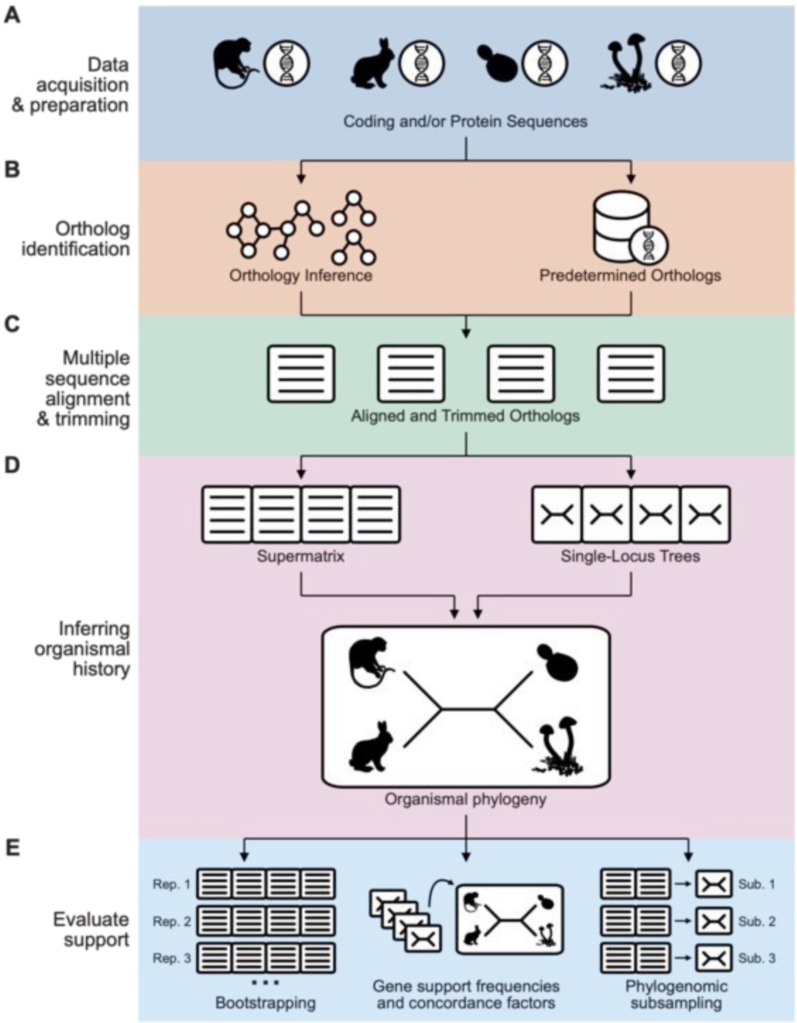
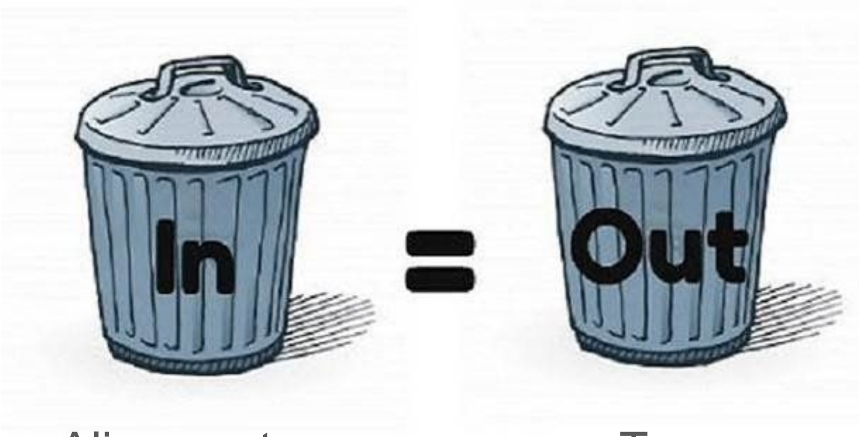
Multiple sequence alignments

Marina Marcet Houben
mmarcet@bsc.es


A
Data acquisition & preparation
Coding and/or Protein Sequences
B
Ortholog identification
Orthology Inference
Predetermined Orthologs
C
Multiple sequence alignment & trimming
Aligned and Trimmed Orthologs
D
Supermatrix
Single-Locus Trees
Inferring organismal history
Organismal phylogeny
E
Evaluate support
Rep. 1
Rep. 2
Rep. 3
Bootstrapping
Gene support frequencies and concordance factors
Sub. 1
Sub. 2
Sub. 3
Phylogenomic subsampling

Sequence alignments

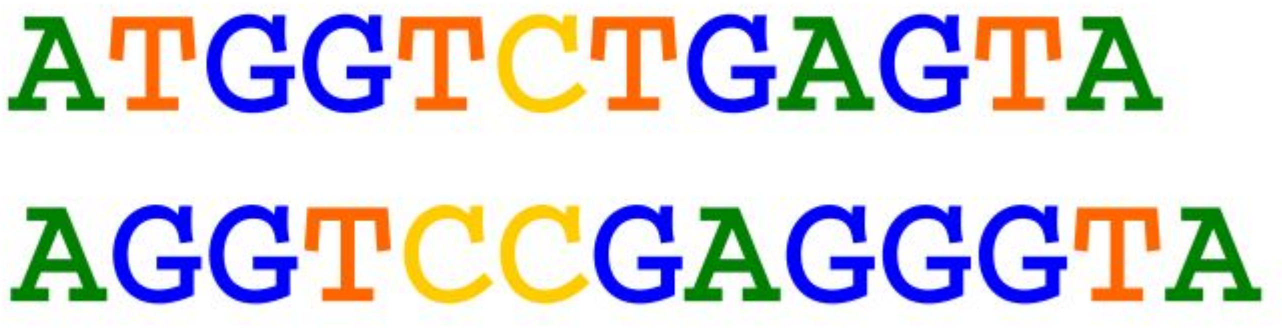
A sequence alignment is a way of arranging two or more sequences to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.



