

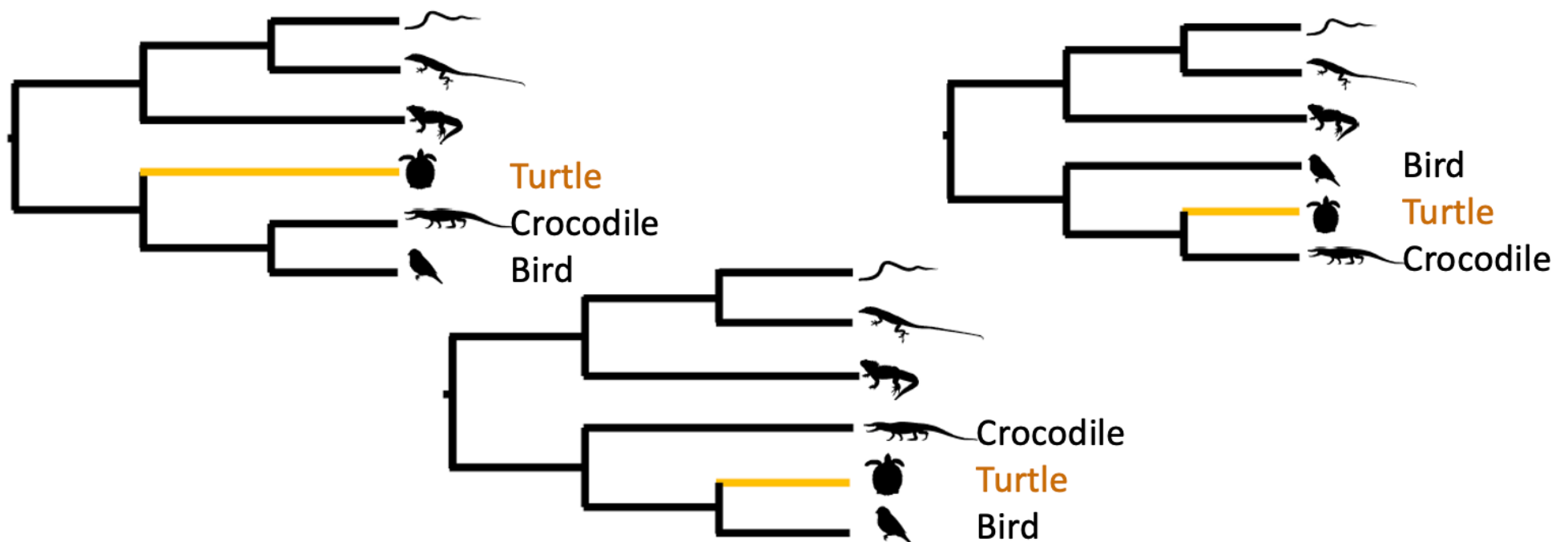
---

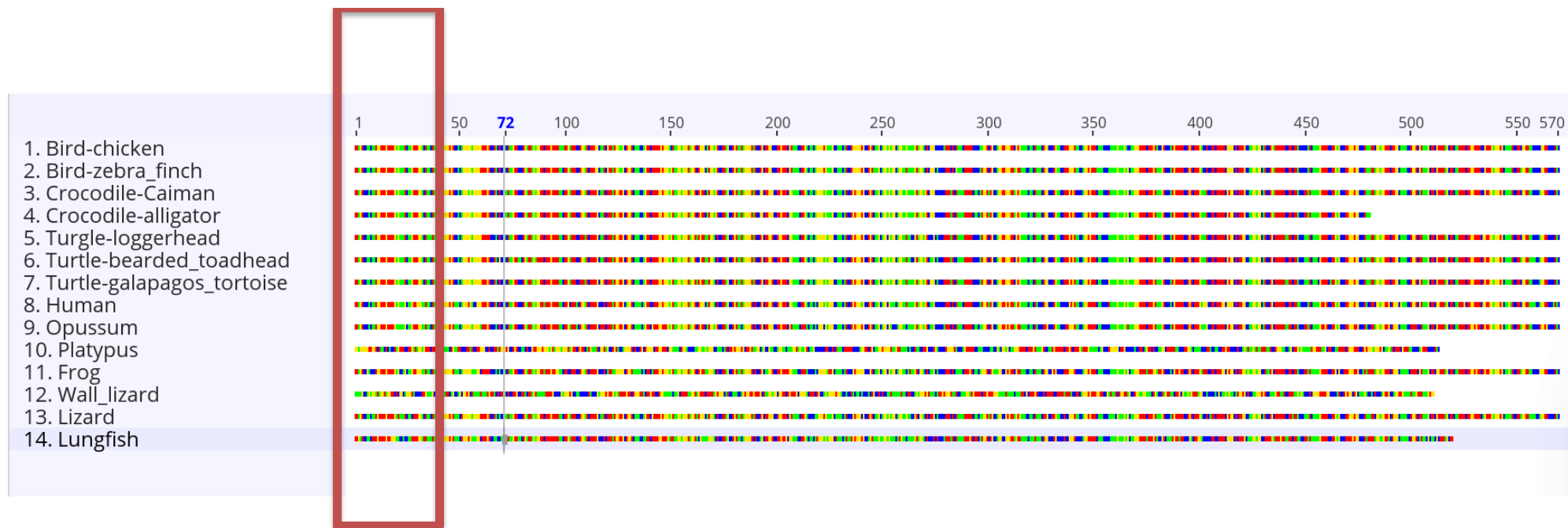
# Alignment

---

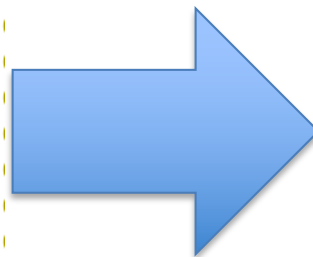
# Where do turtles fit in the tree of life?

- Sequence the same gene from all taxa
- Align the gene sequences to each other
- Estimate a phylogeny





ATGACTTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACTTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACCTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACCTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACTTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACTTCTCGGAAGAAAGTGTTACTGAAA  
 ATGACTTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACCTCTAGGAAGAAAGTGTTGCTGAAG  
 ATGACCTCTAGGAAGAAAGTTTTTCTGAAA  
 GTGGGGAAGACATCACTCATGAACCAAGTAT  
 ATGACGTCCAGGAAGAAAGTGCTGCTGAAG  
 TTTGCTGTGTATGATGGACACGCTGATCC  
 ATGACTTCTAGAAAGAAAGTATTACTCAAAA  
 AGTACTTCAAGAAAGAAAGGTTCTCTAAAA



ATGACTTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACTTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACCTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACCTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACTTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACTTCTCGGAAGAAAGTGTTACTGAAA  
 ATGACTTCTAGGAAGAAAGTGTTACTGAAA  
 ATGACCTCTAGGAAGAAAGTGTTGCTGAAG  
 ATGACCTCTAGGAAGAAAGTTTTCTGAAA  
 ATGACGTCCAGGAAGAAAGTGCTGCTGAAG  
 TTTGCTGTGTATGATGGACACGCTGATCC  
 ATGACTTCTAGAAAGAAAGTATTACTCAAAA  
 AGTACTTCAAGAAAGAAAGGTTCTCTAAAA

