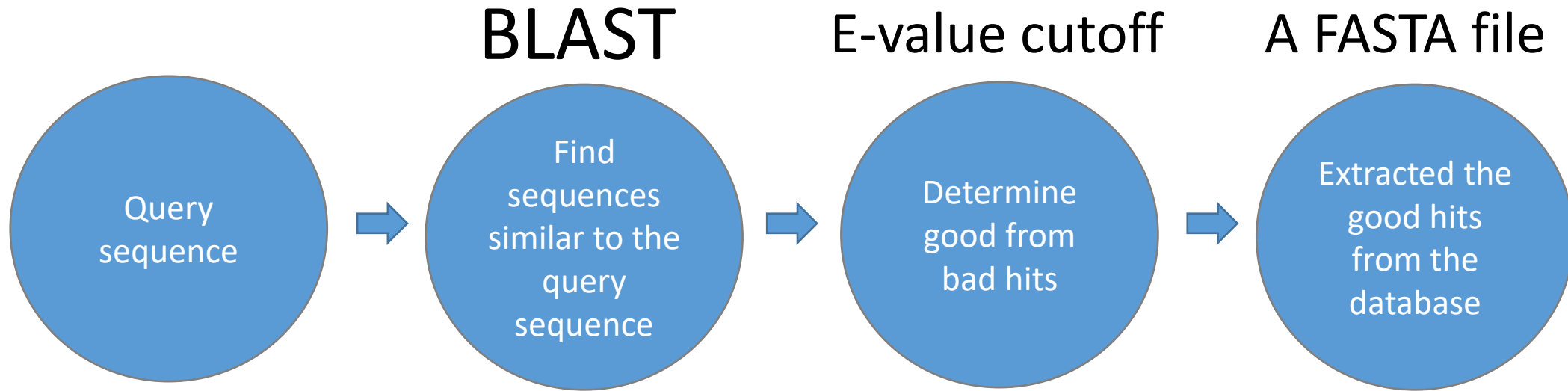


MULTIPLE SEQUENCE ALIGNMENTS

Line up of homologous positions





Why do we need so many sequences?

- Comparative studies, more information
Even if we are only interested in one particular sequence
- What other species has a similar sequence?
 - When did this gene appear?
 - Has it been lost in some species or taxonomic group?
 - When did potential gene duplications happen?



Why do we need to align our sequences?

UNALIGNED

ALIGNED

- Line up of homologous positions

How to align multiple sequences?



Multiple Alignment Algorithms

- ClustalW
- Muscle
- Mafft
- T-coffee
- And more
- <http://www.ebi.ac.uk/Tools/msa/> or Jalview or other



FASTA FILE

```
>SEQ_A
THIESRPSRL
>SEQ_B
TRIESRPSRL
>SEQ_C
HIESRRSR
>SEQ_D
TVIESKPSRL
>SEQ_E
SQHVESRQSR
```

Add the pairwise sequence identities to complete the table.

	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	90				
SEQ_C					
SEQ_D					
SEQ_E					

$$\text{Distance} = 1 - \frac{\text{pairwise sequence identity \%}}{100}$$



FASTA FILE

```
>SEQ_A
THIESRPSRL
>SEQ_B
TRIESRPSRL
>SEQ_C
HIESRRSR
>SEQ_D
TVIESKPSRL
>SEQ_E
SQHVESRQSR
```

$$\text{Distance} = 1 - \frac{\text{pairwise sequence identity \%}}{100}$$

Add the pairwise sequence identities to complete the table.

	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	90				
SEQ_C	87.5	75			
SEQ_D	80	80	62.5		
SEQ_E	66.67	55.56	75	44.44	

DISTANCE

	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B					
SEQ_C					
SEQ_D					
SEQ_E					



FASTA FILE

```
>SEQ_A
THIESRPSRL
>SEQ_B
TRIESRPSRL
>SEQ_C
HIESRRSR
>SEQ_D
TVIESKPSRL
>SEQ_E
SQHVESRQSR
```

$$\text{Distance} = 1 - \frac{\text{pairwise sequence identity \%}}{100}$$

Add the pairwise sequence identities to complete the table.

