Multiple Sequence Alignments: *Clustal Omega*

Susan Parrish
McDaniel College

Last Update: 12/2021

1

Multiple Sequence Alignments & Building Phylogenetic Trees

- Basic *BLAST*: input query and search database for homologous sequences and see pair-wise matches (query/subject)
- However, what if you want to specify the sequences to align?
  1) Two sequences: NCBI *BLAST* — (*bl2seq*)
  2) Three or more sequences: *Clustal Omega*
  3) Building trees from multiple sequence alignments: *Clustal Simple Phylogeny*

2

Clustal: Creating a Multiple Sequence Alignment

- You SELECT sequences to align to identified conserved nucleotides or amino acids
- What conserved sequences might you identify in a multiple sequence alignment of DNA sequences?
- What conserved sequences might you identify in a multiple sequence alignment of amino acid sequences?
- How could you use multiple sequence alignments to build phylogenetic trees?

3

Consensus **Nucleotide** Sequences

- A consensus nucleotide sequence is derived by making a multiple sequence alignment and calculating the most represented nucleotide at each position
- Provides insight into the functional regions of a given sequence (more important for function = conservation through evolution)
- Example: If a DNA binding site is necessary to recruit a specific protein, that DNA sequence could be conserved during evolution

4

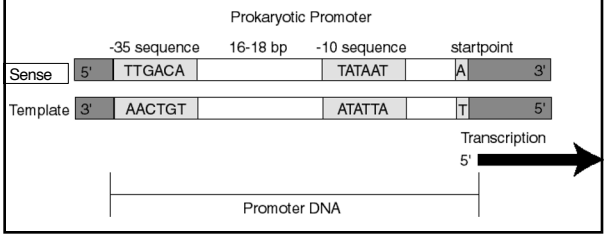
Consensus Nucleotide Sequences: **Promoters**

- Determine consensus sequence found in promoter for RNA polymerase recruitment by multiple sequence alignments of regions upstream of genes
- Closer a promoter is to consensus sequence, stronger the promoter in driving expression of gene (better at recruiting RNA polymerase)
- Individual promoters usually differ from the consensus at one or more positions

5

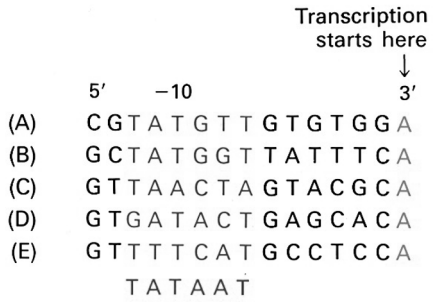
Prokaryotic -10 & -35 Promoter Consensus Sequence

- The prokaryotic promoter consensus sequence
- Closer to this sequence, more easily recognized by prokaryotic RNA polymerase



6

Prokaryotic -10 & -35 Promoter Consensus Sequence



7

A Eukaryotic Consensus Sequence

Deep Thoughts:
Any idea what this consensus sequence controls?

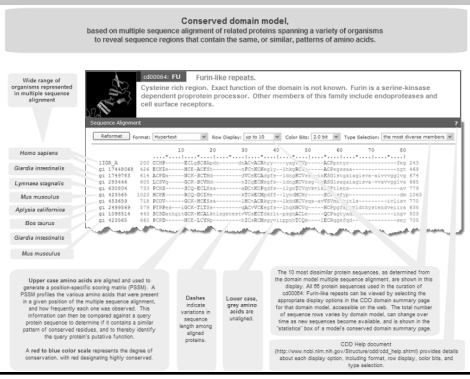
8

Proteins: Conserved Domains

- Can find conserved domains in proteins by performing multiple sequence alignments of amino acid sequences
- If a region of a protein has a particular function that is important (e.g., DNA binding domain/ protein-protein interaction domain/ enzymatic active site) will see conservation of amino acids within that domain
- Conserved domains tend to fold independently of other parts of protein (therefore structure of domain conserved)
- Appears that conserved domains have been "shuffled" during evolution to create new proteins