Alignment

Tree

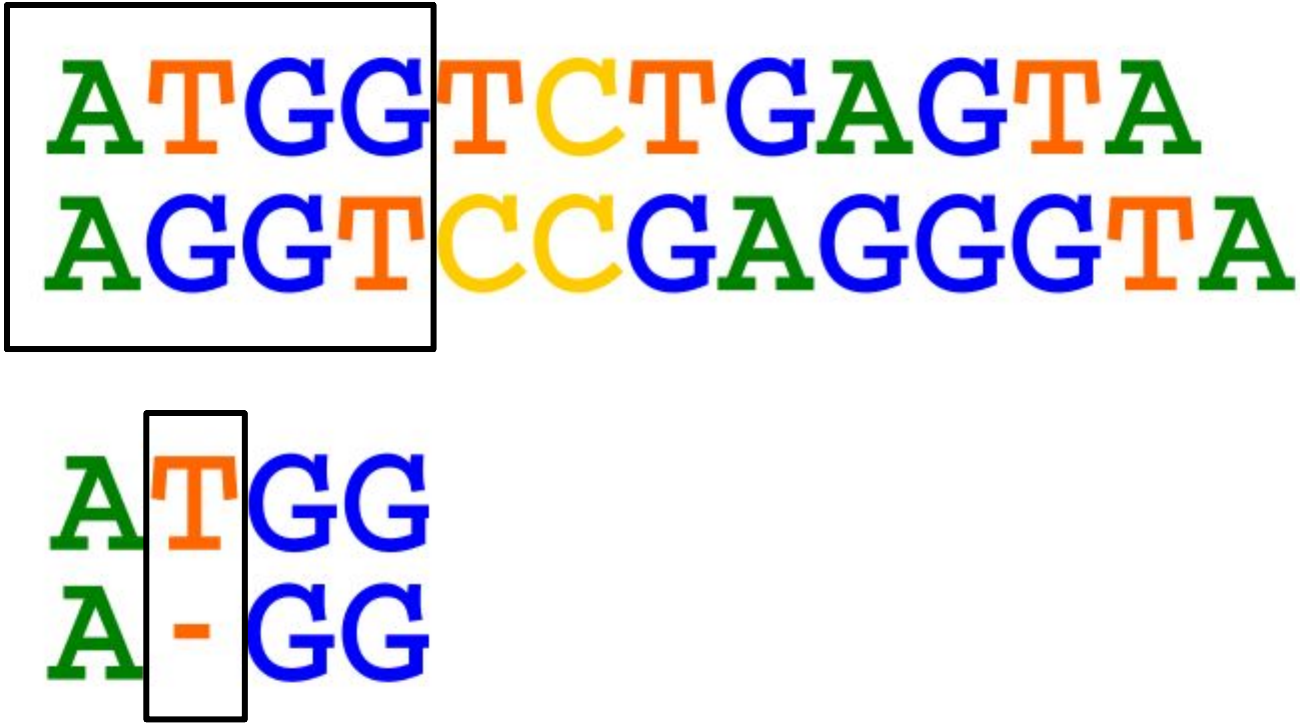
The basics



ATGGTCTGAGTA

AGGTCCGAGGGTA

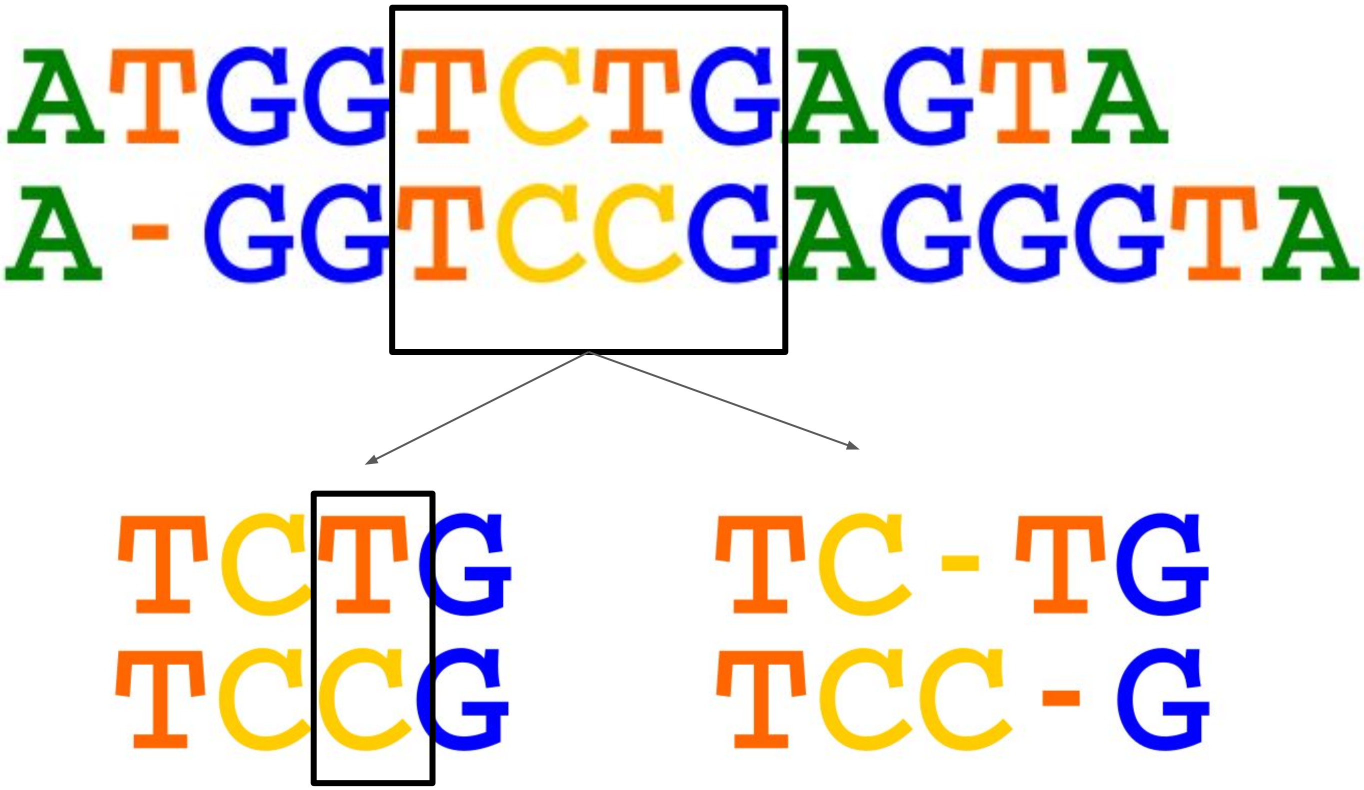
Can you align those two sequences?



```
ATGGTCTGAGTA
AGGTCCGAGGGTA

ATGG
A-GG
```

Gap: feature that shows an insertion / deletion



Point mutation

Which is the right one?

