

Our Animal DNA:

Comparing genes across the Tree of Life

Practical 2: align sequences with Clustal

Step 1: Fetch human sequence

First, we need to export our sequences from Ensembl. We already have our human sequence from Practical 1, which looks like this:

Header **Amino acid 1-letter codes**

```
>ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:protein_coding
MSRQFTCKSGAAAKGGFSGCSAVLSGGSSSSFRAGSKGLSGGFGSRSLSLGGVRSNLVA
SGSGKSGGYGFRGRASGFAGSMFGSVALGPVCPTVCPPGGIHQVTNNESSLAPLNVELD
PEIQKVRQEREQIKALNNKFASFDKVRFLQEQNQVLETKWELLQQLDLNNCKNNLEPI
LEGYISNLRKQLETLSGDRVRLDSELNRNVRDVVEDYKKRYEEEINKRTAAENEFVLLKKD
VDAAYANKVELQAKVESMDQEI KFFRCLFEAEITQIQSHISDMSVILSMDNNRNLDSI
IDEVRTQYEEIALKSKAEAEALYQTKFQELQLAAGRHGDDLKNTKNEISELTRLIQIRIS
EIENVKQASNLETAIADAEQRGDNALKDARAKLDELEGALHQAEEELARMLREYQELMS
LKLALDMEIATYRKLESEECRMSGFPPSPVISIISSTSGGSVYGFRPSMVSGGYVANS
SNCISGVCSVRGGGRSGSANDYKDTLGKSSLSAPSCKTSTR
```

If you have lost this sequence, follow steps 1 and 2 of the first practical to find it again

Copy and paste this into a document. To make it easier to see what's what, change the header (the line beginning >) to:

>human_KRT71

Make sure the header line starts with >, otherwise Clustal won't know it's the header and will try to read it as amino acid sequence. The amino acid lines should not have > at the beginning. Make sure you don't remove the newline between the header and the first line of sequence. Do not add any extra newlines between the header and the first line of the sequence. Do not have multiple lines starting with >.

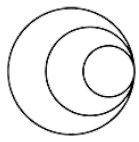
Step 2: Fetch other species sequences

Now you need to find the sequences from sea otter, black swan and Atlantic cod.

Go back to the tab with your BLAST results. Click on the identifier in the first column of the first row in the results table – this is the protein identifier.

Protein

Results table												
Show All entries		Show/hide columns (2 hidden)						Filter				
Subject name	Gene hit	Subject start	Subject end	Subject ori	Genomic Location	Orientation	Query start	Query end	Length	Score	E-val	%ID
ENSGMOP00000026268	ENSGMOP00000014429	101	441	Forward	1:23583241-23621465 (Sequence)	Forward	102	444	343 (Sequence)	424	9e-142	58.31 (Alignment)
ENSGMOP00000039893	ENSGMOP00000014429	239	578	Forward	1:23583241-23621462 (Sequence)	Forward	102	443	342 (Sequence)	424	3e-140	58.48 (Alignment)
ENSGMOP00000038280	ENSGMOP00000014429	239	578	Forward	1:23583241-23621462 (Sequence)	Forward	102	443	342 (Sequence)	424	3e-140	58.48 (Alignment)
ENSGMOP00000043242	ENSGMOP00000014429	239	578	Forward	1:23583241-23621462 (Sequence)	Forward	102	443	342 (Sequence)	424	3e-140	58.48 (Alignment)



If you no longer have the BLAST results open, you can search for the IDs you noted down instead.

This will take you to a protein page. This page shows us the regions of the protein that are known to be involved in particular functions. These have been computed from the sequence, based on those functions in other proteins.

Location: 1:23,582,829-23,622,805 Gene: ENSGMOG00000014429 Trans: ENSGMOT00000033979 BLAST results

Transcript: ENSGMOT00000033979.1

Location: Primary assembly 1: 23,582,829-23,622,805 forward strand.