	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	90				
SEQ_C	87.5	75			
SEQ_D	80	80	62.5		
SEQ_E	66.67	55.56	75	44.44	

DISTANCE

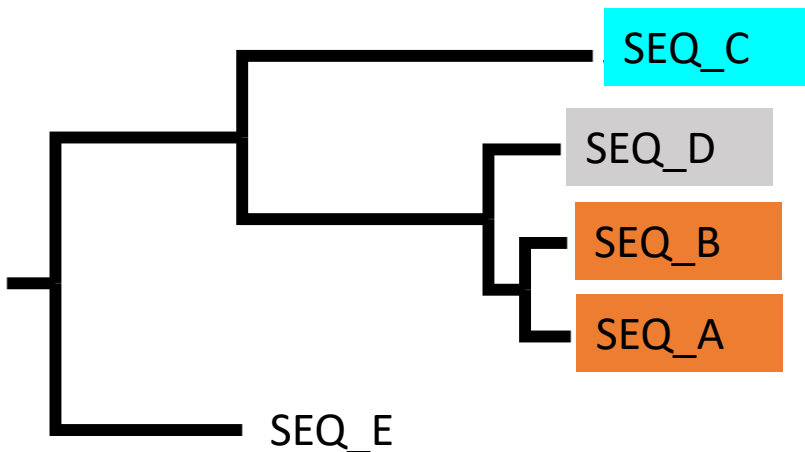
	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	0.1				
SEQ_C	0.125	0.25			
SEQ_D	0.2	0.2	0.375		
SEQ_E	0.3333	0.4444	0.25	0.5556	



FASTA FILE

```
>SEQ_A
THIESRPSRL
>SEQ_B
TRIESRPSRL
>SEQ_C
HIESRRSR
>SEQ_D
TVIESKPSRL
>SEQ_E
SQHVESRQSR
```

Based on the distance matrix, a distance tree is built.



Add the pairwise sequence identities (%) to complete the table.

	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	90				
SEQ_C	87.5	75			
SEQ_D	80	80	62.5		
SEQ_E	66.67	55.56	75	44.44	

DISTANCE MATRIX

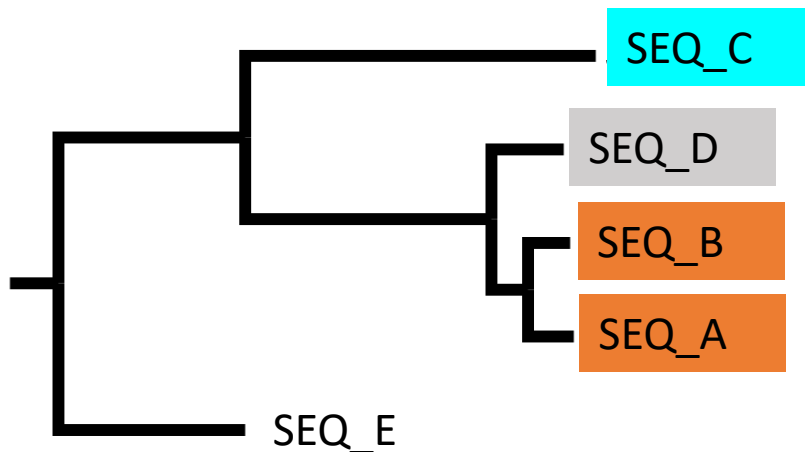
	Distance = $1 - \frac{\text{pairwise sequence identity \%}}{100}$				
	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	0.1				
SEQ_C	0.125	0.25			
SEQ_D	0.2	0.2	0.375		
SEQ_E	0.3333	0.4444	0.25	0.5556	



FASTA FILE

```
>SEQ_A
THIESRPSRL
>SEQ_B
TRIESRPSRL
>SEQ_C
HIESRRSR
>SEQ_D
TVIESKPSRL
>SEQ_E
SQHVESRQSR
```

Based on the distance matrix, a distance tree is built.



Add the pairwise sequence identities (%) to complete the table.

	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	90				
SEQ_C	87.5	75			
SEQ_D	80	80	62.5		
SEQ_E	66.67	55.56	75	44.44	

DISTANCE MATRIX

	$\text{Distance} = 1 - \frac{\text{pairwise sequence identity \%}}{100}$				
	SEQ_A	SEQ_B	SEQ_C	SEQ_D	SEQ_E
SEQ_B	0.1				
SEQ_C	0.125	0.25			
SEQ_D	0.2	0.2	0.375		
SEQ_E	0.3333	0.4444	0.25	0.5556	