Alignment score

There are different ways with which to align two sequences, to formalize it we use an **alignment score.**

Example:

Match = +1
Mismatch = -1
Gap = -1

```
TCTG
TCCG
```

Alignment score = +2

```
TC-TG
TCC-G
```

Alignment score = +1

But this is not feasible at large scale, we cannot always search for all possible sequence alignments to score them

How can we generalize this when having longer sequences: **Dynamic programming**

How much money can you gain?

5	1	9	1
8	1	1	1
1	1	1	1
1	1	12	4

↘ Collect the money in the square

↓ → Lose one coin and don't collect the money

We first draw a second, empty board and start filling it with the amount of money we would win when moving into it. When multiple paths are available you keep the one that gives the most money

Money board

5	1	9	1
8	1	1	1
1	1	1	1
1	1	12	4

Dynamic programming board

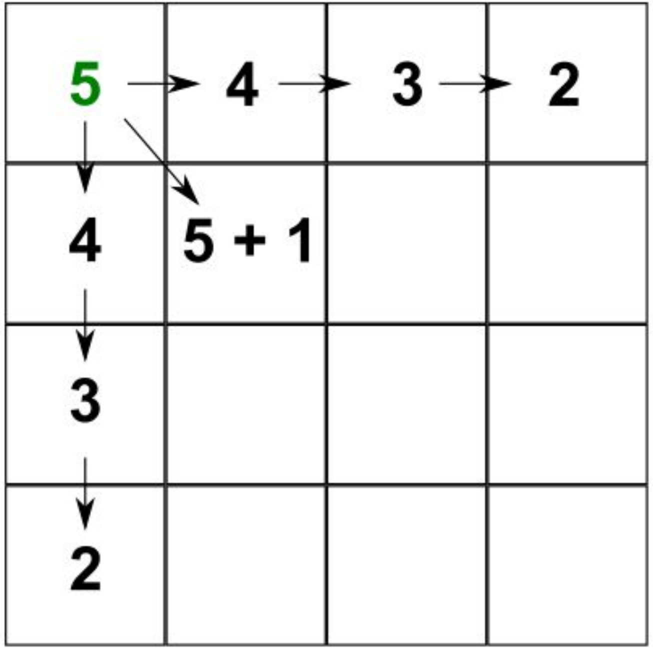
5 →	4 →	3 →	2
↓ 4			
↓ 3			
↓ 2			

The next empty square could be filled coming down (4-1), coming right (4-1) or coming in diagonal (5+1). This last option is the best and it’s the one we keep

Money board

5	1	9	1
8	1	1	1
1	1	1	1
1	1	12	4

Dynamic programming board



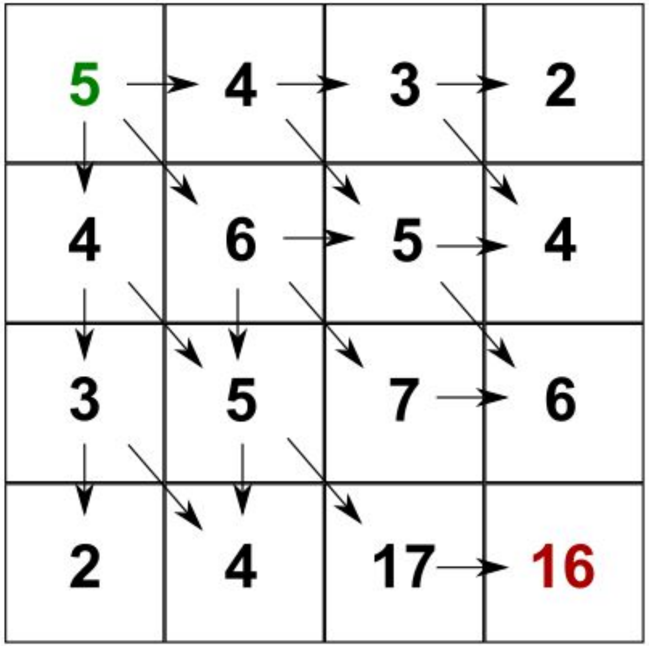
5	4	3	2
4	5 + 1		
3			
2			

Once you have filled the whole board you know the maximum amount of money you can win are 16 coins. You also see that sometimes there are multiple paths you can take with the same results.

Money board

5	1	9	1
8	1	1	1
1	1	1	1
1	1	12	4

Dynamic programming board



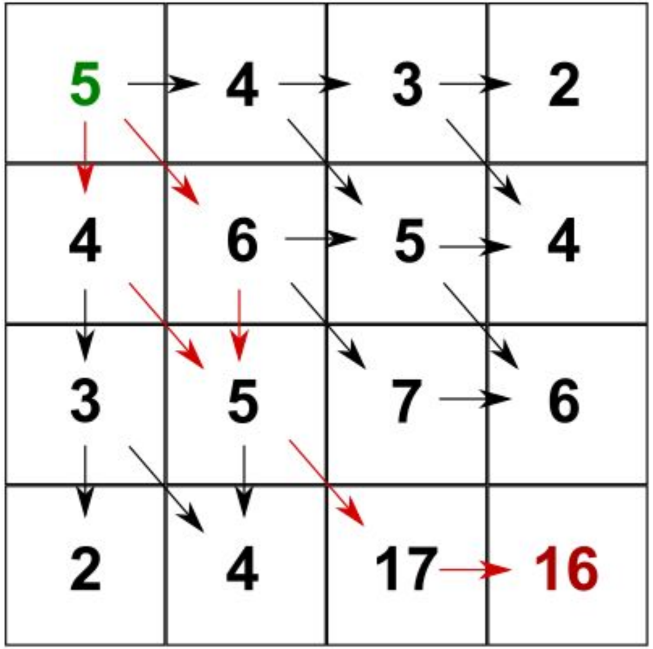
5	4	3	2
4	6	5	4
3	5	7	6
2	4	17	16

Now you can go back and see which path was the best one. You'll notice that there are actually two paths that resulted in the same amount of money gained.