		1,220		1,230		1,240		1,250
Consensus	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T A G G C	C T A C T A G C C		
↗ Balaena mysticetus	A A A A A -	C C C T C C G A	G T G A -	T T A A A A G C	C C T A G G C	C C A C T A G C C		
↗ Balaenoptera acutorostrata	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	C C A C T A G C C		
↗ Caperea marginata	A A A A A -	C C C T C C G A	G T G A -	T T A A A A G C	C C T A G A C	T C A C T A G C C		
↗ Cephalorhynchus eutropia	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T T A G G C	C T A C C A G C C		
↗ Delphinapterus leucas	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	C T A C T G G C C		
↗ Delphinus delphis	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T T A G G C	C T A C T A G C C		
↗ Eschrichtius robustus	A A A A A -	C C C T C C G A	G T G A T	T T A A A A C C	C C T A G G C	T C A C T A G C C		
↗ Eubalaena australis	A A A A A -	C C C T C C G A	G T G A -	T T A A A A G C	C C T A G G C	C C A C C A G C C		
↗ Hippopotamus amphibius	A A G A A -	T C C T C C G A	G T G A -	T A A A A A T C	C T A G A C	T C A C C A G T C		
↗ Hyperoodon ampullatus	A A A A A -	A C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	T T A C C A G C C		
↗ Inia geoffrensis	G A A A A -	A C C T C C G A	G T G A T	T A T A A G C	C C T A G G C	C T A C T A G C C		
↗ Kogia breviceps	A A A A C -	C C C T C C G A	G T G A -	T T A G A G C	C C T A G G C	C T A C T A G C C		
↗ Lipotes vexillifer	A A A A A -	T C C T C C G A	G T G A -	T T A A A G C	C C T A G G C	C T A C C A G C C		
↗ Megaptera novaeangliae	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T T A G G C	C C A C T A G C C		
↗ Mesoplodon peruvianus	A A A A A -	A C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	C T A C C A G C C		
↗ Orcinus orca	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T T A G G C	C T A C C A G C C		
↗ Phocoena phocoena	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	C T A C T A G C C		
↗ Physeter catodon	A A A A A C	C C C T C C G A	G T G A -	T T A A A - C	C C T A G G C	C T A C C A G C C		
↗ Platanista gangetica	A A A A A -	A C C T C C G A	G T G A -	T T A A A G C	C C T A G G C	C T A C C A G C C		
↗ Pontoporia blainvillei	A A A A A -	C C C T C C G A	G T G A -	T T A A A G C	C C T A G G C	C C A C T A G C C		
↗ Pseudorca crassidens	A A A A A -	C C C T C C G A	G T G A -	T C A A A A C	T T A G G C	C T A C T A G C C		
↗ Steno bredanensis	G A A A A -	C C C T C C G A	G T G A -	T T A A A A C	T T A G G C	C T A C C A G C C		
↗ Ziphius cavirostris	A A A A A -	A C C T C C G A	A T G A -	T T A A A A C	T T A G G C	C T A C C A G C C		

A homologous site

(one that was inherited from a common ancestor of all the sequences in the alignment)

**The point of alignment is to maximise the number of sites (or columns)  
for which you can confidently infer homology**

		1,220		1,230		1,240		1,250
Consensus	A A A A A	C C C T C C G A	G T G A	T T A A A A C C T A G G C	C T A C T A G C C			
Balaena mysticetus	A A A A A	C C C T C C G A	G T G A	T T A A A G C C T A G G C	C C A C T A G C C			
Balaenoptera acutorostrata	A A A A A	C C C T C C G A	G T G A	T T A A A A C C T A G G C	C C A C T A G C C			
Caperea marginata	A A A A A	C C C T C C G A	G T G A	T T A A A G C C T A G A C	T C A C T A G C C			
Cephalorhynchus eutropia	A A A A A	C C C T C C G A	G T G A	T T A A A A C T T A G G C	C T A C C A G C C			
Delphinapterus leucas	A A A A A	C C C T C C G A	G T G A	T T A A A A C C T A G G C	C T A C T G G C C			
Delphinus delphis	A A A A A	C C C T C C G A	G T G A	T T A A A A C T T A G G C	C T A C T A G C C			
Eschrichtius robustus	A A A A A	C C C T C C G A	G T G A	T T A A A A C C T A G G C	T C A C T A G C C			
Eubalaena australis	A A A A A	C C C T C C G A	G T G A	T T A A A G C C T A G G C	C C A C C A G C C			
Hippopotamus amphibius	A A G A A	T C C T C C G A	G T G A	T A A A A A T C T A G A C	T C A C C A G T C			
Hyperoodon ampullatus	A A A A A	A C C T C C G A	G T G A	T T A A A A C C T A G G C	T T A C C A G C C			
Inia geoffrensis	G A A A A	A C C T C C G A	G T G A	T A T A A G C C T A G G C	C T A C T A G C C			
Kogia breviceps	A A A A C	C C C T C C G A	G T G A	T T A G A G C C T A G G C	C T A C T A G C C			
Lipotes vexillifer	A A A A A	T C C T C C G A	G T G A	T T A A A G C C T A G G C	C T A C C A G C C			
Megaptera novaeangliae	A A A A A	C C C T C C G A	G T G A	T T A A A A C T T A G G C	C C A C T A G C C			
Mesoplodon peruvianus	A A A A A	A C C T C C G A	G T G A	T T A A A A C C T A G G C	C T A C C A G C C			
Orcinus orca	A A A A A	C C C T C C G A	G T G A	T T A A A A C T T A G G C	C T A C C A G C C			
Phocoena phocoena	A A A A A	C C C T C C G A	G T G A	T T A A A A C C T A G G C	C T A C T A G C C			
Physeter catodon	A A A A A	C C C T C C G A	G T G A	T T A A A - C C T A G G C	C T A C C A G C C			
Platanista gangetica	A A A A A	A C C T C C G A	G T G A	T T A A A G C C T A G G C	C T A C C A G C C			
Pontoporia blainvillei	A A A A A	C C C T C C G A	G T G A	T T A A A G C C T A G G C	C C A C T A G C C			
Pseudorca crassidens	A A A A A	C C C T C C G A	G T G A	T C A A A A C T T A G G C	C T A C T A G C C			
Steno bredanensis	G A A A A	C C C T C C G A	G T G A	T T A A A A C T T A G G C	C T A C C A G C C			
Ziphius cavirostris	A A A A A	A C C T C C G A	A T G A	T T A A A A C T T A G G C	C T A C C A G C C			

