

Lab 4.1: Multiple Sequence Alignment

```
AKEELERQACWL SKHPEVKQTVEGHC DERGTREYNL ALGERRAAAAKQFL ANKG IAHNPL  
NLT ILASLVPHLRKSPKTTLYIEGHTDERGAAAYNL ALGARRANAKQYL IKQG IASDFL  
IHKTLRGVARILVE YPDTSLVIEGHTDSTGSDTTNOVL SEKRAE SURE SILL SQGVAAGRA  
GMQTUQKIGRILSDVG IKHMRUDGHTDSUGKDDYNOQL SYORAL AVADTL ATUG IPKSN I  
SYGDVKNLAD FMAQYPATWVEVAGHTDSIGPDAYNOQL SQPRADREKQVLKEDGVAPSP I
```

Jennifer Gardy

Molecular Biology & Biochemistry

Simon Fraser University



creativecommons
COMMONS DEED

Attribution-ShareAlike 2.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works
- to make commercial use of the work

Under the following conditions:

 **BY: Attribution.** You must give the original author credit.

 **Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the [Legal Code \(the full license\)](#).

[Disclaimer](#) 

Goals

- Learn the basics of multiple sequence alignments (MSAs) and the Clustal program
- Understand how alignment settings can significantly affect an alignment
- Complete questions 1 & 2 in the phylogeny assignment

Outline

- MSAs:
 - Purpose
 - Automated alignment considerations
 - Clustal's alignment strategy
 - Manual editing
- Research Question
- ClustalX with default parameters
- Varying alignment settings
- Deleting sequences/regions of sequences

MSAs: A Quick Review

- Why perform an MSA?
 - Visualize trends between homologous sequences
 - Shared regions of homology
 - Regions unique to a sequence within a family
 - Structural/functional motif
 - As the first step in a phylogenetic analysis
 - Useful for improving accuracy of structure predictions
- How does one perform an MSA?
 - By hand: too hard!
 - Automated alignment: Fast, but doesn't necessarily produce the "correct" alignment

