Practical 1: *Xenoturbella*

Placing enigmatic taxa in the Tree of Life

This practical was primarily written by **Rob Lanfear** (Australian National University).

Aims

The point of this practical is to learn how to align DNA sequences in the software *MEGA*, compare the fit of different nucleotide substitution models, and perform a phylogenetic analysis to figure out what kind of animal *Xenoturbella* is.

Background

Next time you're in the fjords of Scandinavia, try dredging up some mud from the sea floor and putting it in a bucket, and then sit back and wait for a bit. If you're lucky you might see one or two small, nondescript pale worms slowly crawling up the side of the bucket. They're probably *Xenoturbella bocki*.

Xenoturbella was first discovered in 1915, but not described until 1949. Ever since then its position in the Tree of Life has been the cause of intense debate. The biggest problem is that *Xenoturbella* doesn't have much in the way of morphology: it's really just a ciliated bag of cells with a mouth and no anus. Nobody quite knows how it breeds, and nobody has been able to keep it for more than a single generation in captivity. This is exactly the kind of thing that DNA and molecular phylogenetics are supposed to be useful for – every organism has DNA, and this should provide plenty of information to understand where *Xenoturbella* fits into the Tree of Life.

The first molecular phylogenetic analysis of *Xenoturbella* was big news (Noren and Jondelius 1997, *Nature*). That paper concluded that *Xenoturbella* was a protostome, and in particular that was almost certainly a bivalve mollusc. Other biologists then started to think about the morphology of *Xenoturbella* in new ways. One scientist thought that its embryology agreed very strongly with the molecular data: "I report here the previously unknown embryology of *Xenoturbella* that **unequivocally corroborates a bivalve relationship**" (Israelsson, 1999). So that's it, all sorted. Right? Both molecular and morphological evidence show that *Xenoturbella* is a mollusc. We'll see …

Not everyone was convinced about the embryology data (that's putting it a bit lightly). Because of this, some scientists at University College London decided to see if they could extract additional DNA data to follow up more on the DNA evidence. They used meticulous laboratory techniques to extract and sequence DNA from *Xenoturbella bocki*. We will be analysing these DNA data in this practical.