Informative for topology  
and amount of evolution

Not informative for inferring the tree topology  
Informative for inferring amount of evolution  
(i.e. branch lengths, times, evolutionary rates,  
substitution model parameters)

**Both of these kinds of sites are useful in phylogenetics**

		1,220	1,230	1,240	1,250
Consensus	A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C C T A C T A G C C				
Balaena mysticetus	A A A A A - C C C T C C G A G T G A - T T A A A A G C C T A G G C C C A C T A G C C				
Balaenoptera acutorostrata	A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C C C A C T A G C C				
Caperea marginata	A A A A A - C C C T C C G A G T G A - T T A A A A G C C T A G A C T C A C T A G C C				
Cephalorhynchus eutropia	A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C C A G C C				
Delphinapterus leucas	A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C C T A C T G G C C				
Delphinus delphis	A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C T A G C C				
Eschrichtius robustus	A A A A A - C C C T C C G A G T G A T T T A A A A C C T A G G C T C A C T A G C C				
Eubalaena australis	A A A A A - C C C T C C G A G T G A - T T A A A A G C C T A G G C C C A C C A G C C				
Hippopotamus amphibius	A A G A A - T C C T C C G A G T G A - T A A A A T C T A G A C T C A C C A G T C				
Hyperoodon ampullatus	A A A A A - A C C T C C G A G T G A - T T A A A A C C T A G G C T T A C C A G C C				
Inia geoffrensis	G A A A A - A C C T C C G A G T G A T T A T A A G C C T A G G C C T A C T A G C C				
Kogia breviceps	A A A A C - C C C T C C G A G T G A - T T A G A G C C T A G G C C T A C T A G C C				
Lipotes vexillifer	A A A A A - T C C T C C G A G T G A - T T A A A A G C C T A G G C C T A C C A G C C				
Megaptera novaeangliae	A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C C A C T A G C C				
Mesoplodon peruvianus	A A A A A - A C C T C C G A G T G A - T T A A A A C C T A G G C C T A C C A G C C				
Orcinus orca	A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C C A G C C				
Phocoena phocoena	A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C C T A C T A G C C				
Physeter catodon	A A A A A C C C C T C C G A G T G A - T T A A A - C C T A G G C C T A C C A G C C				
Platanista gangetica	A A A A A - A C C T C C G A G T G A - T T A A A A G C C T A G G C C T A C C A G C C				
Pontoporia blainvillei	A A A A A - C C C T C C G A G T G A - T T A A A A G C C T A G G C C C A C T A G C C				
Pseudorca crassidens	A A A A A - C C C T C C G A G T G A - T C A A A A C T T A G G C C T A C T A G C C				
Steno bredanensis	G A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C C A G C C				
Ziphius cavirostris	A A A A A - A C C T C C G A A T G A - T T A A A A C T T A G G C C T A C C A G C C				



An Indel



Probably a deletion

(A site where an insertion or a deletion has happened)

Indels are informative, but most ML and Bayesian phylogenetics don't use them sensibly