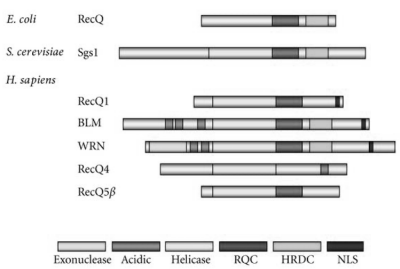
9

Proteins: Conserved Domains



10

Many Proteins have Multiple Conserved Domains



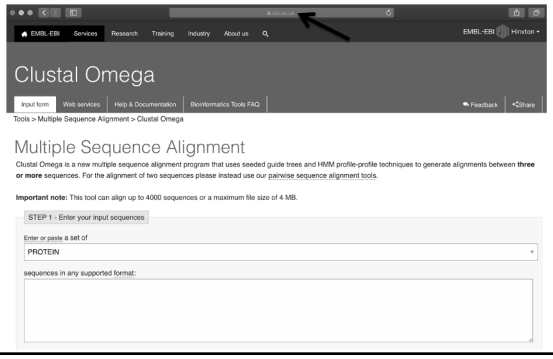
11

Clustal Omega: Creating a Multiple Sequence Alignment

- Select sequences to align:
  - from NCBI or other sequence databases (e.g., FlyBase)
  - Nucleotide or amino acid
  - Obtain FASTA sequences and paste into text document
- Click on the link for the EMBL-EBI *Clustal Omega* page:
  - https://www.ebi.ac.uk/Tools/msa/clustalo/
- Copy and paste selected FASTA formatted sequences into *Clustal Omega* window

12

Clustal Omega: Creating a Multiple Sequence Alignment



13

Clustal Omega: Creating the Initial Alignment

14

Generating the Sequence Files

- Create a new text file
  - Using WordPad in MS Windows, TextEdit in macOS
- For each sequence you like to align:
  - Retrieve the sequence in FASTA format
  - Copy and paste it into the text file, then hit return
  - Save the text file
- Repeat the previous step until the text file includes all the sequences you like to align

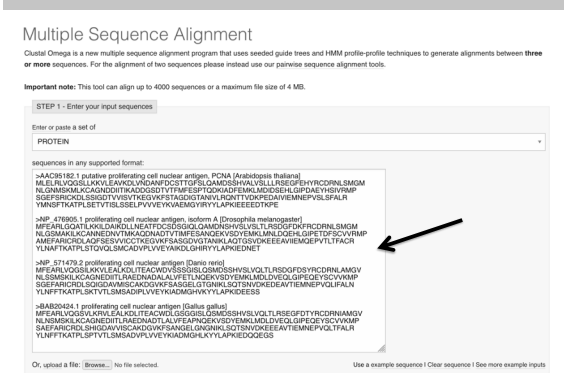
15

Generating the Sequence Files

- Prudent to save the text file after you add each sequence to the file (in case of malfunction)
- After you have created the sequence file:
  - Select all the FASTA formatted sequences in the text document
  - Copy and paste the sequences into the "Enter your input sequences" text box in the Clustal Omega window

16

Clustal Omega: Enter Your Input Sequences



17

Clustal Omega: The Alignment Report

18

Clustal Omega: The Alignment Report

- Consensus symbols for protein alignments

Symbol	Description
*	Position with a fully conserved residue
:	Conservation between groups of strongly similar properties
.	Conservation between groups of weakly similar properties
	Position with residues that have different properties

- Consensus symbols for nucleotide alignments

Symbol	Description
*	Position with fully conserved nucleotide
	Position with different nucleotides

Ignore the ":" and "." consensus symbols for nucleotide alignments

19

Clustal Omega: The Alignment Report

Results for job clustalo-I20191224-074035-0184-47913577-p1m

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

Download Alignment File | Show Colors

```
CLUSTAL O(1.2.4) multiple sequence alignment


AAC95182.1      MLELRLVQGSLLKKVLEAVKDLVNDANFDCSTTGFSLQAMDSSHVALVSLLLRSEGFEHY   60
NP_476905.1     MFEARLGQATILKKILDAIKDLLNEATFDCSDSGIQLQAMDNSHVSLVSLTLRSDGFDKF   60
NP_571479.2     MFEARLVQGSILKKVLEALKDLITEACWDVSSSGISLQSMDSSHVSLVQLTLRSDGFDSY   60
BAB20424.1      MFEARLVQGSVLKRVLEALKDLITEACWDLGSGGISLQSMDSSHVSLVQLTLRSEGFDTY   60
                *:* ** *.::**::*:*:***:.:* :* .  *:.**:**.***:**.* ***:**: :

AAC95182.1      RCDRNLSMGMNLGNMSKMLKCAGNDDIITIKADDGSDTVTFMFESPTQDKIADFEMKLMD   120
NP_476905.1     RCDRNLSMGMNLGSMAKILKCANNEDNVTMKAQDNADTVTIMFESANQEKVSDYEMKLMN   120
NP_571479.2     RCDRNLAMGVNLSSMSKILKCAGNEDITTLRAEDNADALALVFETLNQEKVSDYEMKLMD   120
BAB20424.1      RCDRNIAMGVNLNSMSKILKCAGNEDITTLRAEDNADTLALVFEAPNQEKVSDYEMKLMD   120
                *****::**:**..*:*:****.*:* :*::*:.:*::::::** .*:*::*:*****:
```

