

# **Bioinformatics**

**CS300**

**Chap 3**

**Sequence Alignment  
with ClustalW**

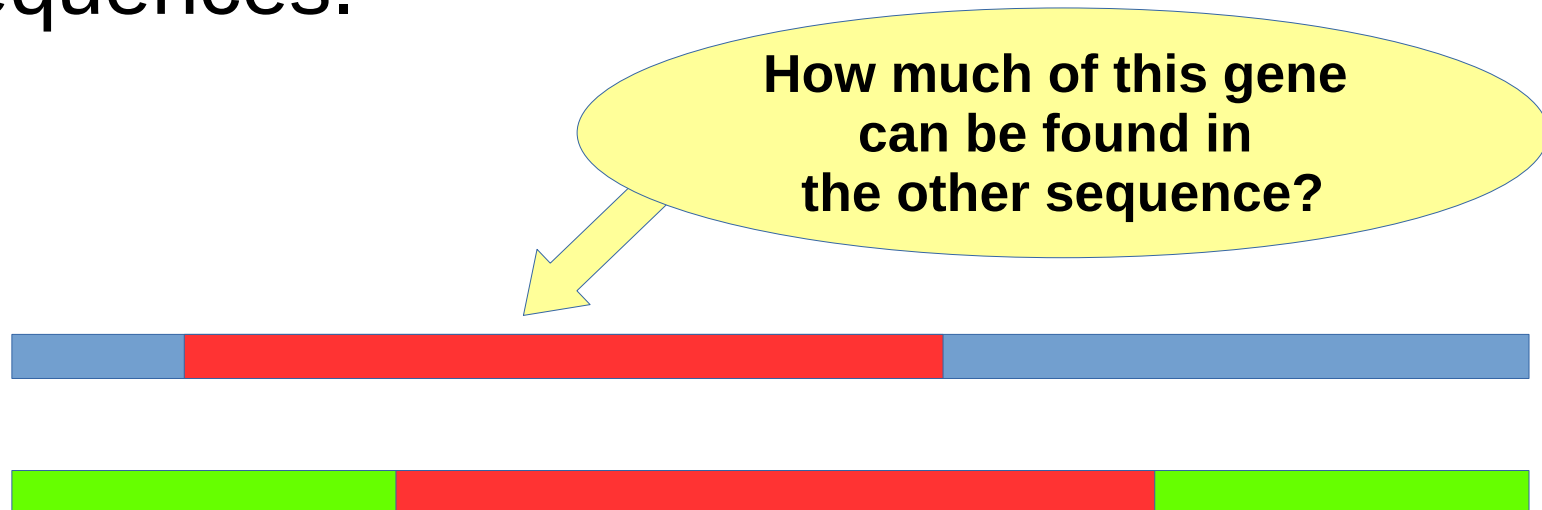
**Fall 2021**

**Oliver BONHAM-CARTER**



# What is Sequence Alignment?

- Sequence alignment is a way of arranging the sequence of genetic material (DNA, RNA or protein) to identify regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences.



# Types of Alignment

<i>A</i>	<i>T</i>	<i>C</i>	<i>G</i>	<i>A</i>	<i>A</i>	<i>C</i>	<i>T</i>	<i>G</i>	<i>G</i>	<i>C</i>	<i>C</i>	–	–
.	.			.									
<i>T</i>	<i>A</i>	<i>C</i>	<i>G</i>	<i>C</i>	<i>A</i>	<i>C</i>	<i>T</i>	–	–	<i>C</i>	<i>C</i>	<i>A</i>	<i>A</i>

(a) Global alignment example

–	<i>A</i>	<i>T</i>	<i>C</i>	<i>G</i>	<i>A</i>	<i>A</i>	<i>C</i>	<i>T</i>	<i>G</i>	<i>G</i>	<i>C</i>	<i>C</i>	–	–
					.									
<i>T</i>	<i>A</i>	–	<i>C</i>	<i>G</i>	<i>C</i>	<i>A</i>	<i>C</i>	<i>T</i>	–	–	<i>C</i>	<i>C</i>	<i>A</i>	<i>A</i>