Best approach = Automated alignment with manual editing

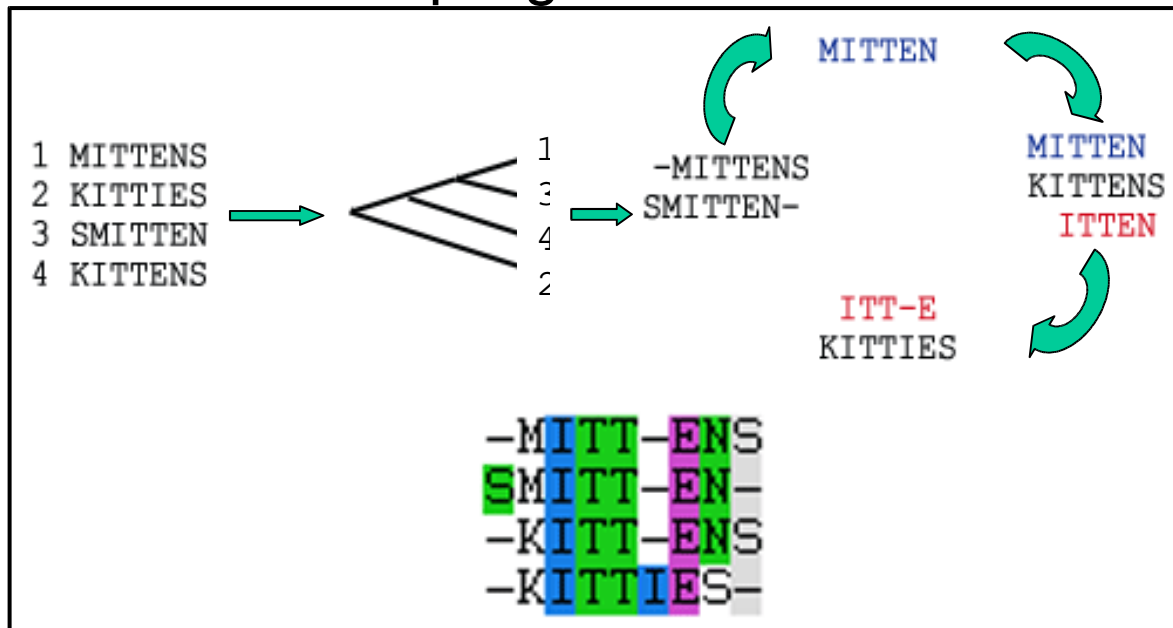
Automated alignment

- Technical considerations:
 - Select sequences carefully
 - Homologous over length, no unrelated sequences
 - The algorithm will align everything you give it!
 - Use an appropriate objective function
 - Most common = simple sum-of-pairs w/ gap penalties
 - Not evolutionarily ideal, but shown to perform well
 - Computational intensity
 - No current methods guarantee full optimization
 - 3 categories of heuristics:
 - Exact: close to optimal, can only use small number of sequences and sum-of-pairs OF
 - Progressive: most common, adds sequences to an alignment one-by-one, fast, no great potential for optimization
 - Iterative: produces an alignment, refines it through a series of cycles until no more improvements can be made

“Recent progress in MSAs: a survey. C. Notredame. Pharmacogenomics. PMID: 11

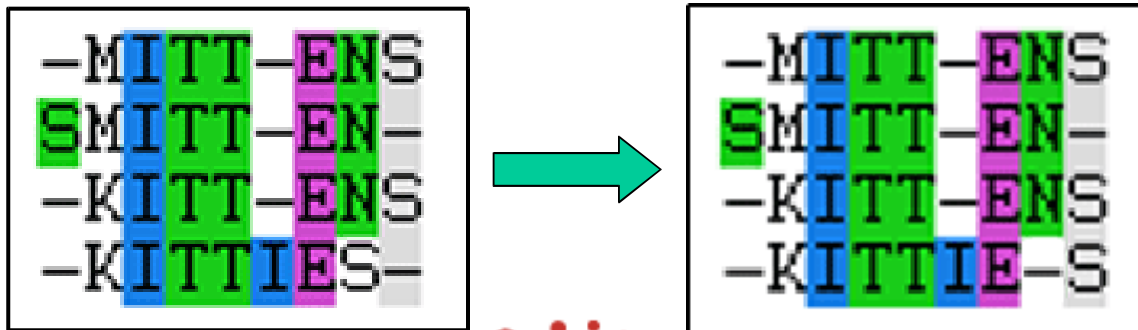
Clustal

- One of the most common MSA tools
- Uses sum-of-pairs with gaps OF
- Progressive alignment strategy:
 - Sequences used to make guide tree
 - Least dissimilar 2 seqs aligned, make consensus
 - Next closest seq aligned to consensus