20

Clustal Omega: The Alignment Report

- **COLOUR**

Show Colors
A button labeled 'Show Colors' will be displayed in the Alignment section of results page. If you press this button the alignment will be show in color according to the table below.

NOTE: This option only works when you have chosen ALN or GCG as the output format.

AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + Amine + Basic - Q
Others	Gray	

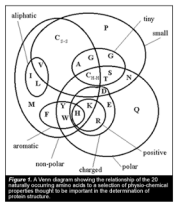


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties thought to be important in the determination of protein structure.

21

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

Download Alignment File | Hide Colors

```
CLUSTAL O(1.2.4) multiple sequence alignment


AAC95182.1      MLELRLVQGSLLKKVLEAVKDLVNDANFDCSTTGFSLQAMDSSHVALVSLLLRSEGFEHY   60
NP_476905.1     MFEARLGQATILKKILDAIKDLLNEATFDCSDSGIQLQAMDNSHVSLVSLTLRSDGFDKF   60
NP_571479.2     MFEARLVQGSILKKVLEALKDLITEACWDVSSSGISLQSMDSSHVSLVQLTLRSDGFDSY   60
BAB20424.1      MFEARLVQGSVLKRVLEALKDLITEACWDLGSGGISLQSMDSSHVSLVQLTLRSEGFDTY   60
                *:* ** *.::**::*:*:***:.:* :* .  *:.**:**.***:**.* ***:**: :

AAC95182.1      RCDRNLSMGMNLGNMSKMLKCAGNDDIITIKADDGSDTVTFMFESPTQDKIADFEMKLMD   120
NP_476905.1     RCDRNLSMGMNLGSMAKILKCANNEDNVTMKAQDNADTVTIMFESANQEKVSDYEMKLMN   120
NP_571479.2     RCDRNLAMGVNLSSMSKILKCAGNEDITTLRAEDNADALALVFETLNQEKVSDYEMKLMD   120
BAB20424.1      RCDRNIAMGVNLNSMSKILKCAGNEDITTLRAEDNADTLALVFEAPNQEKVSDYEMKLMD   120
                *****::**:**..*:*:****.*:* :*::*:.:*::::::** .*:*::*:*****:

AAC95182.1      IDSEHLGIPDAEYHSIVRMPSGEFSRICKDLSSIGDTVVISVTKEGVKFSTAGDIGTANI   180
NP_476905.1     LDQEHLGIPETDFSCVVRMPAMEFARICRDLAQFSESVVICCTKEGVKFSASGDVGTANI   180
NP_571479.2     LDVEQLGIPEQEYSCVVKMPSGEFARICRDLSQIGDAVHISCAKDGVKFSASGELGTGNI   180
BAB20424.1      LDVEQLGIPEQEYSCVVKMPSAEFARICRDLSHIGDAVVISCAKDGVKFSANGELGNGNI   180
                :* *:****: :: .:*:**: **:***:**: :.::*:*. :*:***** *::*..**

AAC95182.1      VLRQNTTVDKPEDAIVIEMNEPVSLSFALRYMNSFTKATPLSETVTISLSSELPVVVEYK   240
NP_476905.1     KLAQTGSVDKEEEAVIIEMQEPVTLTFACRYLNAFTKATPLSTQVQLSMCADVPLVVEYA   240
NP_571479.2     KLSQTSNVDKEDEAVTIEMNEPVQLIFALNYLNFFTKATPLSKTVTLSMSADIPLVVEYK   240
BAB20424.1      KLSQTSNVDKEEEAVTIEMNEPVQLTFALRYLNFFTKATPLSPTVTLSMSADVPLVVEYK   240
                 * *. .*** ::*: ***:*** * ** .*:* ********  * :*:.::*:****

AAC95182.1      VAEMGYIRYYLAPKIEEEEDTKPE       264
NP_476905.1     IKDLGHIRYYLAPKIEDNET----       260
NP_571479.2     IADMGHVKYYLAPKIDEESS----       260
BAB20424.1      IADMGHLKYYLAPKIEDQQEGS--       262
                : ::*:::******:::.
```