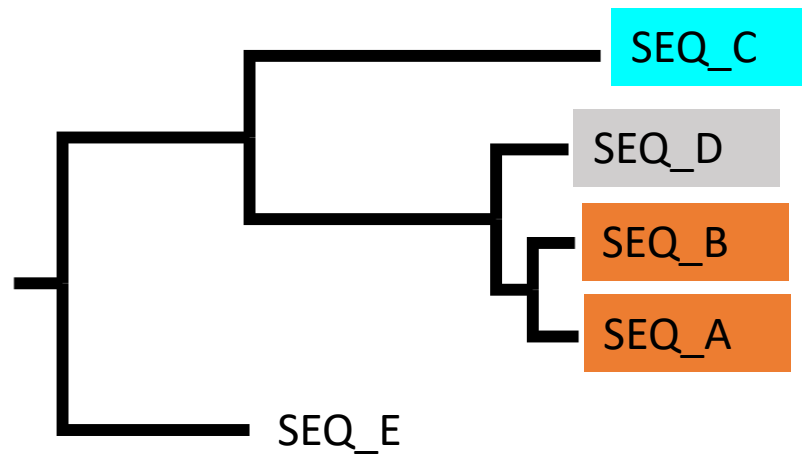
(We'll get back to tree building. For now, this is a distance tree from Jalview.)

The tree is rooted at mid-point.)



FASTA FILE

```
>SEQ_A  
THIESRPSRL  
>SEQ_B  
TRIESRPSRL  
>SEQ_C  
HIESRRSR  
>SEQ_D  
TVIESKPSRL  
>SEQ_E  
SQHVESRQSR
```

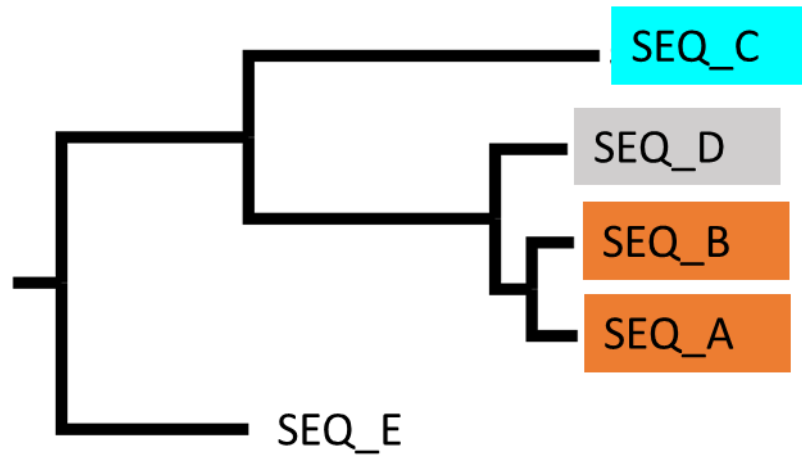


1. Start the progressive alignment with SEQ_A and SEQ_B:



FASTA FILE

```
>SEQ_A  
THIESRPSRL  
>SEQ_B  
TRIESRPSRL  
>SEQ_C  
HIESRRSR  
>SEQ_D  
TVIESKPSRL  
>SEQ_E  
SQHVESRQSR
```



1. Start the progressive alignment with SEQ_A and SEQ_B:

SEQ_A	T H I E S R P S R L
SEQ_B	T R I E S R P S R L

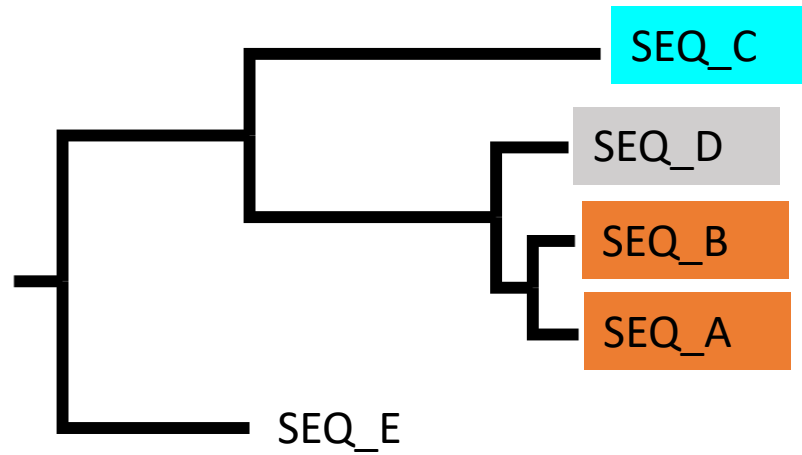
2. Add sequence D:

SEQ_A	T H I E S R P S R L
SEQ_B	T R I E S R P S R L
SEQ_D	T V I E S K P S R L



FASTA FILE

```
>SEQ_A  
THIESRPSRL  
>SEQ_B  
TRIESRPSRL  
>SEQ_C  
HIESRRSR  
>SEQ_D  
TVIESKPSRL  
>SEQ_E  
SQHVESRQSR
```