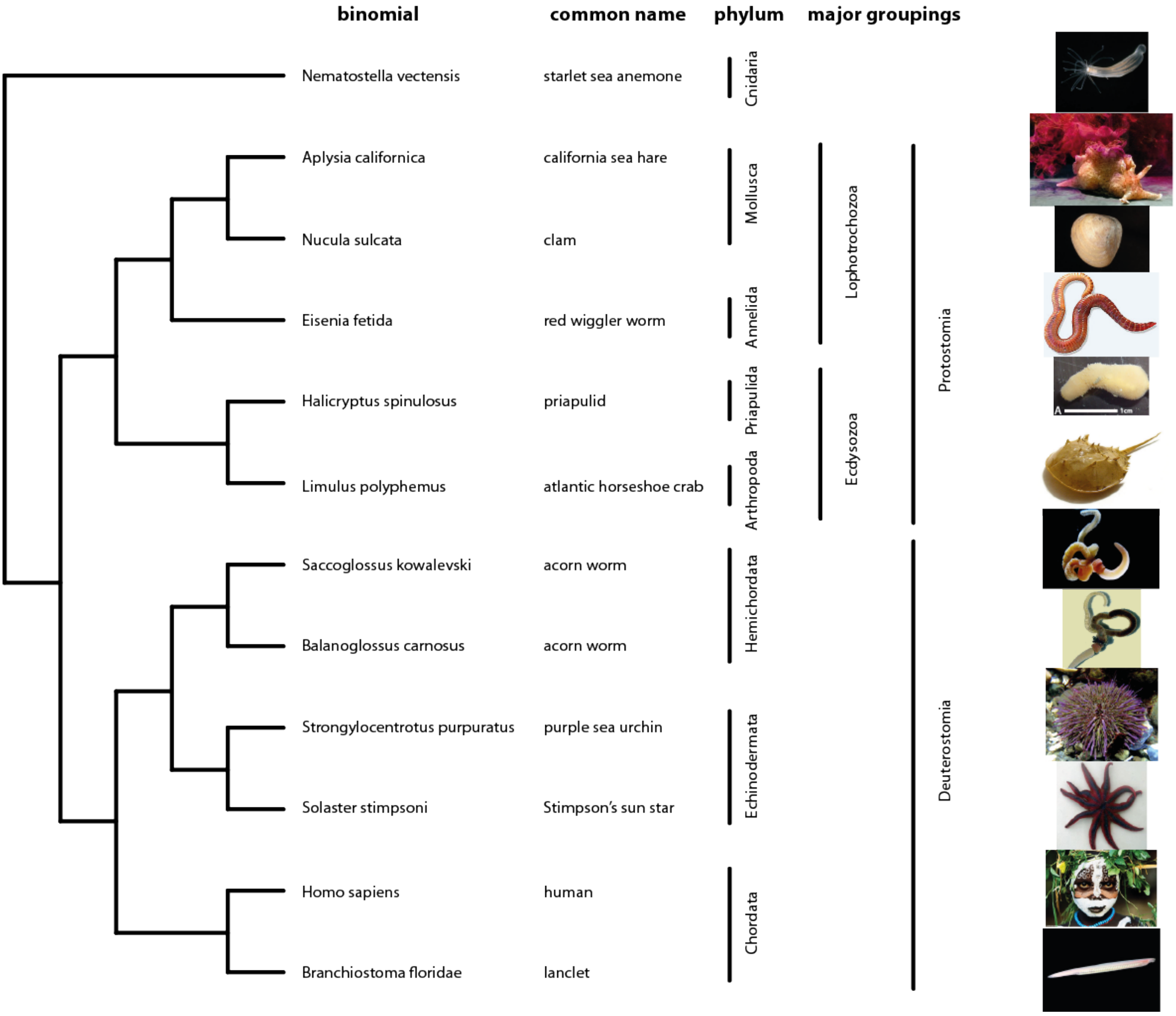
Outline of the practical

Before you begin, check that you have a recent version of MEGA (version 6 or 7) and the two data files, **animals.fasta** and **animals_cut.fasta**.

The file **animals.fasta** contains 13 DNA sequences from the *18S* gene. *18S* is a ribosomal RNA gene, which codes not for a protein but for a structural RNA molecule. It is the most commonly used molecule for resolving deep relationships in the Tree of Life. The taxa in the file have been carefully selected to represent a broad sampling of animal diversity. We know from lots of other molecular and morphological evidence what the relationships of most of these taxa should be, and these are shown below. By aligning and analysing the *18S* sequences from these taxa along with that from *Xenoturbella*, we should be able to get a pretty good first idea of where *Xenoturbella* fits in the phylogeny.

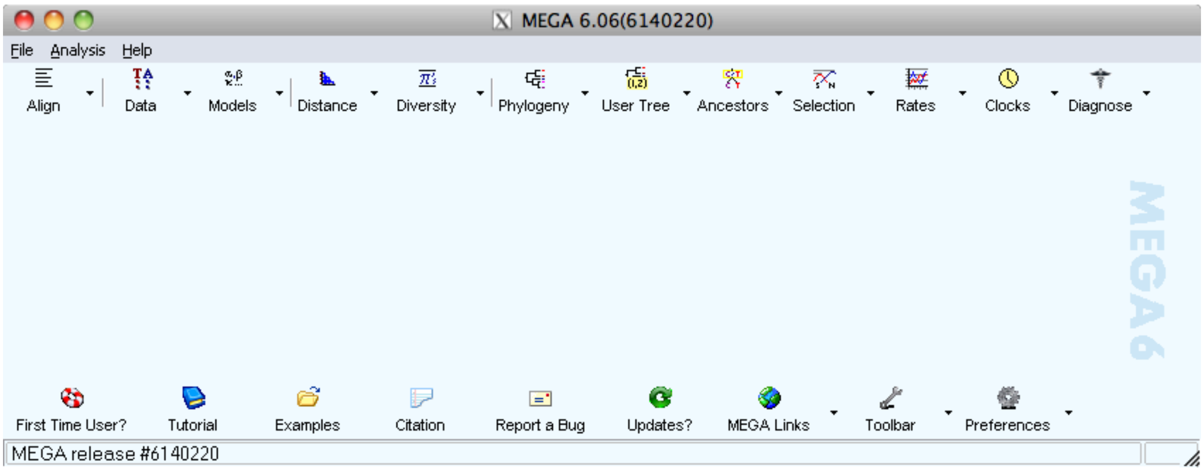
There are three parts to the practical:

a) DNA sequence alignment
b) Evolutionary model selection
c) Phylogenetic analysis