Money board

5	1	9	1
8	1	1	1
1	1	1	1
1	1	12	4

Dynamic programming board

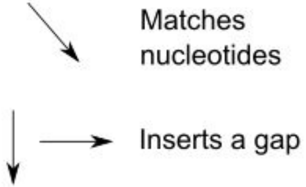


5	4	3	2
4	6	5	4
3	5	7	6
2	4	17	16

Now we apply the same to sequences

TCTG
TCCG

		T	C	T	G
	0				
T					
C					
C					
G					



Matches nucleotides

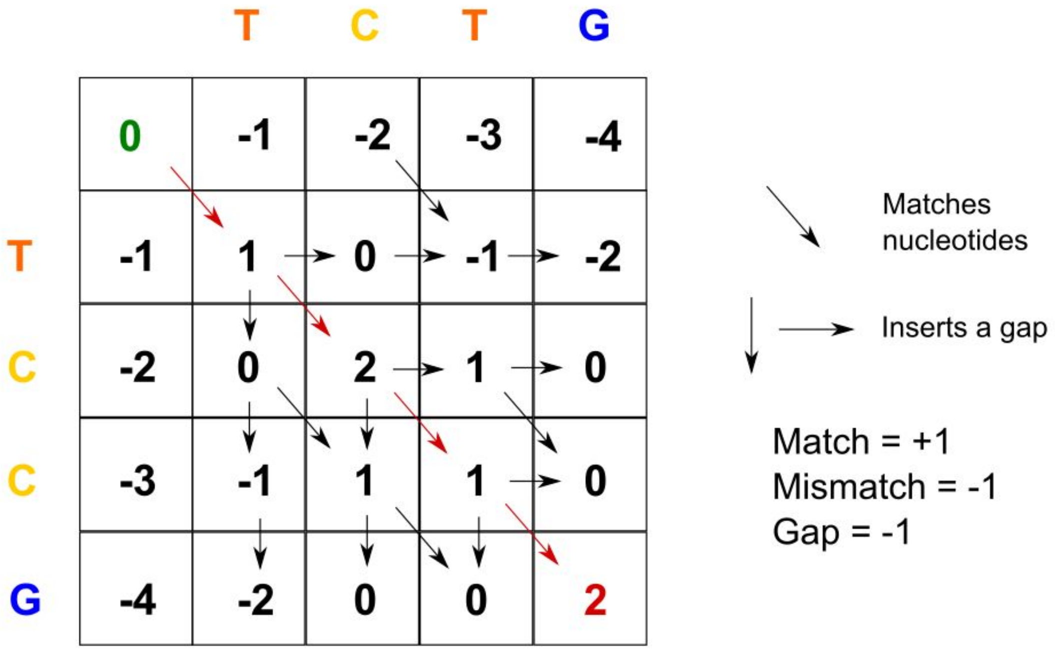
Inserts a gap

Match = +1
Missmatch = -1
Gap = -1

Now, instead of adding coins, when you move in diagonal you add a +1 when the residues match or -1 if the residues do not match. Inserting a gap also subtracts 1

Now we apply the same to sequences

TCTG
TCCG



		T	C	T	G
	0	-1	-2	-3	-4
T	-1	1	0	-1	-2
C	-2	0	2	1	0
C	-3	-1	1	1	0
G	-4	-2	0	0	2

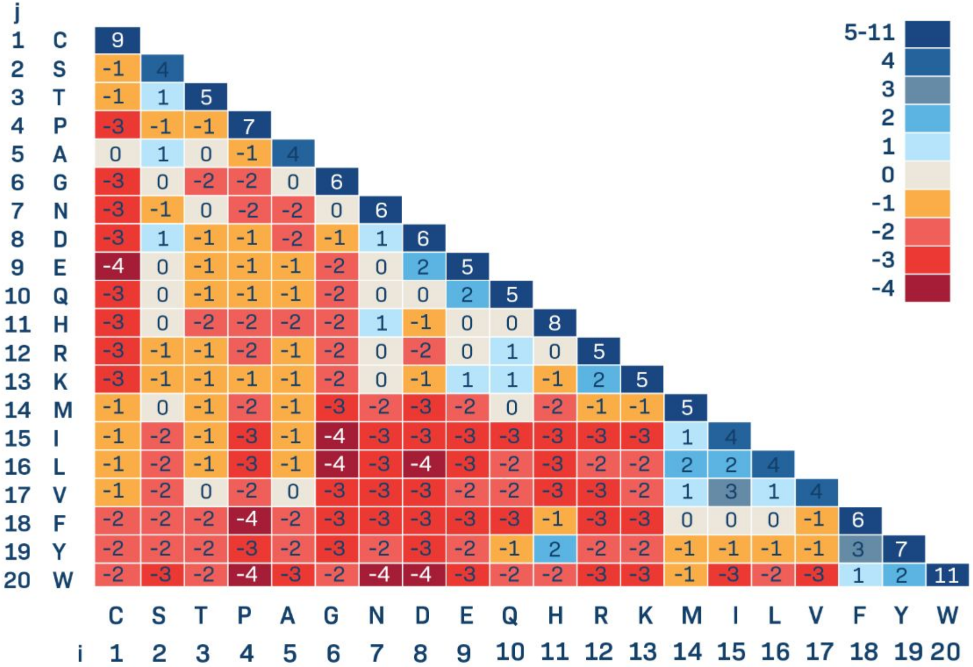
Match = +1
Mismatch = -1
Gap = -1

In this case the best path is the one which does not insert gaps.

The gap penalty and the mismatch penalty can be altered.

