNGNKLDVYGKVKAMHYMSDYDSKDG DQSYVR
PhoEseedSen2      --MNKSTLAI-VVSI IASASVHAAEVY NKNGNKLDVYGKVKAMHYMSDYDSKDG DQSYVR
                   .  ::  .  . : : *.:***:***:****:*****:* **:* . *:..*** :*:*
```

```
LCseedSf1          LGFKGETQINDQLTGFGQWEYEFKGNRAESQGSSKDKTRLAFAGLKF GDYGSIDYGRNYG
PhoEseedEco2      FGFKGETQINDQLTG YRWEAEFAGNKAESDT-AQQKTRLAFAGLKYKDLGSFDYGRNLG
PhoEseedEco1      FGFKGETQINDQLTG YRWEAEFAGNKAESDT-AQQKTRLAFAGLKYKDLGSFDYGRNLG
PhoEseedEco4      FGFKGETQINDQLTG YRWEAEFAGNKAESDT-AQQKTRLAFAGLKYKDLGSFDYGRNLG
PhoEseedSen1      FGFKGETQINDQLTG YRWEAEFAGNKAESDS-SQQKNRLAFAGLKLKDIGSFDYGRNLG
PhoEseedSen2      FGFKGETQINDQLTG YRWEAEFAGNKAESDS-SQQKTRLAFAGLKLKDIGSFDYGRNLG
                   :*****:*** ** **:*: :*:***** * **:****** *
```

As with DNA:

- “-” still means a gap.
- “*” means conserved.
- “.” and “:” still mean slightly conserved.

However, the . and : indicate different properties of amino acids conserved at a position.

Multiple Sequence Alignment: FASTA Format

Below is a MSA in FASTA format. It looks just like a regular FASTA file, but the sequences are all the same length and there are gap characters to show instances of deletions and insertions. FASTA MSAs are not interleaved.

```
>NucSeq2
---ATGGCTGATTATCCTTTTACTGACAAAGCCGCAAAGACATTGTCTGATGCGTACTCA
ATTGCACAATCTTATGGTCATTCACAATTAACCCCTATTCACATTGCTGCTGCTCTTTG
TCCGACAGTGACAGTAACGGTAC
>NucSeq1
ATGAACGACGAAACACAATTTACAAATAAGGCCAACGAAATTATCCGTTTGGCCCAGAAA
TTGGCTCAGGATCACAGACATGCTCAGTTACAACCAATTCACTTACTTGCTGCATTTGTT
GAGCAAACGAGGATG---GTTC
>NucSeq3
ATGAACGACGAAACGAAGTTTACGAACAAAGCTCTCGATATCATCACCATTGCACAGAAA
CTAGCACAGGACCACCAGCATTTCGACGCTGGTGCCTCTACACGTGCTTGCAGCGTTTCGTA
GAGACACCTGCTGATG---GTAG
```



Multiple Sequence Alignment: Phylip Format

Below is a MSA in Phylip format. This format is very commonly used for phylogenetic analysis. It was designed for the Phylip phylogenetic software tools, which are still widely used to create phylogenetic trees.

The format is interleaved like clustal. At the top of the file, the first line includes information on the number of sequences in the file, and the number of nucleotide (or amino acid positions) in the alignment.

```
3 143
NucSeq2 ---ATGGCTGATTATCCTTTTACTGACAAAGCCGCAAAGACATTGTCCTGA
NucSeq1 ATGAACGACGAAACACAATTTACAAATAAGGCCAACGAAATTATCCGTTT
NucSeq3 ATGAACGACGAAACGAAGTTTACGAACAAAGCTCTCGATATCATCACCAT

TGCGTACTCAATTGCACAATCTTATGGTCATTCACAATTAACCCCTATTC
GGCCCAGAAATTGGCTCAGGATCACAGACATGCTCAGTTACAACCAATTC
TGCACAGAAACTAGCACAGGACCACCAGCATTCGACGCTGGTGCCTCTAC

ACATTGCTGCTGCTCTTTTGTCCGACAGTGACAGTAACGGTAC
ACTTACTTGCTGCATTTGTTGAGCCAAACGAGGATG---GTTC
ACGTGCTTGCAGCGTTCGTAGAGACACCTGCTGATG---GTAG
```