Part A: DNA sequence alignment

a) Open the program *MEGA*. This software provides an integrated framework for running various analyses of DNA sequence data.



b) Open the data file, **animals.fasta**, and select “Align” in the box that appears.

c) Have a look at the sequences, notice that they are unaligned. Move the slider at the bottom all the way to the right. Notice that the sequences differ in length.

Q. *Before we conduct any phylogenetic analyses, we need to align these sequences. What is the purpose of sequence alignment?*

. . . .

. . . .

. . . .

. . . .

Q. *What are ‘indels’?*

. . . .

. . . .

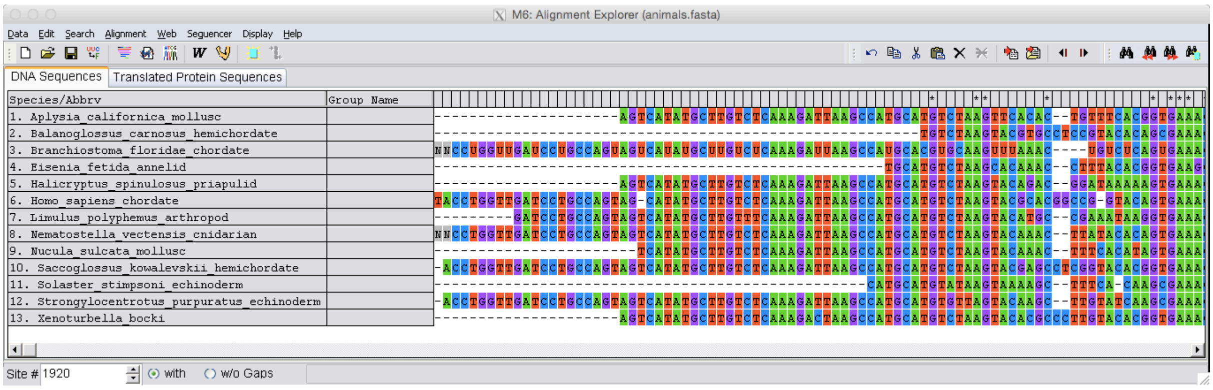
d) Try for a few minutes to align these sequences by eye. To do this, use your mouse to highlight some columns in one or more sequences, then use the left and right arrow symbols at the top of the screen to move that block of columns left and right. You can also use the keyboard: in Windows use alt + the arrow keys; on a Mac use alt + command + the arrow keys.

To get you started, try taking the sequence of *Aplysia californica*, highlight the first base, and move it 21 columns to the right. Notice that the first 50 or so bases of *Aplysia californica* now look like they are well aligned with those of *Branchiostoma floridae* (n.b.

you can move sequences up and down in the list by dragging with your mouse). Now try moving the first base of *Strongylocentrotus purpuratus* 1 column to the right and compare with the two sequences mentioned above. Notice anything?

e) Now we'll get the computer to align the sequences for us. In the **Alignment** menu, select "Align by ClustalW". ClustalW is one of the available alignment algorithms. The other algorithm available in *MEGA* is Muscle. These two are the most widely used alignment algorithms. *MEGA* will show the settings for ClustalW, including the penalties for postulating gaps. We will simply use the default values here. Click "OK" and wait for the sequences to be aligned.

f) Have a look at the alignment to see where gaps have been inserted. Check that the alignment hasn't made any obvious errors.