For protein alignments you usually have a substitution matrix instead of a fixed match / mismatch value

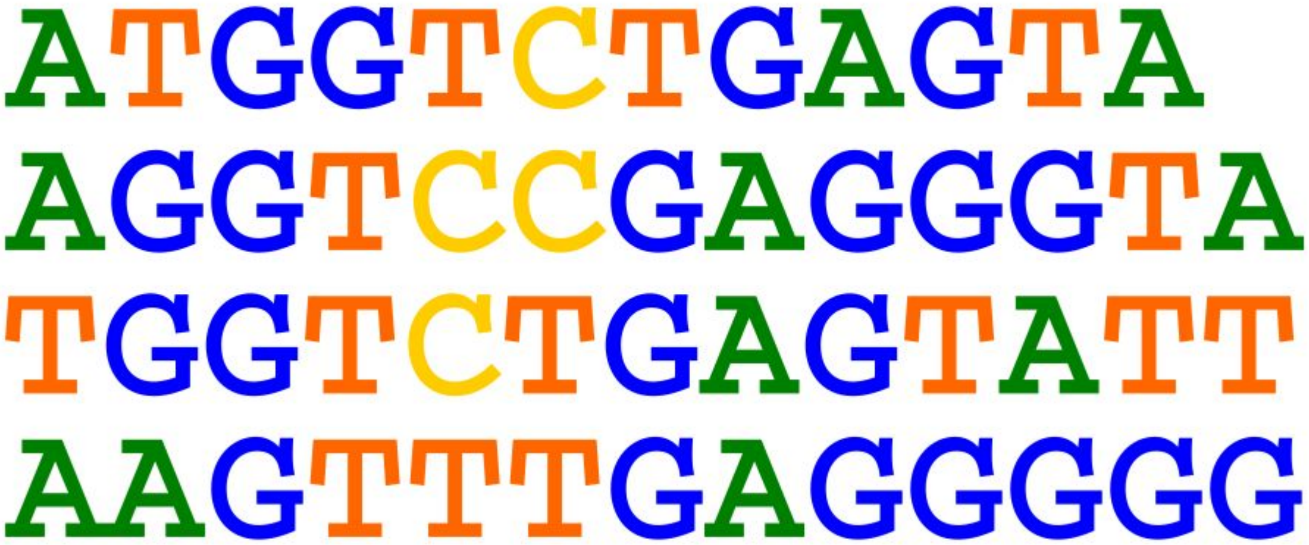
Blossum62 substitution matrix



j																					
1	C	9																			
2	S	-1	4																		
3	T	-1	1	5																	
4	P	-3	-1	-1	7																
5	A	0	1	0	-1	4															
6	G	-3	0	-2	-2	0	6														
7	N	-3	-1	0	-2	-2	0	6													
8	D	-3	1	-1	-1	-2	-1	1	6												
9	E	-4	0	-1	-1	-1	-2	0	2	5											
10	Q	-3	0	-1	-1	-1	-2	0	0	2	5										
11	H	-3	0	-2	-2	-2	-2	1	-1	0	0	8									
12	R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
13	K	-3	-1	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
14	M	-1	0	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
15	I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
16	L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
17	V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
18	F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
19	Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
20	W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

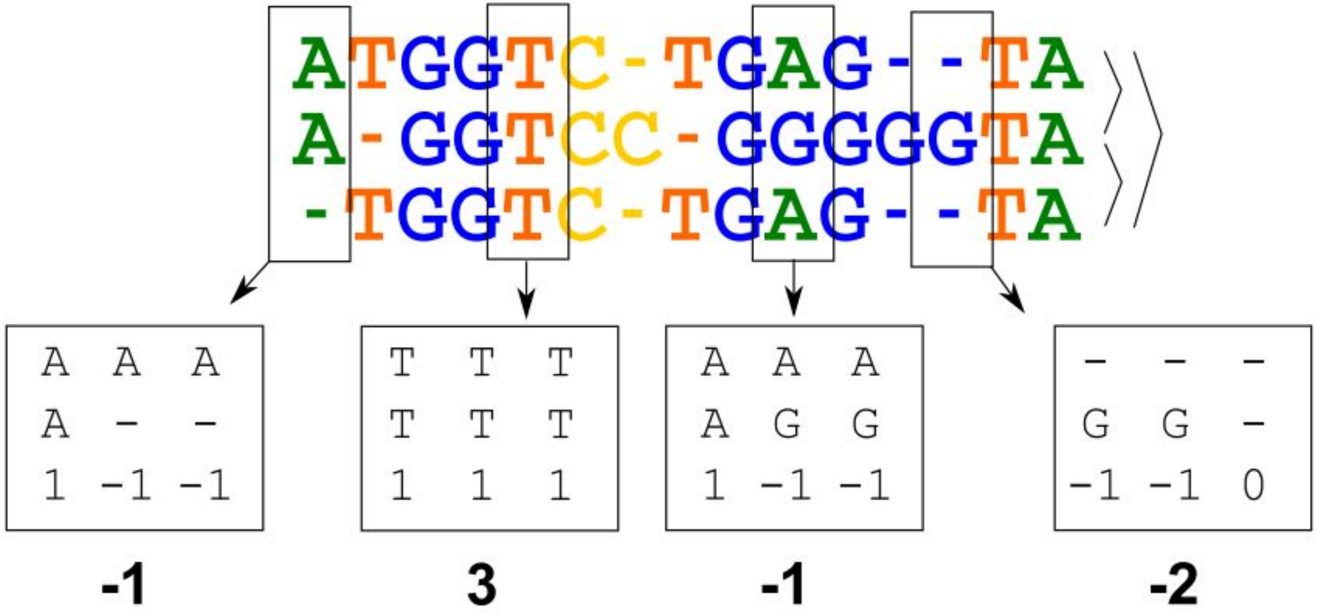
Calculated based on alignments of conserved protein regions without gaps. The score is based on the substitution probabilities between the different amino acids.