(b) Semi-global alignment example

<i>A</i>	<i>T</i>	<i>C</i>	<i>G</i>	<i>A</i>	<i>A</i>	<i>C</i>	<i>T</i>	<i>G</i>	<i>G</i>	<i>C</i>	<i>C</i>		
				.									
<i>T</i>	<i>A</i>	<i>C</i>	<i>G</i>	<i>C</i>	<i>A</i>	<i>C</i>	<i>T</i>	–	–	<i>C</i>	<i>C</i>	<i>A</i>	<i>A</i>

(c) Local alignment example



# Needleman-Wunsch Algorithm Background

- **Alignment:** Used to determine which parts of a sequence are in common with another sequence; used to measurement similarity between sequences.
- Developed by Saul B. Needleman and Christian D. Wunsch in 1970.
- Dynamic programming to find optimal solution for matching the characters of the two sequences.

# Ex: Pairwise Alignment

## Alignment of a gene from two closely related viruses

Hemagglutinin gene from virus A: ATGAACGCAATACTCGTAGTT...

||||| ||||| |||||

Hemagglutinin gene from virus B: ATGAAGGCAATACTAGTAGTT...

Few Mismatches



## Alignment of a gene from two distantly related viruses

Hemagglutinin gene from virus A: ATGAACGCAATACTCGTAGTT...

||| ||| ||| |||| | |

Hemagglutinin gene from virus C: ATGCACGAAATGCTCGGACCT...

Lots of Mismatches





# What is Global Sequence Alignment?

- We search for matches, matches and gaps between two sequences to determine their **relatedness**.
- (\*) indicate matches or similar nucleotides (bases) along sequence
- Here, the sequences may have a common ancestor

ACGTACT  
ACTACGT  
\*\* \*

ACGTAC-T  
AC-TACGT  
\*\* \*\*\* \*

ACGTACT- - - -  
- - - -ACTACGT  
\*\*\*\*



# Ex: Comparing DNA

- We compare DNA samples from several different organisms.

		850		860		870		880								
<i>Gallus_gallus</i> /1-2533	CT	CAG	AAAA	CT	GCTTT	AAAT	GAA	GCC	CAT	CCA	GCA	GCTT	GG	AG	GG	GC
<i>Mus_musculus</i> /1-2491	CT	GGG	AAAA	CT	GTTTT	AAAT	CAA	GCT	ATT	TTT	ACA	GCTT	GG	AG	GG	AC
<i>Rattus_norvegicus</i> /1-2601	CT	GGG	AAAA	CT	GTTTT	AAAT	CAA	GCT	ATT	ATT	ACA	GCTT	GG	AG	GG	AC
<i>Dasyus_novemcinctus</i> /1-2306	CT	GGG	AAAA	CT	GCTTT	AAAT	CAA	GCT	ATT	TTT	GCA	AACTT	GG	AG	GG	AC
<i>Loxodonta_africana</i> /1-2443	CT	GGG	AAAG	CT	GCTTT	GAAT	CAA	ACT	GTT	TTT	TCA	AACTT	GG	AG	GG	GC
<i>Oryctolagus_cuniculus</i> /1-2522	CT	GGG	AAAA	CT	GCTTT	AAAT	CAA	GCT	GTT	ATT	GCA	AACTT	GG	AG	GG	AC
<i>Equus_caballus</i> /1-2583	CT	GGG	AAAA	CT	GCTTT	AAAT	CAA	GCT	GTT	ATT	GCA	AACTT	GG	AG	GG	AC
<i>Gorilla_gorilla</i> /1-4513	CT	GGG	AAAA	CT	GCTTT	AAAT	CAA	GCT	ATT	ATT	GCA	AACTT	GG	AG	GG	AT
<i>homo_sapiens</i> /1-4639	CT	GGG	AAAA	CT	GCTTT	AAAT	CAA	GCT	ATT	ATT	GCA	AACTT	GG	AG	GG	AT
<i>Macaca_mulatta</i> /1-2393	CT	GGG	AAAA	CT	GCTTT	AAAT	CAA	GCT	ATT	ATT	GCA	AACTT	GG	AG	GG	AT
<i>Bos_taurus</i> /1-2527	CT	GGG	AAAA	CT	GCTTT	AAGT	CAT	GCC	CAT	ATT	GCA	AACTT	GG	AG	GG	AC
<i>Tursiops_truncatus</i> /1-2513	CT	GGG	AAAA	CT	GCTTT	AAGT	CAA	GCT	GTT	ATT	GCA	AACTT	GG	AG	GG	AC
<i>Canis_lupus_familiaris</i> /1-2513	CT	GGG	AAAA	CT	GCTTT	AAAT	CAA	GCT	ATT	TTT	GCA	AACTT	GG	AG	GG	AC
<i>Felis_catus</i> /1-2884	CT	GGG	AAAA	CT	GCTTT	AAAT	CAA	GCT	ATT	ATT	GCA	AACTT	GG	AG	GG	AC