M6: Alignment Explorer (animals.fasta)
Data
Edit
Search
Alignment
Web
Sequencer
Display
Help
DNA Sequences
Translated Protein Sequences

Species/Abbrv	Group Name	
1. Aplysia_californica_mollusc		---------------------AGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTTCACAC--TGTTTCACGGTGAAA
2. Balanoglossus_carnosus_hemichordate		---TGTCTAAGTACGTGCCTCCGTACACAGCGAAA
3. Branchiostoma_floridae_chordate		NNCCUGGUUGAUCCUGCCAGUAGUCAUAUGCUUGUCUCAAAGAUUAAGCCAUGCACGUGCAAGUUUAAAC----UGUCUCAGUGAAA
4. Eisenia_fetida_annelid		---TGCATGTCTAAGCACAAAC--CTTTACACGGTGAAG
5. Halicryptus_spinulosus_priapulid		---------------------AGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTACAGAC--GGATAAAAAGTGAAA
6. Homo_sapiens_chordate		TACCTGGTTGATCCTGCCAGTAG-CATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTACGCACGGCCG-GTACAGTGAAA
7. Limulus_polyphemus_arthropod		---------GATCCTGCCAGTAGTCATATGCTTGTTTCAAAGATTAAGCCATGCATGTCTAAGTACATGC--CGAAATAAGGTGAAA
8. Nematostella_vectensis_cnidarian		NNCCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTACAAAC--TTATACACAGTGAAA
9. Nucula_sulcata_mollusc		-----------------------TCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTACAAAC--TTTCACATAGTGAAA
10. Saccoglossus_kowalevskii_hemichordate		-ACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTACGAGCCTCGGTACACGGTGAAA
11. Solaster_stimpsoni_echinoderm		---CATGCATGTATAAGTAAAAGC--TTTCA-CAAGCGAAA
12. Strongylocentrotus_purpuratus_echinoderm		-ACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTTAGTACAAGC--TTGTATCAAGCGAAA
13. Xenoturbella_bocki		---------------------AGTCATATGCTTGTCTCAAAGACTAAGCCATGCATGTCTAAGTACACGCCCTTGTACACGGTGAAA

Site # 1920 ● with ○ w/o Gaps

g) Write down how long your alignment is in the space below (scroll to the right, highlight a base in the last column, and look at the "Site #" box at the bottom of the screen). Go back to the **Alignment** menu, select "Align by ClustalW" again. This time, go to the options for "Multiple Alignment" and increase the "Gap Opening Penalty" from 15 to 100. This will tend to make gaps less frequent in the resulting alignment. Click "OK" and wait for the sequences to be aligned.

Q. *How long are the two alignments? How has increasing the Gap Opening Penalty altered the results of the alignment algorithm?*

. . . .

h) Now go back to the **Alignment** menu and select "Align by Muscle". We will simply use the default values here. Click "Compute" and wait for the sequences to be aligned.

Q. *Which of the two alignment algorithms was faster?*

. . . .

Unfortunately, neither MUSCLE nor CLUSTAL did a perfect job of aligning our sequences. This is almost always the case, especially for noncoding sequences. The way to fix this is to refine your alignment by hand. The aim here is to **maximise the number of columns for which you can confidently infer homology among sites**.

First we'll rearrange the sequences relative to one another without deleting anything at all. Then we'll delete columns without rearranging anything.

a) Start with the MUSCLE alignment and look at site 71.

b) Focus on *Solaster stimpsoni*, which should have the sequence "TTTCACA" here.

Q. *What could be changed to improve the alignment of this stretch of 7 bases?*

. . . .

c) Try to re-align these 7 bases to improve how well they fit with the rest of the sequences in the alignment. Remember that you are aiming to maximise the number of columns for which you can confidently infer homology.

d) Spend a few minutes working on this section of the alignment, trying to improve it. The main thing to note here is that we cannot completely trust alignments that a computer gives us.

e) There are some parts of the sequences that are just not possible to align reliably, and/or that are not particularly useful for phylogenetics (e.g., parts of the alignment for which we have sequences for only one or two taxa). Regions of genes can be impossible to align for many reasons, but most commonly it is because rates of evolution are so high, or the timescale of evolution so long, that any historical signal of homology has been erased. We should not use these sections of an alignment in an analysis, as all of our methods assume that each column is homologous. If we cannot be sure of homology, we need to delete those columns.