About this transcript: This transcript has 9 exons and is annotated with 27 domains and features.

Gene: This transcript is a product of gene ENSGMOG00000014429.2 [Show transcript table](#)

Protein summary

Protein domains for ENSGMOP00000026268.1

Functional regions of the protein

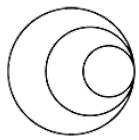
Export button

As before, we can export the sequence by clicking on the [Export data](#) button and selecting [Peptide sequence](#) only, ensuring that you select [None](#) under Genomic sequence. Refer back to step 2 of Practical 1 for more details and a screenshot of what to select.

Step 3: Make your sequence file

Copy and paste the sequence you get into the same document as the human protein sequence and change the name to give you the species again, making sure that you have [>](#) at the beginning of the header line and not at the beginning of the sequence lines.

Now do the same for the other two species. You should end up with a document containing four protein sequences, like this:



```
>human_KRT71
MSRQFTCKSGAAAKGGFSGCSAVLSGGSSSSFRAGSKGLSGGFGSRSLYSLGGVRS LNVA
SGSGKSGGYGFGGRGRASGFAGSMFGSVALGPVCPTVCPGGI HQVT VNESLLAPLNVELD
PEIQKVRAQEREQIKALNNKFASFIDKVRFL EQNQVLETKWELLQQLDLNNCKNNLEPI
LEGYISNLRKQLETLSGDRVRLDSELNRVDDVEDYKKRYEEEINKRTAAENEFVLLKKD
VDAAYANKVELQAKVESMDQEI KFFRCLFEAEITQIQSHISDMSVILSMDNNRNLDLDSI
IDEVRTQYEEIALKSKAEAEALYQTKFQELQLAAGR HGDDLKNTKNEISELTRLIQRI
EIENVKKQASNLETAIADAEQRGDNALKDARAKLDELEGALHQAKEELARMLREYQELMS
LKLALDMEIATYRKLLSEECRMSGEFPSVSIISSTSGGSVYGF RPSMVSGGYVANS
SNCISGVC SVRGGEGRSRGSANDYKDTL GKGSSLSAPSKKTSR

>swan
MSRQSTVRIQRGRSGFSAASAIVPNTCRTSFSSRSVTRV GSCNAGSGFSRVGGGFGSKSL
YNVGGCKRISVAGRGGGFYGPAGFGGAGSVYGC GGGFGMPANLVS VNQSLLKPLNLEID
PSIQIRKEEKEQIKTLNNKFASFIDKVRFL EQNQVLETKWSLLQE QGMKTVKNNLEPL
FETYINNLRMQLNSLLSDKGRLEGELVNTQYLVEDFKKKYED EINNRTVAENEFVTLKKD
VDASYMNKVELQAKVDALTEEINFLRALYEAELSQM QTQISDTSVLTMDNNRNLDLDSI
ISEVKAQYEDIANRSRAEAESWYQTKYEELQATAGRHGDDL RNTKQEISELNRHVQRLRS
EIDSVKKQCANLKAATADAEERGELALKDAKAKLAELE DALQAKADLARQLREYQELMN
VKLALDIEIATYRKLLGE EECRWGNTTLLVLVLEDPTQVL VMLGISEMHMLNGVSRGRG
KGCLVNSVTL

>otter
MSRQFTCKSGAATKGGFSGCSAVLSGGSSSSYRAGGKGIS GGFSGSRSLYSLGGIRNISFN
MTSGSGKSGGYGFGGRGRASGFAGSMFGSVALGPVCPSV CPGGIHQVT VNESLLAPLNVE
LDPEIQKVRAQEREQIKALNNKFASFIDKVRFL EQNQVLETKWELLQQLDLNNCKNNLE
PILEGYISNLRKQLETLSGDRVRLDSELRSVRDVEDYKKK YEEEINNRTAAENEFVLLK
KD VDAAYANKVELQAKVDSMDQEI KFFKCLYEA EIAQIQSHISDMSVILSMDNNRDLNLD
SIIDEVRAQYEEIALKSKAEAEALYQTKFQELQLAAGR HGDDLKNTKNEISELTRLIQRI
RSEIENVKKQASNLETAIADAEQRGDNALKDARAKLDELES ALHQAKEELARMLREYQEL
MSLKLALDMEIATYRKLLSEECRMSGEFPSVSIISSTSGSGGYGFRPSSVSGGYVA
SSSSCISGVC SVRGGDVRGRGSSSDYKDTLGRGSSQSTSSKKASR

>cod
MSRQFTCKSGAATKGGFSGCSAVLSGGSSSSYRAGGKGIS GGFSGSRSLYSLGGIRNISFN
MTSGSGKSGGYGFGGRGRASGFAGSMFGSVALGPVCPSV CPGGIHQVT VNESLLAPLNVE
LDPEIQKVRAQEREQIKALNNKFASFIDKVRFL EQNQVLETKWELLQQLDLNNCKNNLE
PILEGYISNLRKQLETLSGDRVRLDSELRSVRDVEDYKKK YEEEINNRTAAENEFVLLK
KD VDAAYANKVELQAKVDSMDQEI KFFKCLYEA EIAQIQSHISDMSVILSMDNNRDLNLD
SIIDEVRAQYEEIALKSKAEAEALYQTKFQELQLAAGR HGDDLKNTKNEISELTRLIQRI
RSEIENVKKQASNLETAIADAEQRGDNALKDARAKLDELES ALHQAKEELARMLREYQEL
MSLKLALDMEIATYRKLLSEECRMSGEFPSVSIISSTSGSGGYGFRPSSVSGGYVA
SSSSCISGVC SVRGGDVRGRGSSSDYKDTLGRGSSQSTSSKKASR
```