Consensus

CTGGGAAAACTGCTTTAAATCAAGCTATATTGCAACTTGGAGGAC

# Ex: Comparing Protein

- We compare protein samples from several different organisms.

	*																																*	
Human	W	N	Q	S	T	A	R	W	L	R	R	L	V	F	Q	H	S	R	A	W	P	L	L	Q	T	F	A	F	S	A	W	W	H	G
Pig	W	N	H	S	T	A	Q	W	L	R	R	L	V	F	Q	Q	G	R	T	W	P	L	L	Q	T	F	V	F	S	A	W	W	H	G
Cow	W	N	Q	S	T	A	R	W	L	R	R	L	V	F	Q	Q	R	R	T	W	P	L	L	Q	T	F	L	F	S	A	W	W	H	G
Dog	W	N	Q	S	T	A	R	W	L	R	R	L	V	F	Q	Q	R	R	T	W	P	L	L	Q	T	F	L	F	S	A	W	W	H	G
Rat	W	N	R	S	T	A	Q	W	L	K	R	L	V	F	Q	R	S	R	R	W	P	V	L	Q	T	F	A	F	S	A	W	W	H	G
Mouse	W	N	R	S	T	A	L	W	L	R	R	L	V	F	R	K	S	R	R	W	P	L	L	Q	T	F	A	F	S	A	W	W	H	G
Chicken	W	N	R	S	T	S	L	W	L	R	R	L	V	F	Q	R	C	P	V	Q	P	L	L	A	T	F	A	F	S	A	W	W	H	G
Zebrafish	W	N	Q	T	T	V	D	W	L	R	K	I	V	F	N	R	T	S	R	S	P	L	F	M	T	F	G	F	S	A	L	W	H	G