1. Start the progressive alignment with SEQ_A and SEQ_B:

SEQ_A	T	H	I	E	S	R	P	S	R	L
SEQ_B	T	R	I	E	S	R	P	S	R	L

2. Add sequence D:

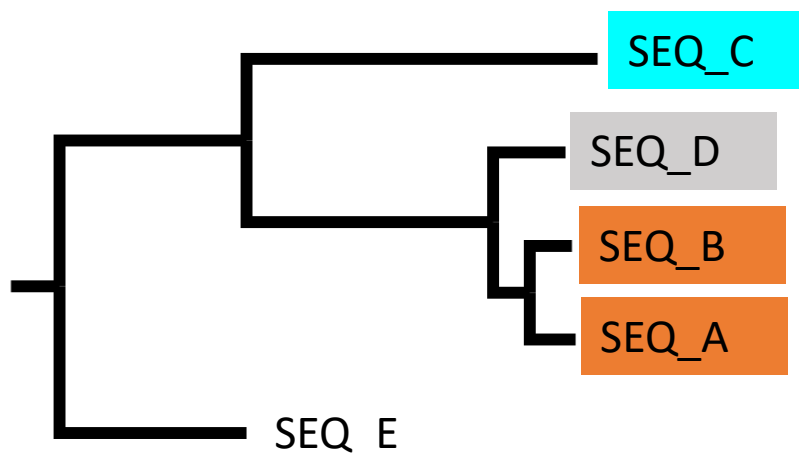
SEQ_A	T	H	I	E	S	R	P	S	R	L
SEQ_B	T	R	I	E	S	R	P	S	R	L
SEQ_D	T	V	I	E	S	K	P	S	R	L

3. And so on...



FASTA FILE

```
>SEQ_A
THIESRPSRL
>SEQ_B
TRIESRPSRL
>SEQ_C
HIESRRSR
>SEQ_D
TVIESKPSRL
>SEQ_E
SQHVESRQSR
```



1. Start the progressive alignment with SEQ_A and SEQ_B:

SEQ_A	T	H	I	E	S	R	P	S	R	L
SEQ_B	T	R	I	E	S	R	P	S	R	L

2. Add sequence D:

SEQ_A	T	H	I	E	S	R	P	S	R	L
SEQ_B	T	R	I	E	S	R	P	S	R	L
SEQ_D	T	V	I	E	S	K	P	S	R	L

3. And so on...



**A PROGRESSIVE
MULTIPLE
SEQUENCE
ALIGNMENT!**

SEQ_C	-	-	H	I	E	S	R	R	S	R	-
SEQ_A	-	T	V	I	E	S	K	P	S	R	L
SEQ_B	-	T	H	I	E	S	R	P	S	R	L
SEQ_D	-	T	R	I	E	S	R	P	S	R	L
SEQ_E	S	Q	H	V	E	S	R	Q	S	R	-



Clustal W

- ## 1. Fast approximate alignments or full dynamic programming

$$\text{Distance} = 1 - \frac{\text{pairwise sequence identity \%}}{100}$$

- 2. Unrooted NJ tree is rooted mid-point => guide tree**

- 3.** Sequence weights from branch lengths. Highly similar sequence have low weights.

- #### 4. Aligning alignments...

Early alignment errors have no way of being corrected.

4.

**Progressive alignment:
Align following
the guide tree**

```

-----VILPDEKAAVTAALDKW-----DEVQZALGZALVVFSTQVFFESQDLST
-----VQLSDEKAAVTAALDKW-----KEVOGZALGZLVVFSTQVFFDSQDLSE
-----VLSADKTHVKAANKVQZAGAGYGAHALEKMFLLPDKTKHFFPFYDLS-----
-----VLSADKTHVKAANKSVQZAGAGYGAHALEKMFLLPDKTKHFFPFYDLS-----
-----VLSGEGHNVLELVKAGVADVAGQDILILKFFKMFLLPDKTKHFFDAFKELKT
PIVDTGSAVPLSAAEKTKIRSANAFVSTSTQVDDILVKKFFKMFLLPDKTKHFFPKQLTT
GALTSGOALVKSASHEEPMKPKFVFFVFFLAVFTHAPAKLFSYFKOTSE