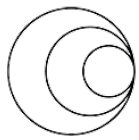
Check that:

1. All header lines start with >.
2. Each sequence has only one line of header.
3. The sequence lines do not start with >.
4. There are no extra lines between the header and sequence.

Step 4: Clustal input

You need to put these sequences into Clustal Ω. Go to

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



wellcome
connecting
science

EMBL-EBI



Darwin
TREE
of
LIFE

[Home](#) [EMBL-EBI](#) [Services](#) [Research](#) [Training](#) [Industry](#) [About us](#) [Search](#) [EMBL-EBI](#) [Hinxton](#)

Clustal Omega

[Input form](#) [Web services](#) [Help & Documentation](#) [Bioinformatics Tools FAQ](#) [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](#).

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:

Input

Or, upload a file: [Choose file](#) No file chosen [Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfill the needs of most users.

[More options...](#) (Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

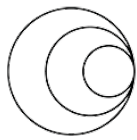
☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

Run

Copy all four sequences into the input box on Clustal. Then click on [Submit](#).

Step 5: Clustal results



Results for job clustalo-l20201015-095701-0397-7099231-p2m

Alignments Result Summary Guide Tree Phylogenetic Tree Results Viewers Submission Details

Download Alignment File Show Colors

View the alignment in different ways

CLUSTAL O(1.2.4) multiple sequence alignment

```
human      MSRQFTCKSGAA---AKGGF-----SGCSAVLSGGSS----SSFRAGSKGLSGG      42
sea        MSRQFTCKSGAA---TKGGF-----SGCSAVLSGGSS----SSYRAGKGISGG      42
Atlantic   MSKRVTQTSYVRSSAPRNYSSASYSGPSMGGSRQSYARSST--FGGASRGMGGGGGGGG      58
black      MSRQ----STVRIQRGRSGFSAASAIVPN--TCRTSFSSRSVTRVSGSCNAGSGFSRVGG      54
          **::      .      :      *      *      *      *      *      *      *
```