		*	:	*	:	:	
Q5E940_BOVIN	-----MPREDRATWKS	NYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVL	76				
RLA0_HUMAN	-----MPREDRATWKS	NYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVL	76				
RLA0_MOUSE	-----MPREDRATWKS	NYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVL	76				
RLA0_RAT	-----MPREDRATWKS	NYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVL	76				
RLA0_CHICK	-----MPREDRATWKS	NYFMKIIQLLDDYPKCFVVGADNVGSKOMQQIRMSLRGK-AVVL	76				
RLA0_RANSY	-----MPREDRATWKS	NYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVL	76				
Q7ZUG3_BRARE	-----MPREDRATWKS	NYFLKIIQLLDDYPKCFIVGADNVGSKOMQTIRLSLRGK-AVVL	76				
RLA0_ICTPU	-----MPREDRATWKS	NYFLKIIQLLDDYPKCFIVGADNVGSKOMQTIRLSLRGK-AIVL	76				
RLA0_DROME	-----MVRENKA	AWKAQYFIKVVELFDEFPKCFIVGADNVGSKOMQNIRTSLRGI-AVVL	76				
RLA0_DICDI	-----MSGAG-SKR	KKLFIEATKLFTTYDKMIVA	75				
Q54LP0_DICDI	-----MSGAG-SKR	KNVFIEATKLFTTYDKMIVA	75				
RLA0_PLAF8	-----MAKLSKQ	QKKQMYIEKLSSLIQQYSKILIVHVDNVGSKOMASVRKSLRGK-ATIL	76				
RLA0_SULAC	-----MIGLAVT	TTTKKIAKWKVDEVAELTEKLKTHKTI	79				
RLA0_SULTO	-----MRIMAVIT	QERKIAKWKIEEVKELEOKLREYHTII	80				
RLA0_SULSO	-----MKRLALAL	QQRKVASWGLEEVKELTELKINSNTILIGN	80				
RLA0_AERPE	MSVVS	SLVGQMYKREKPIPEWKTMLMLELEELFSKHRYVFLDLTGPTIFVQVRVKKLWKK-YPMVAVKKRIILFRAMKAAGLE---LDDN	86				
RLA0_PYRAE	-MMLAIG	KRRYVRTQYPARKVIIVSEATELQKYPYVFLDHLGLSSRIILHEYRYRLARY-GVIKIKPTLLFKIAFTKVYGG---IPAE	85				
RLA0_METAC	-----MAEERH	HTEHIPQWKKDEIENIKELIQSHKVF	78				
RLA0_METMA	-----MAEERH	TEHIPQWKKDEIENIKELIQSHKVF	78				
RLA0_ARCFU	-----MAAVRG	S---PPEYKVRAVEEIKRMISSKPVVAIVSF	75				
RLA0_METKA	MAYKAKG	QPPSGYEPKVAEWKRREVKELEKELMDEYENVGLVDLEGIPAPOLQEIRAQLRERTTIIRMSRNTLMRIAEEKLDER--PELE	88				
RLA0_METHH	-----MAHVAE	WKKKEVQELHDLIKGYEVVGIANLADIPAROLOQMROTLRDS-ALIRMSKKTLLISALEKAGREL--ENV D	74				
RLA0_METT L	-----MITAESE	HKIAPWKIEEYNKLKELLKNGQIVALVDMMEVPAROLOEIRDKIR-CTMTLKSRNTLIERAIKEVAEETGNPEFA	82				
RLA0_METVA	-----MIDAKSE	HKIAPWKIEEYNKLKELLKANVALIDMMEVPAPVOLQEIRDKIR-DQMTLKSRNTLIKRAEEVAEETGNPEFA	82				
RLA0_METJA	-----METKVKA	HVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPOLQEIRDKIR-DKVKLRSRNTLIIRALKEAAEELNPNKLA	81				
RLA0_PYRAB	-----MAHVAE	WKKKEVEELANLIKSYPIVIALVDVSSMPAYPLSQMRRLIRENGCLLRVSRTNLIELAIKKAAGELGKPELE	77				
RLA0_PYRHO	-----MAHVAE	WKKKEVEELAKLIKSYPIVIALVDVSSMPAYPLSQMRRLIRENGCLLRVSRTNLIELAIKKAAGELGKPELE	77				
RLA0_PYRFU	-----MAHVAE	WKKKEVEELANLIKSYPIVIALVDVSSMPAYPLSQMRRLIRENGCLLRVSRTNLIELAIKKAAGELGKPELE	77				
RLA0_PYRKO	-----MAHVAE	WKKKEVEELANLIKSYPIVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKKAAGELGQPELE	76				
RLA0_HALMA	-----MSAESER	KTETIPEWKQEEVDVAIVEMIESYESVG V NVGTGPSROLQDMRDDLHG T-AELRVSRNTLIERALDDVD---DGLE	79				
RLA0_HALVO	-----MSESEVR	QTEVIPQWKREEVDDELDFIESYESVG V NVGTGPSROLQSMRRELHGS-AAVRMSRNTLVNRALDEVN---DGFE	79				
RLA0_HALSA	-----MSAEEQ	RTTEEVPWK R Q E V A E L V D L L E T Y D S V G V N V T G I P S K O L Q D M R R L H G Q - A A L R M S R N T L L V R A L E E A G ----DGLD	79				
RLA0_THEAC	-----MKEVSQ	KKELVNEITORIKASRSVAIVDTAGIRTROIQDIRGKNRGK-INLKVIKKTLLFKALENLGD---EKLS	72				
RLA0_THEVO	-----MRKINP	KKKEIVSELADITKSKAVAIVDTKGVRTROMODIRAKNRDK-VKIKVVKKTLLFKALDSIND---EKLT	72				
RLA0_PICTO	-----MTEPAQ	WKIDFVNKLENI NSRKVAAIVSIGKLRNNEFOKIRNSIRDK-ARIKVSARILLRLAIENTGK---NNIV	72				
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90						