22

Simple Phylogeny: Making a Phylogenetic Tree

- *Simple Phylogeny* can only assemble a phylogenetic tree from multiple sequence alignments generated by *Clustal Omega*

23

Simple Phylogeny: Making a Phylogenetic Tree

Tree format

Determines the outputs that the Simple Phylogeny tool produces.

Format	Description	Value
Default	Newick/PHYLIP format tree file	phylip
Clustal	Clustal format file in addition to the PHYLIP tree	nj
Distance Matrix	Distance matrix file in addition to the PHYLIP tree	dist
NEXUS	NEXUS format file in addition to the PHYLIP tree	nexus

Default value is: Default [phylip]

Distance correction

Controls whether Simple Phylogeny attempts to correct for multiple substitutions at the same site. This is recommended to be set 'on' for more divergent sequences and has the effect of stretching branch lengths. For very divergent sequences the distances cannot be reliably corrected.

Type	Value
off	false
on	true

Default value is: off [false]

Exclude gaps

With this option enabled columns where any of the sequences in the input have a gap will be excluded, forcing the alignment to use only positions where information can be included from all sequences.

Type	Value
off	false
on	true

Default value is: off [false]

24

PHYLIP: **Phyl**ogeny **I**nference **P**ackage

PHYLIP

A new release of PHYLIP, version 3.698, is now available as source code. This release differs in correcting the consensus tree bug that was recently pointed out, and in its license -- from version 3.696 on, we have had an open source license, so that PHYLIP can be distributed with other software that has commercial licenses or has a restrictive open-source source license. MacOS executables are at version 3.695, with the old license, but I will update them soon.

Note -- recently users of 64-bit Windows systems have had difficulty running the Java front ends for the Drawgram and Drawtree tree-drawing programs. This was because this was 32-bit code. With the version 3.698 executables, we now have a 64-bit distribution available and these should run properly. The 3.695 Windows executables that are 32-bit code is also still available, for people who have older 32-bit Windows systems.

PHYLIP is a *free* package of programs for inferring phylogenies. It is distributed as source code, documentation files, and a number of different types of executables. These Web pages, by Joe Felsenstein of the Department of Genome Sciences and the Department of Biology at the University of Washington, contain information on PHYLIP and ways to transfer the executables, source code and documentation to your computer.

- A general description of PHYLIP.
- Programs in the PHYLIP package
- About the Executables
- About the Source code ... compiling it yourself
- The documentation web pages for PHYLIP can be read here
- Get me PHYLIP
- How to install PHYLIP
- Frequently asked questions
- PHYLIP's Facebook page for discussing problems.
- An excellent guide to using PHYLIP with molecular data is available here.
- PHYLIP on the web (HTML documentation, server services)
- Current and future versions of PHYLIP (including new features)
- Older versions of PHYLIP, including version 3.5
- Bugs in the package, known or recently fixed
- Phylogeny programs available elsewhere
- Credits (people, grants etc.)

25

Simple Phylogeny: Making a Phylogenetic Tree

Clustering method

Clustering Methods

Method	Description	Value
Neighbour-joining	Uses the neighbour-joining algorithm to construct trees from the distance matrix	Neighbour-joining
UPGMA	Uses the faster UPGMA tree construction algorithm	UPGMA

Default value is: Neighbour-joining

Percent identity matrix

Output the percentage identity matrix

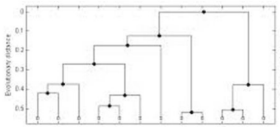
Type	Value
off	false
on	true

Default value is: off [false]

26

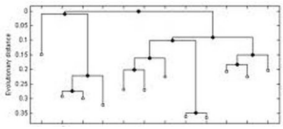
Neighbor Joining vs. UPGMA

UPGMA



- **Assumption: Divergence of sequences is assumed to occur at a constant rate**
- **Distance to root is equal**