Aligned sequences

```
human      FGSRSLYSLGGVRSLL--NVASGSGKSGGYGFGGRGRASGFAGSMFGSVALGPVCPTVCPPG      100
sea        FGSRSLYSLGGIRNISFNMTSGSGKSGGYGFGGRGRASGFAGSMFGSVALGPVCPVCPG      102
Atlantic   FISSSA-YGG-----MGLGSSMAGGGGGMGMGGGFGGGMGMGMG-----A      99
black      FGSKSLYNVGGCKRISV-AGRGGGYGFGAGFGGGAGSVY--CGGGGFG-----M      100
          * * *      **      *..      *      *      *      *      *      *
```

```
human      GIHQVTVNESLLAPLNVELDPEIQKVRQEREQIKALNNKFASFIDKVRFLEQQNQVLET      160
sea        GIHQVTVNESLLAPLNVELDPEIQKVRQEREQIKALNNKFASFIDKVRFLEQQNQVLET      162
Atlantic   PITAVTVNKSLLAPLNLAIDPNIQAVRTHEKEQIKLNNRFASFIDKVRFLEQQNQVLET      159
black      PANLVSVNQSLKPLNLEIDPSIQIRKEEKEQIKLNNKFASFIDKVRFLEQQNQVLET      160
          *:::***      ***:      ***:      *      *      *      *      *      *
```

Symbols indicate matches or partial matches

```
human      KWELLQQLDLNCKNNLEPILEGYISNLRKQLETLSGDRVRLDSELNRVDRVVEDYKKRY      220
sea        KWELLQQLDLNCKNNLEPILEGYISNLRKQLETLSGDRVRLDSELNRVDRVVEDYKKRY      222
Atlantic   KWNLLQQQT--TTRSNIDAMFEAYITNLRRLDGLGNDKMKLEADLHNMQGLVEDFKNKY      217
black      KWSLLQEQQMKTVMKNLEPLFETYINNLRMQLSLLSDKGRLEGEVNTQYLVEDFKKKY      220
          **::**      .      .:::      :*      **      *      *      *      *      *
```

```
human      EEEINKRTAAENEFVLLKKDVAAYANKVELQAKVESMDQEIKFRCLEAEITQIQSHI      280
sea        EEEINKRTAAENEFVLLKKDVAAYANKVELQAKVESMDQEIKFRCLEAEITQIQSHI      282
Atlantic   EDEINKRASVENEFVLLKKDVGAYMNKIELEAKVDALQDEINFLRAVYAEARELQGGI      277
black      EDEINKRTAAENEFVLLKKDVAAYMNKVELQAKVDALTEEINFLRALYAEARELQGGI      280
          *::**      .      *      *      *      *      *      *      *      *
```

```
human      SDMSVILSMDNNRDLDSIIDEVRTQYEEIALKSKAEAEALYQTKFQELQLAAGRHDG      340
sea        SDMSVILSMDNNRDLDSIIDEVRAQYEEIALKSKAEAEALYQTKFQELQLAAGRHDG      342
Atlantic   KDTSVIVEMDNRGLMDSDIIVAEVRAQYEEIANRTRGEAGWYKQFEEMQTSAGQYGEN      337
black      SDTSVVLTMNNRDLDSIIDSEVKAQYEDIANRSRAEAEWYQTKYEELEQATAGRHDG      340
          . *      *::      ***      .      *      *      *      *      *      *
```

```
human      LKNTKNEISELTRLIQIRSEIENVKKQASNLETAIADAEQRGDNALKDARAKLDELEGA      400
sea        LKNTKNEISELTRLIQIRSEIENVKKQASNLETAIADAEQRGDNALKDARAKLDELESA      402
Atlantic   VRNTKSEIAEINRMISRLQNEIDNVKSQRASLEAQIAEAEERGELAVKDAKLRIELEEA      397
black      LRNTKQEISELNRHVQRLRSEIDSVKKQCANLKAATADAEERGELALKDAKLALELEDA      400
          :::**      *      *      *      *      *      *      *      *      *
```