Q. *How long is your alignment before you delete any sites?*

. . . .

f) Now, delete uninformative and poorly aligned columns by highlighting the column or columns that you want to delete using the mouse, then clicking the 'X' at the top-right of the alignment explorer. Once you have finished deleting sites, you're done. Save your alignment as 'animals.mas' by clicking the disk symbol at the top left.

Q. *How long is your alignment after you have deleted all poorly aligned sites?*

. . . .

Part B: Model selection

Now we will identify the best-fitting model of nucleotide substitution for the data set. To ensure consistency, close your data set and open the file “animals_cut.fasta”. This sequence alignment has already had its ends trimmed for you.

A key purpose of substitution models is to account for multiple substitutions. Perhaps the most widely used is the General Time Reversible (GTR) model, which allows different rates for different substitution types. For example, it allows A↔G substitutions to occur at a different rate from C↔G substitutions. There are six substitution types in a time-reversible model (i.e., one in which forward and reverse substitutions occur at the same rate). Most phylogenetic methods only use time-reversible models. The model also allows the four nucleotides to have unequal frequencies.

Q. *Which phylogenetic methods use an explicit model of nucleotide substitution?*

. . . .

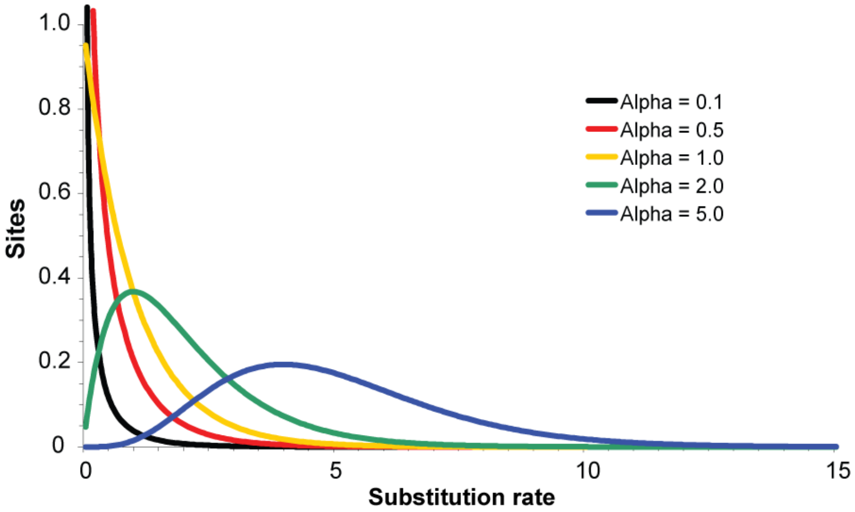
. . . .

Q. *What is the simplest model of nucleotide substitution?*

. . . .

We can also allow the rate of evolution to vary among the sites in the alignment. By doing this, we are assuming that some sites are selectively constrained, whereas others can change more freely without affecting the organism’s fitness.

Rate variation among sites is usually modelled using a gamma distribution, which can take a variety of shapes. The shape of the distribution is determined by a single parameter, alpha. When alpha is small, many sites evolve slowly but a small number of sites evolve quickly. When alpha is large, most sites evolve at about the same rate. For computational reasons, we use discrete rather than continuous gamma distributions. Usually 6 rate categories are used for the discrete gamma.



Conveniently, we can compare different models and select the best-fitting model in *MEGA*. There are several different criteria that can be used for model selection.

a) From the **Models** menu, select “Find Best DNA/Protein Models (ML)” and use the currently active data. This will bring up a box that contains a range of options. Accept the default settings and click on “Compute”.

MEGA is now computing the likelihood (probability of the data given the model) for 24 different substitution models. Have a look at the results of the analysis. The first column shows a list of the models. The second column shows how many parameters each model has. The third and fourth columns show the scores for two model-selection criteria, the Bayesian Information Criterion (BIC) and the corrected Akaike Information Criterion (AICc), respectively. For both of these criteria, lower scores indicate better-fitting models. For further details, have a look at the information below the table.

Q. *How are the BIC and AICc calculated? (look for explanations online)*

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

Q. *What are the best-fitting model(s) according to the BIC and AICc?*

. . . .