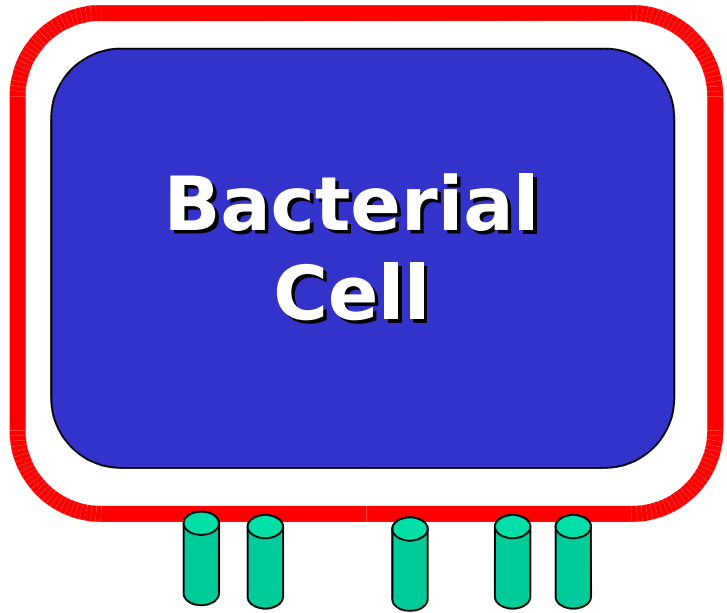


Manual Editing

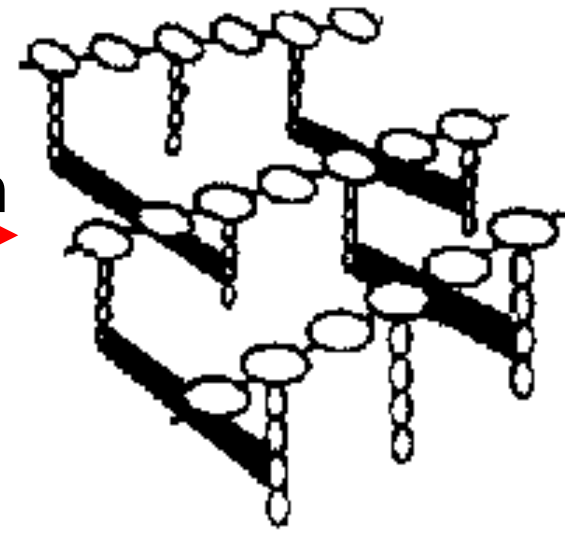
- “Human-assisted quasi-optimization”:
 - Fine adjustment of particular columns
 - May incorporate specific knowledge about sequences
 - Removal of gappy bits
 - Important for phylogenetic analysis
 - Removal of parts of sequences or whole sequences
 - Non-homologous regions
 - Sequence included by error



Research Question: Background



Peptidoglycan →



Peptidoglycan-associated Lipoproteins (PAL proteins)

What part of the PAL protein is involved in peptidoglycan binding?

Research Question: Strategy

- Used 1 PAL protein you identified to search NCBI databases for more PAL family proteins
- Found 4 more proteins from different bacteria

Next Step = Multiple Sequence Alignment

Do all 5 sequences contain a domain that may be involved in peptidoglycan binding?

Where in these proteins is this domain located?

Which residues in particular would you potentially target for further laboratory study for their possible role in PG binding?