Spp1	A		Spp1	A		Spp1	A		Spp1	A		Spp1	A
Spp2	A		Spp2	A		Spp2	A		Spp2	A		Spp2	A
Spp3	A		Spp3	A	+	Spp3	A	+	Spp3	A	+	Spp3	A
Spp4	–	=	Spp4	T		Spp4	C		Spp4	G		Spp4	A
Spp5	A		Spp5	A		Spp5	A		Spp5	A		Spp5	A

$$\ln L_{-} = \ln L_T + \ln L_C + \ln L_G + \ln L_A$$

Spp1	A		Spp1	A
Spp2	A		Spp2	A
Spp3	A	=	Spp3	A
Spp4	–		Spp4	N
Spp5	A		Spp5	A

$$\ln L_{-} = \ln L_N$$

**Most ML and Bayesian methods treat indels as missing data (N's)**

**Indels are informative, but most ML and Bayesian phylogenetics don't use them sensibly**

		1,220		1,230		1,240		1,250
Consensus	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T A G G C	C T A C T A G C C		
↗ Balaena mysticetus	A A A A A -	C C C T C C G A	G T G A -	T T A A A A G C	C C T A G G C	C C A C T A G C C		
↗ Balaenoptera acutorostrata	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	C C A C T A G C C		
↗ Caperea marginata	A A A A A -	C C C T C C G A	G T G A -	T T A A A A G C	C C T A G A C	T C A C T A G C C		
↗ Cephalorhynchus eutropia	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T T A G G C	C T A C C A G C C		
↗ Delphinapterus leucas	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	C T A C T G G C C		
↗ Delphinus delphis	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T T A G G C	C T A C T A G C C		
↗ Eschrichtius robustus	A A A A A -	C C C T C C G A	G T G A T	T T A A A A C C	C C T A G G C	T C A C T A G C C		
↗ Eubalaena australis	A A A A A -	C C C T C C G A	G T G A -	T T A A A A G C	C C T A G G C	C C A C C A G C C		
↗ Hippopotamus amphibius	A A G A A -	T C C T C C G A	G T G A -	T A A A A A T C	C T A G A C	T C A C C A G T C		
↗ Hyperoodon ampullatus	A A A A A -	A C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	T T A C C A G C C		
↗ Inia geoffrensis	G A A A A -	A C C T C C G A	G T G A T	T A T A A G C	C C T A G G C	C T A C T A G C C		
↗ Kogia breviceps	A A A A C -	C C C T C C G A	G T G A -	T T A G A G C	C C T A G G C	C T A C T A G C C		
↗ Lipotes vexillifer	A A A A A -	T C C T C C G A	G T G A -	T T A A A G C	C C T A G G C	C T A C C A G C C		
↗ Megaptera novaeangliae	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T T A G G C	C C A C T A G C C		
↗ Mesoplodon peruvianus	A A A A A -	A C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	C T A C C A G C C		
↗ Orcinus orca	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	T T A G G C	C T A C C A G C C		
↗ Phocoena phocoena	A A A A A -	C C C T C C G A	G T G A -	T T A A A A C C	C C T A G G C	C T A C T A G C C		
↗ Physeter catodon	A A A A A C	C C C T C C G A	G T G A -	T T A A A - C	C C T A G G C	C T A C C A G C C		
↗ Platanista gangetica	A A A A A -	A C C T C C G A	G T G A -	T T A A A G C	C C T A G G C	C T A C C A G C C		
↗ Pontoporia blainvillei	A A A A A -	C C C T C C G A	G T G A -	T T A A A G C	C C T A G G C	C C A C T A G C C		
↗ Pseudorca crassidens	A A A A A -	C C C T C C G A	G T G A -	T C A A A A C	T T A G G C	C T A C T A G C C		
↗ Steno bredanensis	G A A A A -	C C C T C C G A	G T G A -	T T A A A A C	T T A G G C	C T A C C A G C C		
↗ Ziphius cavirostris	A A A A A -	A C C T C C G A	A T G A -	T T A A A A C	T T A G G C	C T A C C A G C C		

A homologous site

(one that was inherited from a common ancestor of all the species in the alignment)

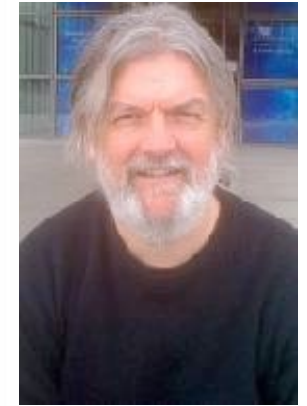
The point of alignment is to maximise the number of sites (or columns)  
for which you can confidently infer homology

(while bearing in mind the limitations of the downstream methods...)