. . . .

The BIC and AICc have selected different best-fitting models. For the purposes of this practical, we will use the GTR+I+G model selected by the AICc. Normally we would need to make a decision regarding which criterion to use.

In the table, the number of free parameters in each model includes the branch lengths.

Q. *How many branches are in the tree? (hint: an unrooted tree of n tips has 2n-3 branches)*

. . . .

The most parameter-rich model (GTR+I+G), there are 10 free parameters (not including the branch lengths). The simplest model (Jukes-Cantor or JC model) has 0 free parameters.

Q. *What assumptions are made in the Jukes-Cantor model?*

. . . .

. . . .

. . . .

Q. *What are the 10 free parameters in the GTR+I+G model?*

. . . .

. . . .

. . . .

Q. *For the GTR+I+G model, what is the estimate of the shape parameter of the gamma distribution for this data set? What does this suggest about the degree of rate variation among sites?*

. . . .

. . . .

Q. *What is the estimate of the shape parameter of the gamma distribution for this data set when using the GTR+G model? Does this differ from the estimate when using the GTR+I+G model?*

. . . .

. . . .

Some researchers have argued that +I should not be used in combination with +G. This is because the two parameters interact strongly.

Q. *Why might +I and +G interact? Is this the case for the current data set?*

. . . .

. . . .

. . . .

Now that we have identified the best-fitting model from the range of available models, we can use this in subsequent phylogenetic analyses.

Part C: Phylogenetic analysis

Now we can estimate some phylogenetic trees. We'll do this in two different ways: a distance-based method and maximum likelihood.

Distance-based methods infer the phylogeny using a matrix of pairwise genetic distances. They are usually very quick because the method does not need to search through tree-space. Instead, distance-based methods use an algorithm to reconstruct the tree based on the distance matrix. The most commonly used algorithm is called 'neighbour-joining', which is what we will use here.

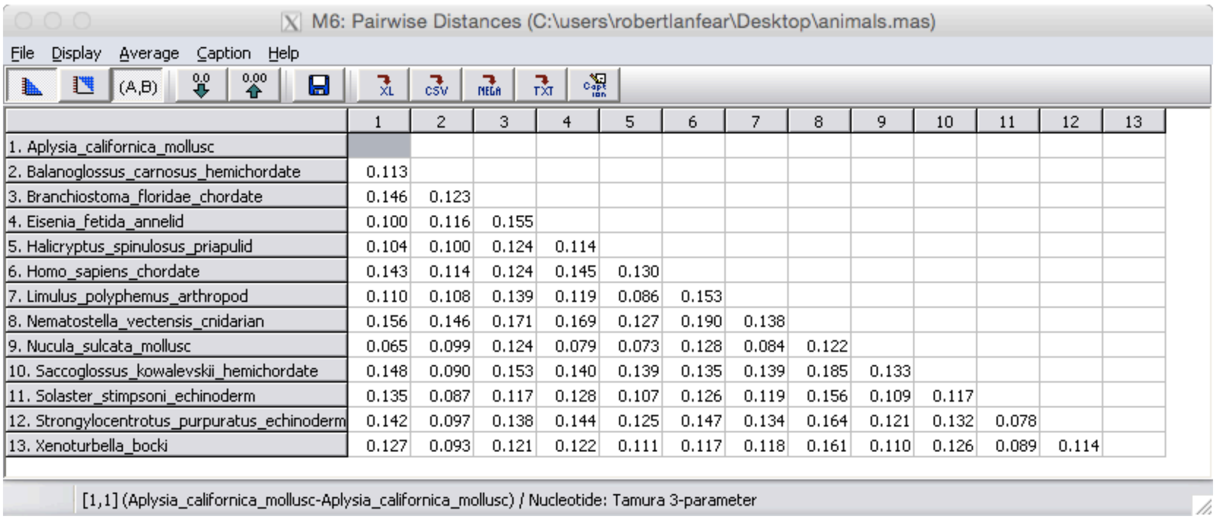
The pairwise distances can be calculated in a number of ways. The simplest method is to calculate the proportion of observed differences between each pair of sequences. This is known as the p-distance.

a) In *MEGA*, load the 'animals_cut.fasta' alignment file if you haven't already done so.

b) From the **Distance** menu, select "Compute Pairwise Distances" and use the currently active data. This will bring up a box that contains a range of options for the neighbour-joining analysis.

c) Accept the default parameters, and click 'Compute'.