Starting up ClustalX

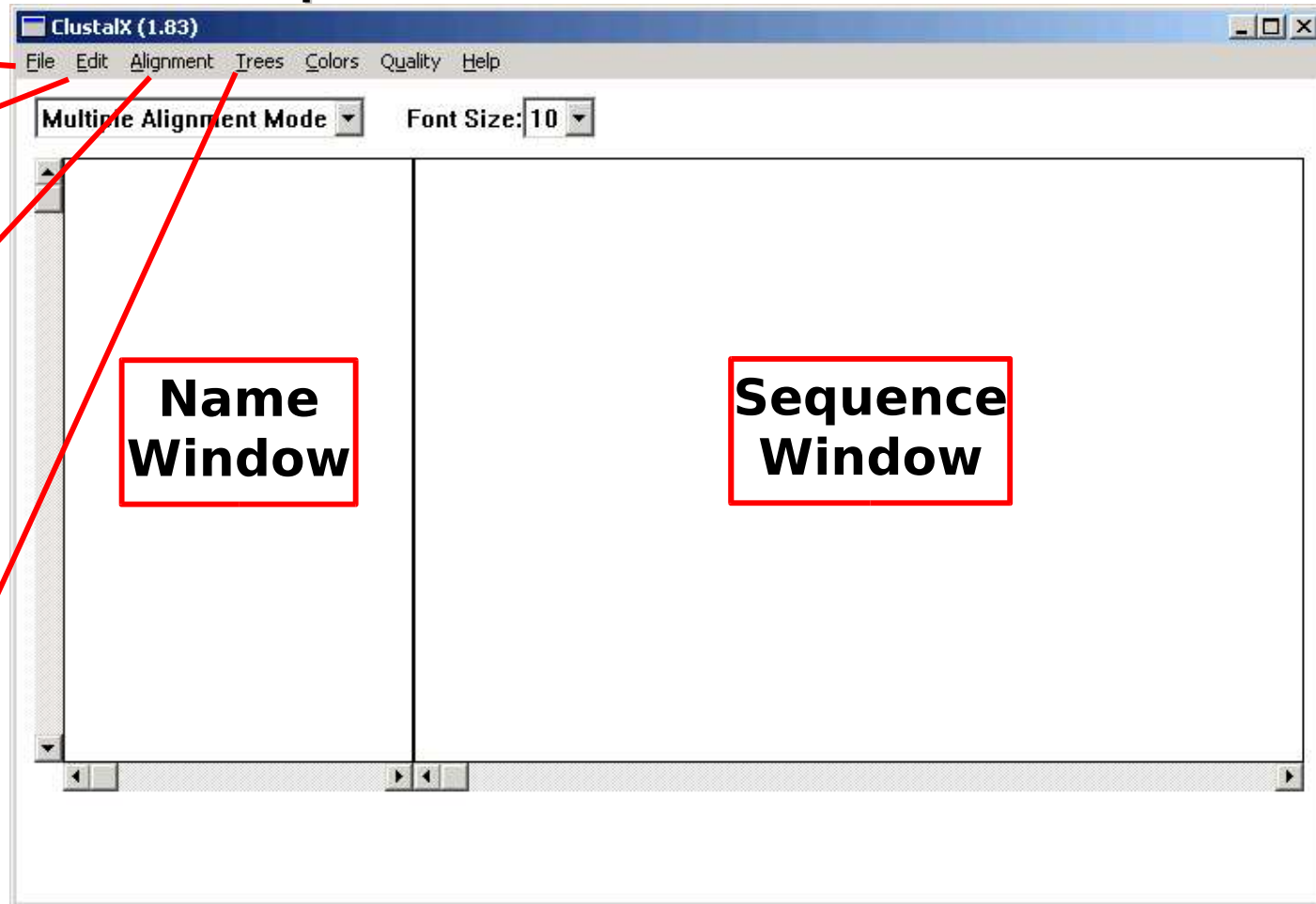
- Day 4 website > **PALproteins.txt**
- Start ClustalX - **\$ clustalx**

File:
-Load sequences

Edit:
-Remove all gaps

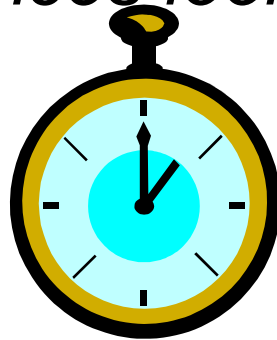
Alignment:
-Do complete alignment
-Alignment parameters