Numbers show the position in the sequence. Gaps mean that these numbers are not the same.

```
human      LHQAKEELARMLREYQELMSKLALDMEIATYRKLESEECRMSGEFPSPVSIISSTS      460
sea        LHQAKEELARMLREYQELMSKLALDMEIATYRKLESEECRMSGEFPSPVSIISSTS      462
Atlantic   LQRAKQDMARQVREYQELMNVKLALDIEIATYRKLEGEECRIGSGTSTAHISISSSS      457
black      LQQAADLARQLREYQELMNVKLALDIEIATYRKLEGEECRWGNVTLLVLVLEDPQV      460
          *::**      :*      *      *      *      *      *      *      *
```

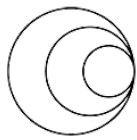
```
human      ----GGSVYGFPRPMVSGGYVANSNCISG-VCSVRGGEGRSGS--ANDYKDTLKGKSS      513
sea        ----GGGYGFRPSSVSGGYVASSSSCISG-VCSVRGGEGRSGS--SSDYKDTLGRGSS      515
Atlantic   SGGGGGGGGGFLGYGGGGGGGGGGYSKSSYSSSGGGGGGGFGSGSMIGGGGGGGGGG      517
black      LVNLG--ISEMHM--LNGVSRGRGKGLVNSVTL-----                     490
          *      :      *      *      *
```

```
human      LS-----A-PSKKTSR--      523
sea        QS-----T-SSKKASR--      525
Atlantic   LSSGGGVSYSSSSSVRSRGF      538
black      -----                     490
```

The symbols underneath the aligned sequences indicate how similar the sequences are.

- A dash indicates a gap, where there is no sequence.
- * An asterisk indicates that all the amino acids at that position are the same.
- : A colon indicates that they are very similar.
- . A dot indicates there is some similarity.
- A blank indicates no similarity at all.

The similarity is based on both chemical similarity between the amino acids at that position and also how many of them match between the different sequences.



Questions:

1. Could you do this without a computer? How long do you think it would take to align all these sequences and spot similarities and differences between them by hand?
2. What if you had to do this for all 20,000 protein coding genes in a species?

Step 6: Change your view

There are other ways to see the alignments. Click on [Show colors](#).



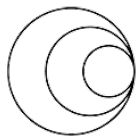
The colours can make it easier to see similarities and differences – it is easy to spot a column of matching colours than to spot a column of matching letters. There are 20 different amino acids, and some are more similar than others. Chemically similar amino acids are coloured the same, so we can most easily identify changes that are likely to affect the protein.

For example, one column shows R, R, K and K, all in magenta. R and K are Arginine and Lysine, respectively, which are both alkaline. In contrast, another column shows E, E (blue), K (magenta) and Q (green), which represent Glutamic acid (acidic), Lysine (alkaline) and Glutamine (with an amide group).



Questions:

- Can you find any sections in the alignment with long runs of matching sequence? What is the longest you can find?
- Look at position 5 in the alignment. What amino acids can you see there? Why have these been coloured differently? This is a great chart to look at for amino acids: <https://www.compoundchem.com/2014/09/16/aminoacids/>
- Why do you think that some parts of the protein have lots of similarity between species, and other parts do not?



Another option is to see a tree constructed from the sequences, showing us which sequences are most similar to each other. Click on [Phylogenetic tree](#) at the top of the page.

Results for job clustalo-l20201015-095701-0397-7099231-p2m

[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



human 0.03735
sea 0.0334
Atlantic 0.25266
black 0.17114

Questions:

6. Which sequences are most similar to each other?

7. Why do you think this is?

Extended practical

Only if you have time or are particularly interested in this.

Why not see if you can find *KRT71* in more species and carry out an alignment of more sequences?

Why not find a different human gene and try searching for that?

Some genes to try:

ADAM30

NOTCH2NLR

PPIAL4A

LIX1L

GJA8