# Terms

- Alignment is divided up into sub problems
- Solutions are scored; the best solutions for char by char comparison are kept in the overall solution.
- **Match** – bases of each sequence at position ARE same
- **Mismatch** – bases of each sequence at position are NOT same
- **Gap** – bases are not the same, some insert or deletion may have occurred.

AGGCTATCACCTGACCTCCAGGCCGATGCCC  
TAGCTATCACGACCGCGGTCGATTGCCCCGAC

–AGGCTATCACCTGACCTCCAGGCCGA––TGCCC––  
TAG–CTATCAC––GACCGC––GGTCGATTGCCCCGAC





# Terms

- **Homology** – Two or more sequences have a common ancestor
- **Similarity** – Two sequences are similar in terms of base arrangements. Note: this similarity does not refer to any specific evolutionary process; the sequences show *similarity* as they are compared.
- **Conserved regions** – Regions in code which are very similar (or the same) across a wide group of organisms. Having code which has not changed, in light of mutations, in all the organisms suggests that the region have been maintained by natural selection (and may serve an important function.)
- **DNA Coding Regions** – Contains code that is more likely to make protein, often less likely to change genetically. Mutations in these areas may cause danger.
- **DNA NonCoding Regions** – Contains DNA that does not necessarily code for protein, but may serve in gene regulation, such as the binding or recognition sites of ribosomes and transcription factors. May still be be conserved within a genome.



ALLEGHENY  
COLLEGE

# Bring the Tool!



**Up Next!**

# Clustal Omega Multiple Sequence Alignment

## Clustal Omega

Input form

Web services

Help & Documentation

Also in this section ▼

Feedback

Share

Tools > Multiple Sequence Alignment > Clustal Omega

## 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 two sequences please instead use our [pairwise sequence alignment tools](#).

Link:

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

# Clustal Omega Multiple Sequence Alignment

Enter DNA  
sequences  
in this **FASTA**  
format.

## STEP 1 - Enter your input sequences

Enter or paste a set of

DNA

sequences in any supported format:

```
>seq_1
ATGCATGCATGCATGC
>seq_2
ATGCATGCATGC AAAA
>seq_3
ATGCATGC AAAAAAAAAA
```

Or, upload a file:  No file chosen

## STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfill the needs of most users.

(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)

Submit your job with  
the *Character  
Counts* option.

# Clustal Omega Results

## Clustal Omega

Input form | Web services | Help & Documentation | Also in this section ▾ | Feedback | Share

Tools > Multiple Sequence Alignment > Clustal Omega

Your job is currently running... please be patient

The result of your job will appear in this browser window.

Job ID: [clustalo-l20210321-210208-0237-5254380-p2m](#)

Please note the following

- You may press Shift+Refresh or Reload on your browser at any time to check if results are ready.
- You may bookmark this page to view your results later if you wish.
- Results are stored for 7 days.

Depending the  
number of seqs,  
you may have  
to wait...

The '\*'s denote  
same bases across  
sequences.