Have a look at the resulting matrix of pairwise distances, which should look something like the screenshot shown below (but perhaps with different numbers).


M6: Pairwise Distances (C:\users\robertlanfear\Desktop\animals.mas)
File
Display
Average
Caption
Help
(A,B)
0.0
0.00
XL
CSV
MEGA
TXT
Caption

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Aplysia_californica_mollusc													
2. Balanoglossus_carnosus_hemichordate	0.113												
3. Branchiostoma_floridae_chordate	0.146	0.123											
4. Eisenia_fetida_annelid	0.100	0.116	0.155										
5. Halicryptus_spinulosus_priapulid	0.104	0.100	0.124	0.114									
6. Homo_sapiens_chordate	0.143	0.114	0.124	0.145	0.130								
7. Limulus_polyphemus_arthropod	0.110	0.108	0.139	0.119	0.086	0.153							
8. Nematostella_vectensis_cnidarian	0.156	0.146	0.171	0.169	0.127	0.190	0.138						
9. Nucula_sulcata_mollusc	0.065	0.099	0.124	0.079	0.073	0.128	0.084	0.122					
10. Saccoglossus_kowalevskii_hemichordate	0.148	0.090	0.153	0.140	0.139	0.135	0.139	0.185	0.133				
11. Solaster_stimpsoni_echinoderm	0.135	0.087	0.117	0.128	0.107	0.126	0.119	0.156	0.109	0.117			
12. Strongylocentrotus_purpuratus_echinoderm	0.142	0.097	0.138	0.144	0.125	0.147	0.134	0.164	0.121	0.132	0.078		
13. Xenoturbella_bocki	0.127	0.093	0.121	0.122	0.111	0.117	0.118	0.161	0.110	0.126	0.089	0.114	

[1,1] (Aplysia_californica_mollusc-Aplysia_californica_mollusc) / Nucleotide: Tamura 3-parameter

Q. *What do these numbers represent, and in what units are they given?*

. . . .

. . . .

Q. *What is the smallest genetic distance between* Xenoturbella *and another taxon?*

. . . .

d) From the **Phylogeny** menu, select “Construct Neighbour-Joining Tree” and use the currently active data. This will bring up a box that contains a range of options for the neighbour-joining analysis.

e) Check that you have the following options selected:
Test of Phylogeny: Bootstrap method
No. of Bootstrap Replications: 100
Model/Method: p-distance
Substitution to Include: d: Transitions + Transversions
Rates among Sites: Uniform rates

f) Click on “Compute” to start the neighbour-joining analysis.

g) The estimate of the phylogeny will appear in a new window. Reroot the tree using the *Nematostella vectensis* (the cnidarian) as the outgroup taxon. To do this, highlight the branch leading to *N. vectensis*, then click the icon on the left-hand side of the window that shows a phylogenetic tree with a green arrow pointing towards it.

Q. *There is a scale bar shown below the tree. What does this measure?*

. . . .

. . . .

Q. *The numbers on the branches are bootstrap support values. What do they represent?*

. . . .

. . . .

In the last analysis, we calculated the pairwise genetic distances using the observed mutational differences between sequences (p-distance). This can be problematic because it doesn’t account for the possibility of multiple mutations occurring at the same site, which would lead to an underestimate of the amount of evolutionary change that has actually occurred. We can correct for multiple hits by using a model of nucleotide substitution.

h) Without closing the window with your first tree, go back to the **Phylogeny** menu in the main *MEGA* window, select “Construct Neighbour-Joining Tree” again.

i) This time, we want to use the model of nucleotide substitution that chose in the previous part of this practical exercise. This should be the K2+G model.
Model/Method: Kimua 2-parameter model
Substitution to Include: d: Transitions + Transversions
Rates among Sites: Gamma Distributed (G)
Gamma Parameter: 0.32

Note that the value for the Gamma Parameter was estimated when we performed the model selection analysis in the previous part of this practical exercise.

j) The estimate of the phylogeny will appear in a new window. Check that the tree is rooted with *Nematostella vectensis*. If it isn’t, reroot it appropriately.