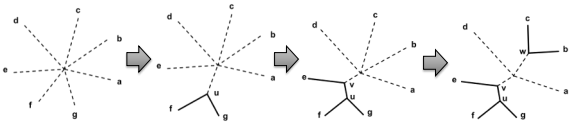
Neighbor Joining



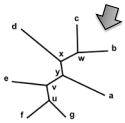
- **Constructs an unrooted guide tree from a distance matrix**
- **We do not assume constant rate of evolution**

27

Neighbor Joining



- Start with an unresolved (star) tree
- Find the two nodes with the shortest distance and join them together
- Repeat the process until the tree is fully resolved



"Neighbor joining 7 taxa start to finish diagram.svg" by Tomfy

28

Simple Phylogeny: Making a Phylogenetic Tree

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

Download Alignment File | Hide Colors

```
CLUSTAL O(1.2.4) multiple sequence alignment

AAC95182.1      MLELRLVQGSLLKKVLEAVKDLVNDANFDCSTTGFSLQAMDSSHVALVSLLLRSEGFEHY   60
NP_476905.1     MFEARLGQATILKKILDAIKDLLNEATFDCSDSGIQLQAMDNSHVSLVSLLTLRSDGFDKF   60
NP_571479.2     MFEARLVQGSILKKVLEALKDLITEACNDVSSSGISLQSMDSSHVSLVQLTLRSDGFDSY   60
BAB20424.1      MFEARLVQGSVLKRVLEALKDLITEACWDLGSGGISLQSMDSSHVSLVQLTLRSEGFDTY   60
                *:* ** *.:;**:;*:*:***:.:* :* .  *:.**:**.***:**.* ***:**: :
```

Simple Phylogeny

A phylogenetic tree has been generated from this alignment. If you want more control over the algorithms and/or parameters you can create a new one:

Send to Simple Phylogeny

29

Simple Phylogeny: Making a Phylogenetic Tree

Simple Phylogeny

This tool provides access to phylogenetic tree generation methods from the ClustalW2 package. Please note this is NOT a multiple sequence alignment tool. To perform a multiple sequence alignment please use one of our MSA tools.

STEP 1 - Enter your multiple sequence alignment

Enter or paste a multiple sequence alignment in any supported format:

clustalo-I20191224-074035-0184-47913577-pfm

Or, upload a file: Browse... No file selected.

Use a example sequence | Clear sequence | See more example inputs

CHANGE to "on"!

STEP 2 - Set your Phylogeny options

TREE FORMAT	DISTANCE CORRECTION	EXCLUDE GAPS	CLUSTERING METHOD	P.I.M.
Default	off	off	Neighbour-joining	off

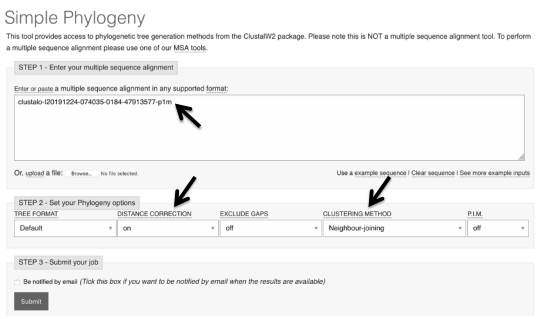
STEP 3 - Submit your job

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

Submit

30

Simple Phylogeny: Making a Phylogenetic Tree

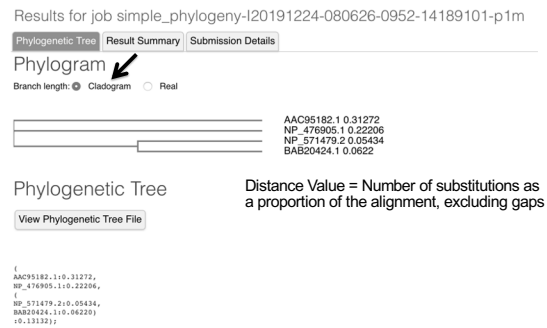


31

- A Cladogram is a branching diagram (tree) assumed to be an estimate of a phylogeny where the branches are of equal length, thus cladograms show common ancestry, but do not indicate the amount of evolutionary "time" separating taxa.

32

Simple Phylogeny: Making a Phylogenetic Tree



Distance Value = Number of substitutions as a proportion of the alignment, excluding gaps

33