# A practical approach to alignment

1. Align sequences with automated methods (e.g. MAFFT)
2. Check alignments by eye (e.g. Geneious)
3. Fix major errors
  - Remove alignments without useful information
  - Remove sequences with uncertain homology
4. Go back to step 1 until you find no more major errors
5. Optional – fix minor errors by hand
  - Can depend on the question and amount of data
  - E.g. realign certain regions
  - E.g. delete poorly aligned columns
6. Use automated methods to clean final alignments



## CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice

Julie D. Thompson, Desmond G. Higgins<sup>+</sup> and Toby J. Gibson<sup>\*</sup>

European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany



## BMC Bioinformatics



Software

**Open Access**

**MUSCLE: a multiple sequence alignment method with reduced time and space complexity**

Robert C. Edgar<sup>\*</sup>

**ALWAYS check and edit your alignments by eye!**

## Robert Edgar's Blog

Just another WordPress.com weblog

HOME



[← Multiple protein alignment is a dead field](#)

[Fishing for significance →](#)

SEARCH IT!

**An unemployed gentleman scholar**

Posted on [May 4, 2010](#) | [7 Comments](#)

RECENT ENTRIES



The key question is, can an [automated] multiple alignment represent homology between letters accurately enough to enable robust inferences to be made by downstream tools? ...

The answer is, **probably not...**

# The Aim

Aplysia_californica	CAGGCGCGCAA	C	TTACCCACTCCCG	G	CAC	G	GGGGAG				
Balanoglossus_carnosus	CAGGCGCGCAA	A	TTACCCA	T	TCCCGACAC	G	GGGGAG				
Branchiostoma_floridae	CAGGCGCGCAA	A	UUACCCACU	C	CCGAC	U	C	GGGGAG			
Eisenia_fetida	CAGGCGCGCAA	A	TTACCCA	A	TCCCGACAC	G	GGGGAG				
Halicryptus_spinulosus	CAGG	C	A	CGCAA	A	TTACCCACTCCCG	G	CAC	G	GGGGAG	
Homo_sapiens	CAGGCGCGCAA	A	TTACCCACTCCCG	A	C	C	G	GGGGAG			
Limulus_polyphemus	CAGGCGCGCAA	A	TTACCCACTCC	C	A	G	A	A	C	G	GGGGAG
Nematostella_vectensis	CAGGCGCGCAA	A	TA	CCCACTCCCG	G	CAC	G	GGGGAG			
Nucula_sulcata	CAGGCGCGCAA	A	TTACCCACTCC	T	G	G	CAC	G	GGGGAG		
Saccoglossus_kowalevskii	CAGGCGCGCAA	A	TTACCCA	T	TCCCGACAC	-	G	GGAG			
Solaster_stimpsoni	CAGGCGCGCAA	A	TTACCCACTCCCG	A	C	C	G	GGGGAG			
Strongylocentrotus_purpuratus	CAGGCGCGCAA	A	TTACCCACTC	T	CGACAC	G	GGGGAG				
Xenoturbella_bocki	CAGGCGCGCAA	A	TTACCCACTCCCG	A	C	C	G	GGGGAG			

A collection of alignments for which you are confident that every column is homologous

A A A - C C C T C C G A G T G A - T T A A A G C C T A G G C  
A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C  
A A A - C C C T C C G A G T G A - T T A A A G C C T A G A C  
A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C  
A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C  
A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C  
A A A - C C C T C C G A G T G A T T T A A A A C C T A G G C  
A A A - C C C T C C G A G T G A - T T A A A G C C T A G G C  
G A A - T C C T C C G A G T G A - T A A A A T C T A G A C  
A A A - A C C T C C G A G T G A - T T A A A A C C T A G G C  
A A A - A C C T C C G A G T G A T T A A G C C T A G G C  
A A C - C C C T C C G A G T G A - T T A G A G C C T A G G C  
A A A - T C C T C C G A G T G A - T T A A A G C C T A G G C  
A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C  
A A A - A C C T C C G A G T G A - T T A A A A C C T A G G C  
A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C  
A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C  
A A A C C C C T C C G A G T G A - T T A A A - C C T A G G C  
A A A - A C C T C C G A G T G A - T T A A A G C C T A G G C  
A A A - C C C T C C G A G T G A - T T A A A G C C T A G G C  
A A A - C C C T C C G A G T G A - T C A A A A C T T A G G C  
A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C  
A A A - A C C T C C G A A T G A - T T A A A A C T T A G G C

Questions?

**Download the workshop practicals here**

**<https://github.com/roblanf/Workshop-MIG/releases/latest>**

**Unzip, then open up the HTML file:**

**1\_Alignment/Alignment.html**