Q. *Does this estimate of the phylogeny differ from the previous estimate based on p-distances? If so, why do you think this is the case?*

. . . .

. . . .

. . . .

. . . .

. . . .

Q. *Which tree do you think is more trustworthy, and why?*

. . . .

. . . .

. . . .

. . . .

. . . .

Now, compare the more trustworthy phylogenetic tree with the one on the second page of this practical. If our alignments were perfect and our models of molecular evolution were accurate, you would have recovered the relationships in that tree for those taxa (obviously we weren’t sure *a priori* where *Xenoturbella* would be placed).

Q. *For each of the following major groups, and ignoring* Xenoturbella *for now, state whether the group is monophyletic in your inferred phylogeny.*

Protostomes: . . . .

Deuterostomes: . . . .

Chordata: . . . .

Echinodermata: . . . .

Q. *Is it (or would it be) a problem for your analysis if you didn't infer all of these groups to be monophyletic? What might it mean about your analysis? Can you trust the biological accuracy of your results?*

. . . .

. . . .

. . . .

. . . .

. . . .

Q. *Describe the position of* Xenoturbella *in your estimate of the phylogeny. Include details such as the group in which* Xenoturbella *is placed, and the amount of bootstrap support for this placement.*

. . . .

. . . .

. . . .

. . . .

. . . .

Q. *According to your analysis, is* Xenoturbella *a mollusc?*

. . . .

. . . .

. . . .

Now we will try analysing the data set using maximum likelihood. This is a statistical method that was first applied to phylogenetic analysis in the 1970s and formalised in the early 1980s. In maximum-likelihood analysis, we aim for maximum-likelihood estimates of the parameters and search for the maximum-likelihood tree. Like distance-based methods, maximum likelihood uses an explicit model of nucleotide substitution.

a) Go back to the main *MEGA* window. From the **Phylogeny** menu, select “Construct Maximum Likelihood Tree” and use the currently active data. This will bring up a box that contains a range of options for the maximum-likelihood analysis.

b) Check that you have the following options selected:
Test of Phylogeny: Bootstrap method
No. of Bootstrap Replications: 100
Substitutions Type: Nucleotide
Model/Method: Kimura 2-parameter model
Rates among Sites: Gamma Distributed (G)
No of Discrete Gamma Categories: 6
Gaps/Missing Data Treatment: Use all sites

c) Click “Compute” to start the maximum-likelihood phylogenetic analysis. See how long it takes to calculate bootstrap support from 100 replicates.

d) The estimate of the phylogeny will appear in a new window. Check that the tree is rooted with *Nematostella vectensis*. If not, reroot the tree appropriately.

Q. *Did the analysis take much longer than the distance-based analyses? Why do you think that this is the case?*

. . . .

. . . .

. . . .

. . . .

. . . .

Q. *For each of the following major groups, and ignoring* Xenoturbella *for now, state whether the group is monophyletic in your estimate of the phylogeny.*

Protostomes: . . . .

Deuterostomes: . . . .

Chordata: . . . .

Echinodermata: . . . .

Q. *You might have noticed that likelihoods are always given in log units, and that log likelihoods are negative. Why is this the case?*

. . . .

. . . .

. . . .

Q. *Describe the position of* Xenoturbella *in your phylogeny. Include details such as the taxon in which* Xenoturbella *is placed, and the amount of bootstrap support for this placement.*

. . . .

. . . .

. . . .

. . . .

Q. *Is* Xenoturbella *in the same place in this phylogeny as it is in the neighbour-joining trees?*

. . . .

. . . .

. . . .

Q. *Based on the bootstrap support values in your tree, what can you confidently say about the evolutionary placement of* Xenoturbella*?*

. . . .

. . . .

. . . .

. . . .

Q. *Of all the trees that you have estimated, which do you think is the most reliable? Why?*

. . . .

. . . .

. . . .

. . . .