Trees:
-Bootstrapped NJ
-Output format options

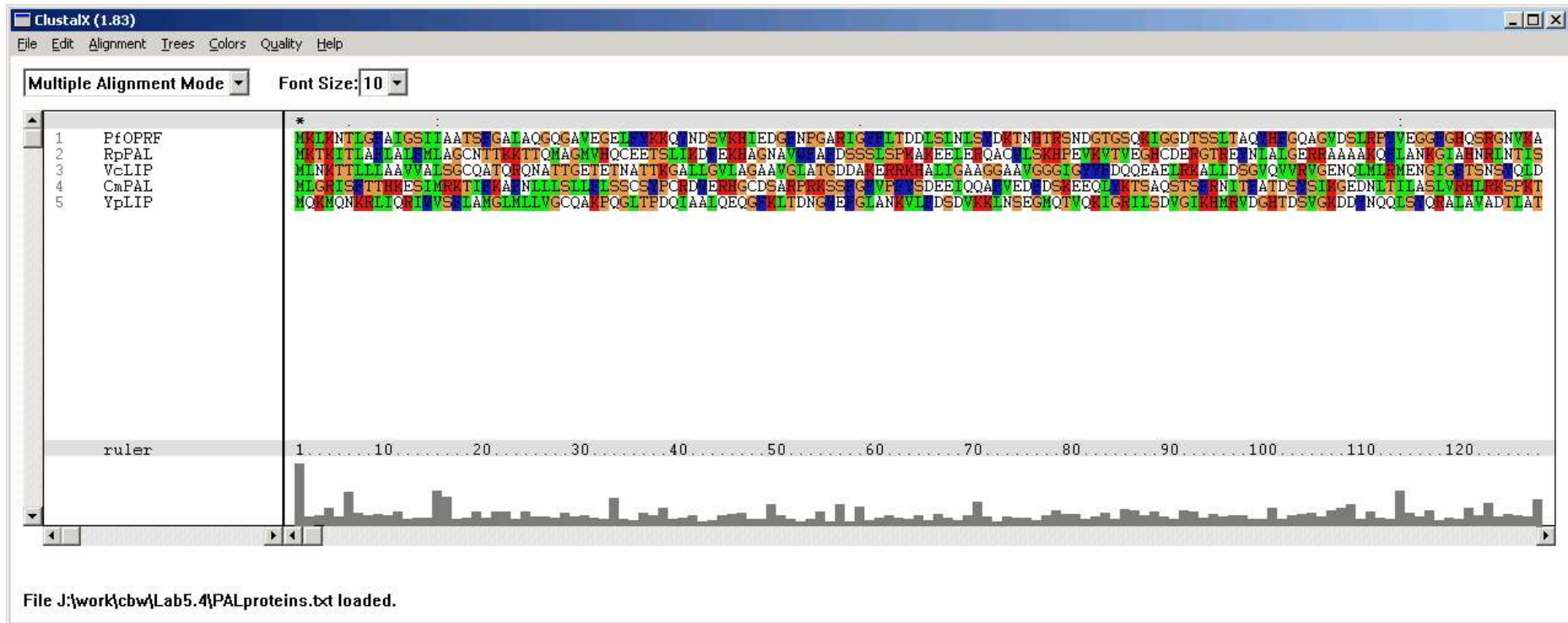


Starting up ClustalX

- **File > Load sequences > PALproteins.txt**
- Examine the sequences:
 - *How are unaligned sequences displayed?*
 - *Do the sequences look similar to each other?*



PAL Proteins in ClustalX



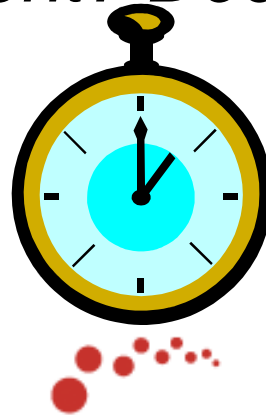
- Left-aligned, in order of input
- Default colouring (identity) – see help file for details
- Conservation score graph
- One long sequence

Let's Do An Alignment!

- **Alignment > Do complete alignment**
- Generates an .aln file

Examine Your Alignment

- *Is there a difference in the order of the sequences?*
- *Could the order of the input sequences affect your alignment?*
- *What effect does the large N-terminal domain have on your alignment?*
- *What effect will increasing the gap penalty have on your alignment? Decreasing it?*



Sequence Order

- Order has changed, & input order affects alignment:
 - Clustal’s “pairwise” strategy generates similarity values for each pair of sequences
 - The most similar pair is selected to build a consensus
 - The consensus is re-compared to the other sequences and new similarity values are generated
 - Lather, rinse, repeat
 - BUT... if two sequences have equal similarity values, Clustal orders them based on the order they were inputted in!

Let’s see that in pictorial form...

Sequence Order

—	A
—	B
—	C
—	D

	A	B	C	D
A	-			
B	.7	-		
C	.8	.2	-	
D	.6	.2	.5	-

—	BC
—	A
—	D

	A	BC
A	-	
BC	.75	-
D	.6	.45

- BC and BD both show the lowest dissimilarity index
- However the BC and BD consensus sequences can be quite different:

B = ELVIS BC = ELVIS BD = ELVIS
C = LIVES LIVES EVILS
D = EVILS --V-S E---S

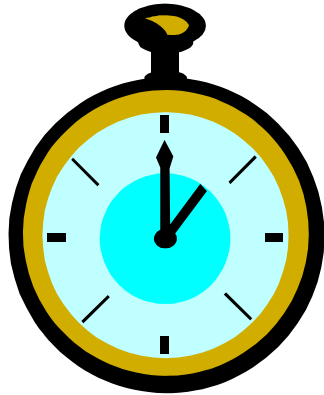
- Affects further similarity calculation

Unusually Long Sequences

- Including 1 much longer sequence may affect the alignment:
 - Evolutionarily, it indicates an insertion or deletion event
 - Not part of the homologous region(s)
 - Program will attempt to align it anyway
 - N-terminal aligned regions are unreliable

Gap Penalties

- **Shift-click** each sequence name to select
- **Edit > Remove all gaps**
- **Alignment > Alignment parameters > Multiple alignment parameters**
- Try a **Gap Opening Penalty** of 1, then 30

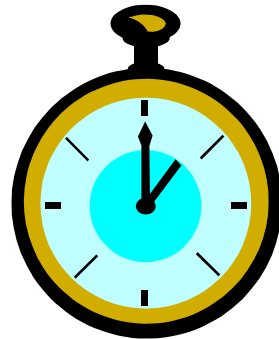


Important: Every time you make a new alignment, a new .aln file will be created. If you do not change the filename, the previous file will be overwritten.

Answer Question 1 in the phylogeny assignment

The Effect of Removing Sequences

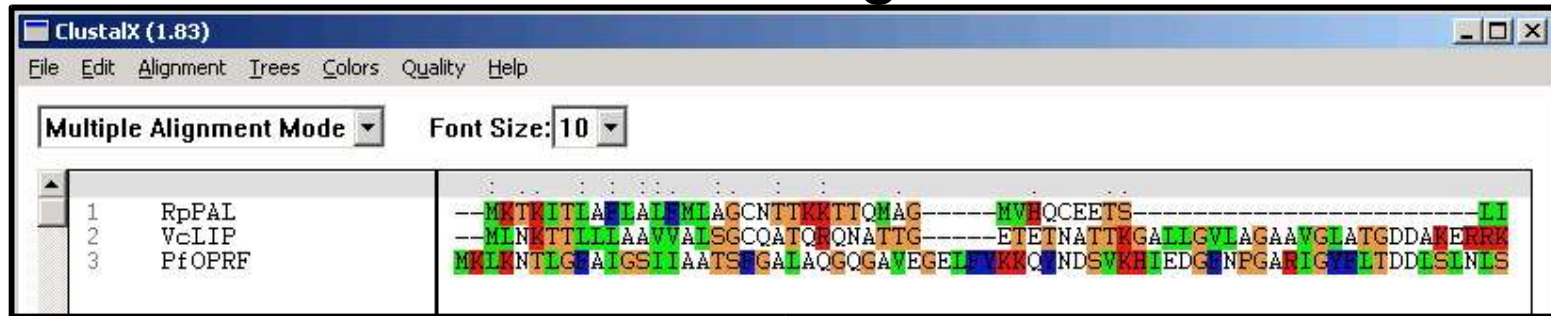
- Open PALproteins.txt in an editor
- Delete CmPAL and YpLIP, save the file
- Load this file in ClustalX
- Do an alignment with the default parameters
- Print this alignment, answer Question 2



– *What effect did removing the sequences have on your alignment?*

The Effect of Removing Sequences

- Increased N-terminal alignment



- What might this indicate?
- Signal peptide
- Not a meaningful homologous sequence
- Best to remove such regions:
 - Signal peptides
 - Other domains

Remainder of Lab Time

- Finish your assignment questions
 - Q1: Effect of changing gap penalties (have your team try out different values)
 - Q2: Annotated printout
- Begin the MSA for Module 3 of the Integrated Assignment (Section 3.2, Task 1)
 - Need to have completed Module 2
 - You have PLENTY of time for the IA and if you'd like to save it for later, that's OK!!!
- Use Clustal to check out your favourite gene/protein family
- Try web-based Clustal:
 - <http://www.ebi.ac.uk/clustalw/>