Multiple sequence alignments (MSA)



```
ATGGTCTGAGTA
AGGTCCGAGGGTA
TGGTCTGAGTATT
AAGTTTGAGGGGG
```

How to score a MSA? Sum of pairs



ATGGTC-TGAG--TA
A-GGTCC-GGGGGTA
-TGGTC-TGAG--TA

A A A
A – –
1 −1 −1

-1

T T T
T T T
1 1 1

3

A A A
A G G
1 −1 −1

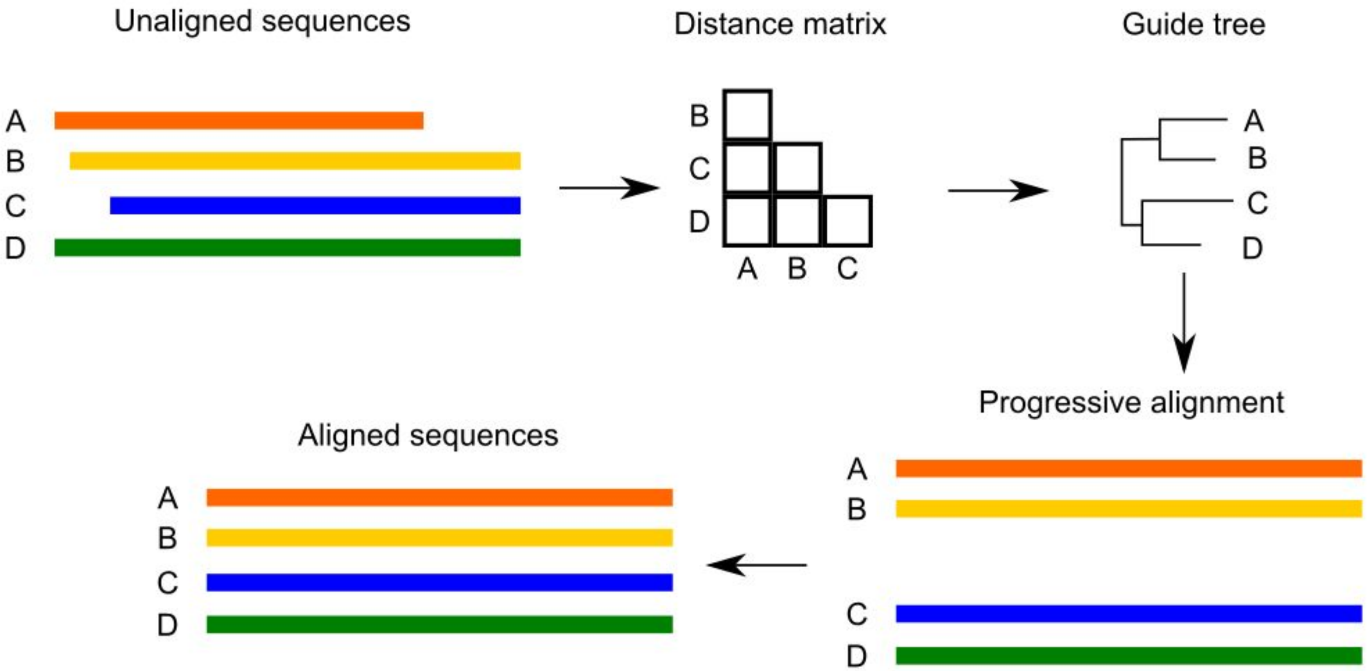
-1

– – –
G G –
−1 −1 0

-2

Sum of Pairs = -1 + (-1) + 3 + 3 + 3 + 3 + (-2) + (-1) + 3 + (-1) + 3 + (-2) + (-2) + 3 + 3 = 14

Multiple sequence alignment methods: Progressive MSA



Alignment methods: Progressive / Iterative

- They are widely used
- They are very fast
- Able to process large numbers of sequences
- They may not achieve the global optimum since errors done at the first alignment stages are kept through the alignment process
- This is improved by iterative alignment methods, where pairwise alignments are improved after having been reconstructed

Muscle5

FAMSA

Kalign

CLUSTAL

DIALIGN [home]

MAFFT version 7
Multiple alignment program for amino acid or nucleotide sequences

T COFFEE

ProbCons

How do I chose the best MSA method?

PROTEINS
RNA
T-COFFEE
SIMPLE MSA
DNA
BAli---Phy
MUMMALS
MASS
Multiple Alignment by Secondary Structures
CLUSTAL
CLUSTAL
ESPript 3.0
EMBL-EBI
MUSCLE
PROBCONS
Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences
PRANK
MAFFT version 7
Multiple alignment program for amino acid or nucleotide sequences

You may be limited by the data

- Which kind of data are you working with? (DNA, proteins, codons, RNA, etc)
- Do you expect to be dealing with isoforms?
- Which percentage of identity do you expect your data to have?
- How many sequences are you trying to align?
- Which computational resources do you have at your disposal?

The alignment challenge

Given three datasets of 13 vertebrate sequences with different levels of average sequence identity. Test four different alignment programs and check what you think works best for each dataset.

Programs can be run on default parameters or you can play with the different parameters depending on your previous familiarity with the programs.

High average identity > 0.90
Medium average identity ~ 0.65
Low average identity ~ 0.4

Data can be found here:
https://drive.google.com/file/d/1uHabUIQJa4J5BXonZSiF0iiy2dtwOyJw/view?usp=sharing

The alignment challenge

- Step1: Build an alignment with each of the methods using the default parameters (MUSCLE, MAFFT, PRANK and PASTA)
- Step2: Visualize each alignment (Aliview, Seaview)
- Step3: Build phylogenetic trees for each alignment (Fasttree)
- Step4: Compare the topology obtained for each alignment+tree to the reference topology (you can visualize trees using this website https://phylo.io/)

Tips for running the different programs:

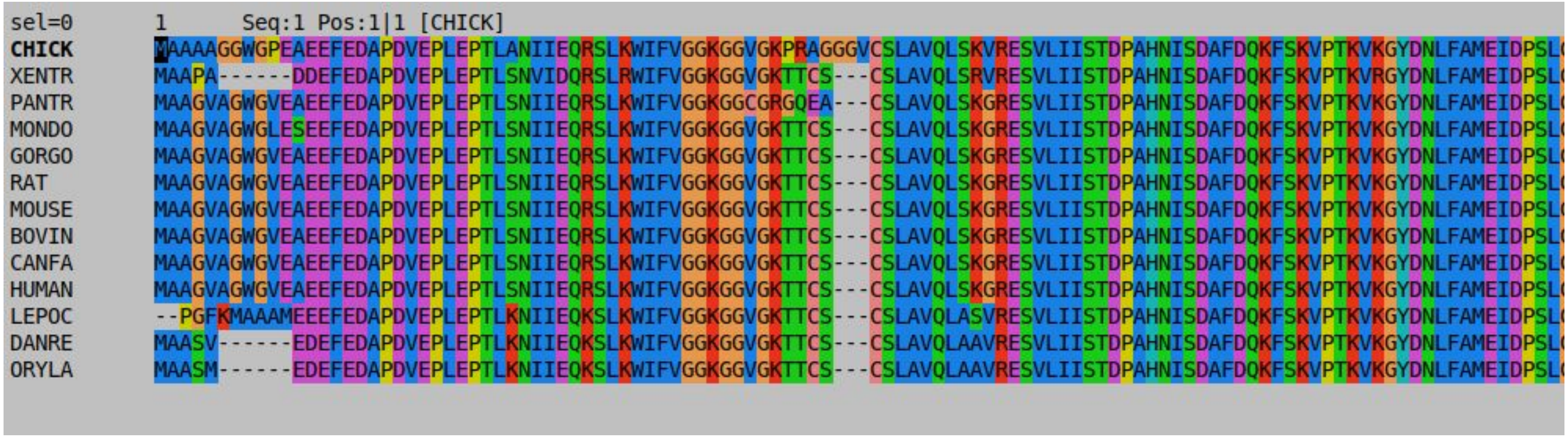
- `muscle5 -align NameSequenceFile -output NameOutputFile`
- `mafft NameSequenceFile >NameOutputFile`
- `prank -d=NameSequenceFile -o=NameOutputFile`
- `run_pasta.py -i NameSequenceFile -d protein -o outputFolder`

Where NameSequenceFile is the file containing the unaligned sequences. And the NameOutputFile is the name of the file where you want the alignment to be printed. Pasta outputs a lot of files, therefore the output is redirected to a folder (outputFolder)

!! Note that Prank and Pasta will modify your output names. For PRANK your output will be renamed to NameOutputFile.best.fas and for PASTA your output can be found in NameOutputFile.marker001.NameSequenceFile.aln

Visualize alignments

You can visualize alignments using the program seaview or aliview, both installed in the guacamole server. Executing either will open a graphical interface where you can upload and view your alignments



```
sel=0     1      Seq:1 Pos:1|1 [CHICK]
CHICK     MAAAAGGWGPEAEEFEDAPDVEPLEPTLANIIEQRSLKWIFVGGKGGVGKPRAGGGVCSLAVQLSKVRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
XENTR     MAAPA------DDEFEDAPDVEPLEPTLSNVIDQRSLRWIFVGGKGGVGKTTCS---CSLAVQLSRVRESVLIISTDPAHNISDAFDQKFSKVPTKVRGYDNLFAMEIDPSL
PANTR     MAAGVAGWGVEAEEFEDAPDVEPLEPTLSNIIEQRSLKWIFVGGKGGCGRGQEA---CSLAVQLSKGRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
MONDO     MAAGVAGWGLESEEFEDAPDVEPLEPTLSNIIEQRSLKWIFVGGKGGVGKTTCS---CSLAVQLSKGRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
GORGO     MAAGVAGWGVEAEEFEDAPDVEPLEPTLSNIIEQRSLKWIFVGGKGGVGKTTCS---CSLAVQLSKGRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
RAT       MAAGVAGWGVEAEEFEDAPDVEPLEPTLSNIIEQRSLKWIFVGGKGGVGKTTCS---CSLAVQLSKGRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
MOUSE     MAAGVAGWGVEAEEFEDAPDVEPLEPTLSNIIEQRSLKWIFVGGKGGVGKTTCS---CSLAVQLSKGRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
BOVIN     MAAGVAGWGVEAEEFEDAPDVEPLEPTLSNIIEQRSLKWIFVGGKGGVGKTTCS---CSLAVQLSKGRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
CANFA     MAAGVAGWGVEAEEFEDAPDVEPLEPTLSNIIEQRSLKWIFVGGKGGVGKTTCS---CSLAVQLSKGRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
HUMAN     MAAGVAGWGVEAEEFEDAPDVEPLEPTLSNIIEQRSLKWIFVGGKGGVGKTTCS---CSLAVQLSKGRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
LEPOC     --PGFKMAAAMEEEFEDAPDVEPLEPTLKNIIEQKSLKWIFVGGKGGVGKTTCS---CSLAVQLASVRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
DANRE     MAASV------EDEFEDAPDVEPLEPTLKNIIEQKSLKWIFVGGKGGVGKTTCS---CSLAVQLAAVRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
ORYLA     MAASM------EDEFEDAPDVEPLEPTLKNIIEQKSLKWIFVGGKGGVGKTTCS---CSLAVQLAAVRESVLIISTDPAHNISDAFDQKFSKVPTKVKGYDNLFAMEIDPSL
```

Discussion on Alignments

When viewing the alignments pay attention to the following:

1.- Do you see differences in the quality of the alignments reconstructed using the three different datasets?

2.- Do the alignment programs give different alignments? (Note that the order of the sequences in the alignment is not relevant in this case)

3.- Can you identify any other trend that could differentiate the programs?

4.- What about the time it takes to run each program, could that be a problem with larger datasets?

Assess how well your MSA has been built using trees

One common way to assess the performance of MSA is by building a tree based on it and checking whether it keeps the same topology as the reference phylogeny.

We are going to use fasttree for that. Execute fasttree for each of your alignments.