## Results for job clustalo-l20210321-210208-0237-5254380-p2m

Alignments

Result Summary

Guide Tree

Phylogenetic Tree

Results Viewers

Submission Details

Download Alignment File

seq_3	ATGCATGCAAAAAAA	16
seq_1	ATGCATGCATGCATGC	16
seq_2	ATGCATGCATGCAAAA	16
	***** *	

# Percent Identity Matrix

## How similar are the seqs?

Results for job clustalo-l20210321-210208-0237-5254380-p2m

Alignments Result Summary Guide Tree Phylogenetic Tree Results Viewers

Submission Details

Download Alignment File

seq_3	ATGCATGCAAAAAAA	16
seq_1	ATGCATGCATGCATGC	16
seq_2	ATGCATGCATGCAAAA	16
	***** *	

Find percent  
Identity results here

How similar are  
the sequences?.

```
#
#
# Percent Identity Matrix - created by Clustal2.1
#
#
```

	Seq_3	Seq_1	Seq 2
1: seq_3	100.00	62.50	81.25
2: seq_1	62.50	100.00	81.25
3: seq_2	81.25	81.25	100.00





# Percent Identity Matrix

## How similar are the seqs?

Results for job clustalo-l20210321-210208-0237-5254380-p2m

Alignments Result Summary Guide Tree Phylogenetic Tree Results Viewers

Submission Details

Download Alignment File

seq_3	ATGCATGCAAAAAAA	16
seq_1	ATGCATGCATGCATGC	16
seq_2	ATGCATGCATGCAAAA	16
	***** *	

Find similarities  
as trees here

View similarity  
in a tree from.

### Phylogram

Branch length: ☒ Cladogram ☐ Real



seq\_3 0.15625  
seq\_1 0.09375  
seq\_2 0.09375

### Guide Tree

### Phylogram

Branch length: ☐ Cladogram ☒ Real



seq\_3 0.15625  
seq\_1 0.09375  
seq\_2 0.09375

### Guide Tree

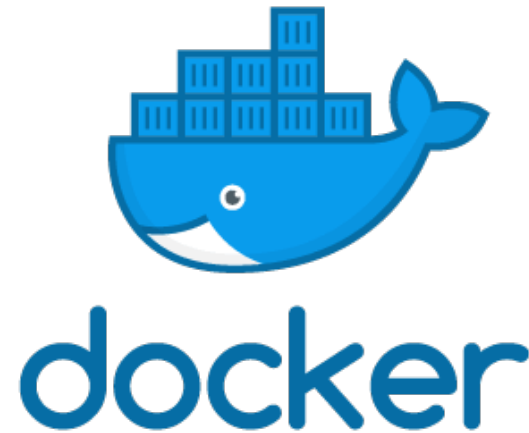


Please rebuild  
your container!!

# New Dockerfile

(for Docker  
Desktop)

Adds a tool called  
ClustalW for  
sequence  
comparison.



```
FROM ubuntu:20.04
```

```
RUN apt-get update && apt-get -y  
install git htop vim python3 python3-pip  
clustalw
```

```
RUN \  
pip3 install --upgrade pip \  
pip install bioPython
```

```
WORKDIR /root
```

```
CMD ["bash"]
```



# ClustalW On Your Own Machine

Enter sequences in a FASTA format:

```
>seq_1
ATGCATGCATGCATGC
>seq_2
ATGCATGCATGCAAAA
>seq_3
ATGCATGCAAAAAAAAAA
```

```
root@ea16b8965382:~# clustalw samples.fasta
```

```
CLUSTAL 2.1 Multiple Sequence Alignments
```

```
Sequence format is Pearson
```

```
Sequence 1: seq_1      16 bp
```

```
Sequence 2: seq_2      16 bp
```

```
Sequence 3: seq_3      16 bp
```

```
Start of Pairwise alignments
```

```
Aligning...
```

```
Sequences (1:2) Aligned. Score: 81
```

```
Sequences (1:3) Aligned. Score: 56
```

```
Sequences (2:3) Aligned. Score: 75
```

```
Guide tree file created: [samples.dnd]
```

```
There are 2 groups
```

```
Start of Multiple Alignment
```

```
Aligning...
```

```
Group 1: Sequences: 2      Score:256
```

```
Group 2: Sequences: 3      Score:214
```

```
Alignment Score 236
```

```
CLUSTAL-Alignment file created [samples.aln]
```