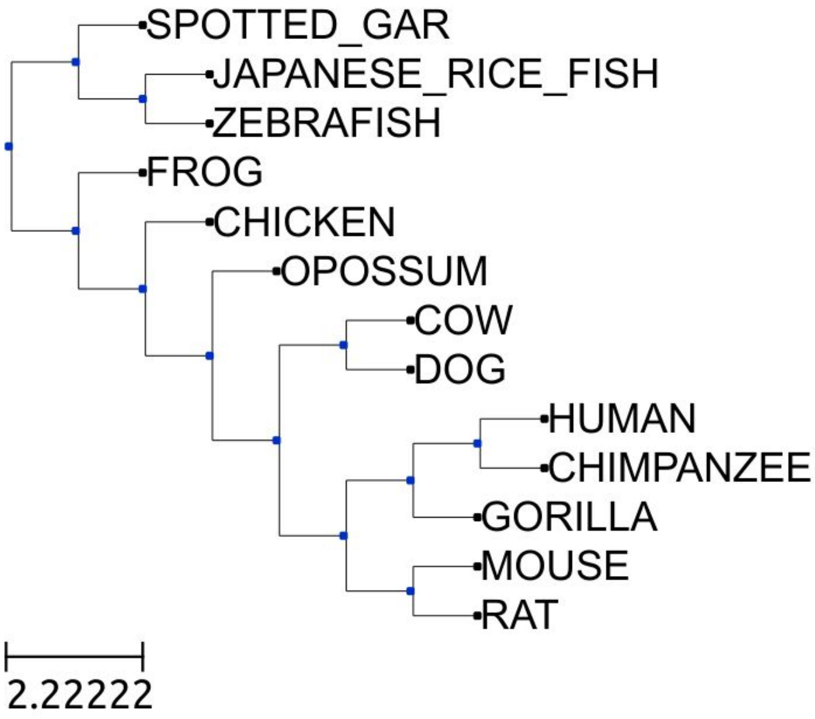
Tip: To execute fasttree you just need to run:

```
FastTree alignmentFile >OutputTree
```

You can also use a loop as long as all your alignments have the -alg termination:

```
for i in *.alg; do FastTree $i > ${i/.alg/.tre}; done
```

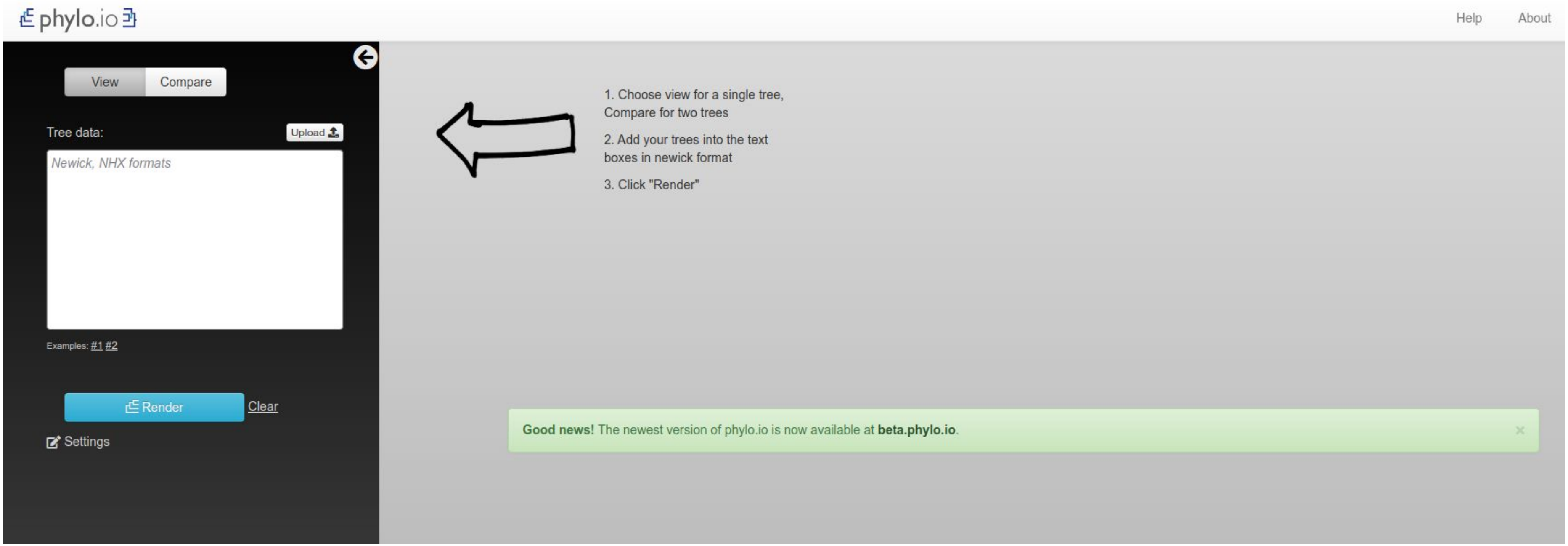
Reference Topology



In order to compare topologies you need to visualize your tree and then compare whether the species are grouped in the same way in the two trees.

Note that in order to effectively compare two trees you will need them to be oriented in the same way

Comparing trees with phylo.io



1. Choose view for a single tree, Compare for two trees

2. Add your trees into the text boxes in newick format

3. Click "Render"

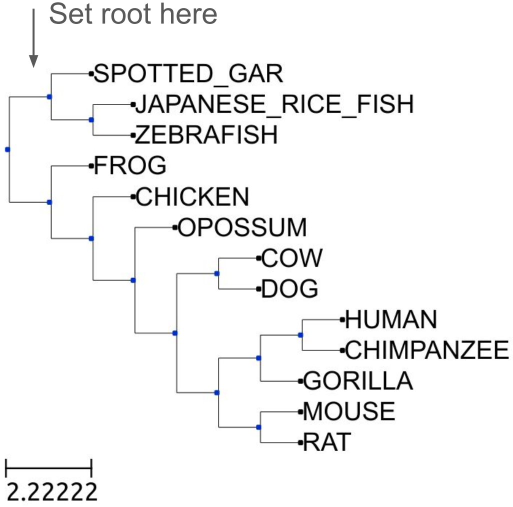
Good news! The newest version of phylo.io is now available at **beta.phylo.io**.

Best is to open this in the local browser, avoid using Chrome as some functions don’t seem to work

Comparing trees with phylo.io

Compare pairs of topologies with phylo.io

Note that both trees should be rooted at the point of divergence between fishes and other animals for the comparison to be meaningful



Set root here

Discussion on trees

- Did all trees for one dataset follow the same topology?
- Where did you find the most differences?
- How was the support for those node?
- Were there trees that were the same across methods but did still disagree with the species tree topology?
- Do you think this is a problem with the alignment programs or could there be biological reasons?

Share your results!

https://docs.google.com/spreadsheets/d/1aSrX__0vrfIk7Za8SHP9C0CfortqFI9nl-Nho0r1WjQ/edit?usp=sharing

Please, add the results of your observations in this excel sheet! Please, try not to overwrite someone else’s result