```

PDVAVMGNPKVKANGKQKVLQAFSDQAHLD---KLQFFATLSKLCSDKLHVPDHFHFL
PQAVMGNPKVKANGKQKVLFSFGQVHLD---KLQFFATLSKLCSDKLHVPDHFHFL
----HQSQVKRGGKQKVDALTKAVHVD---DLQFALSLSDLKAKLHVPDHFHFL
----HQSQVKRGGKQKVDALTKAVHLD---DLQFALSLSDLKAKLHVPDHFHFL
EAMNLSLDLKLKQVTVVLTAQALTKRG---EAKLKLRLQSHATKFKIPKYLEF
ADQLKLSADQVGHAEIRIIVHVDASMDDT---EKQKSLKRLDSQKHAHQFQVDPQYKF
VP---QKQKELQAKGKQKVELVYKAKLQVTVGVVTVATLKLHLSQVHVKQG-VZADHFV

```

LQNVLCVLAHPPCKEFTFPVQAYQKVVAGVAMALSKYH
LQNVLVVLAHPPGQDFTPELQAYQKVVAGVAMALSKYH
LSBCLLVTLAALPAEFTPAVAHSLDKFLASVSTVLSKYR
LSBCLLSTLAVLPNDPFTPAVAHSLDKFLSVSTVLSKYR
ISRAIINVLSHPGQDFGADAGQAMHKAELFRKDIAKYKELGYQG
LAAVIADTVAAG-----DAVFKEKELMHCICILLHAY
VKAALVETIKIKVQAQKSKELHSLMSTAYDNLATUTLHMNDAA

```

1.

Pairwise alignment: Calculate distance matrix

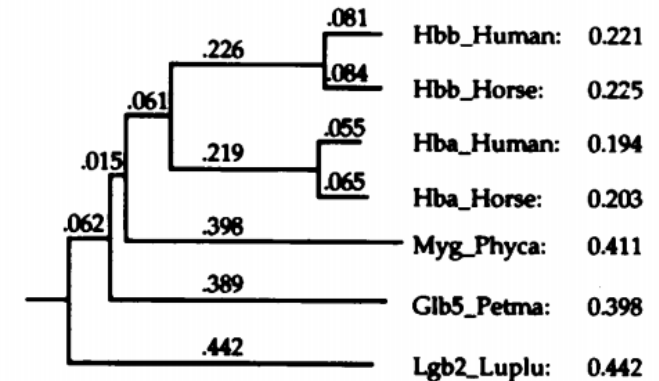
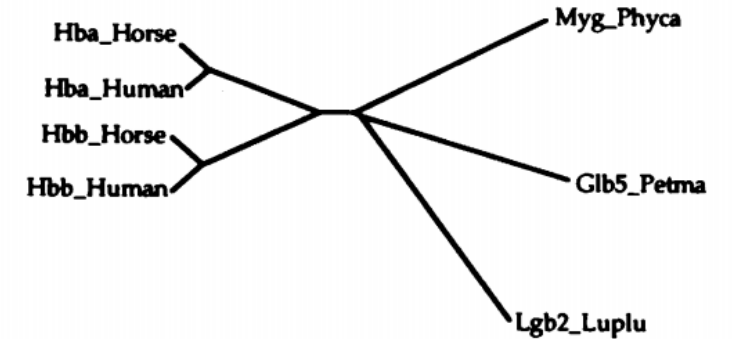
2.

Unrooted Neighbor-Joining tree

3.

**Rooted NJ tree (guide tree)
and sequence weights**

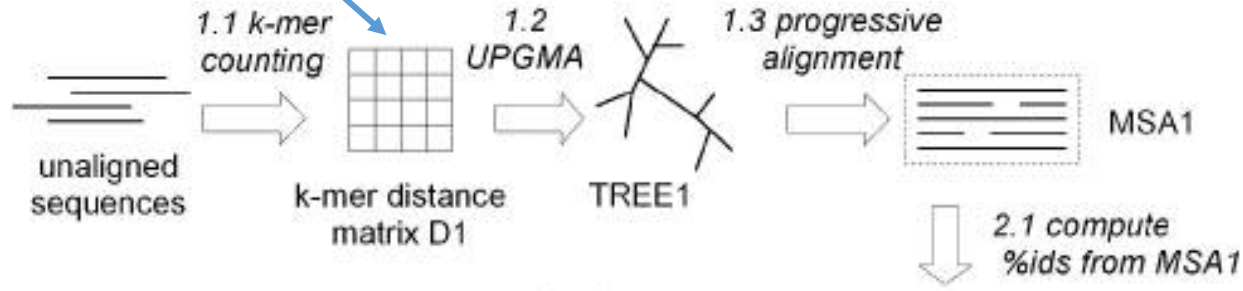
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6



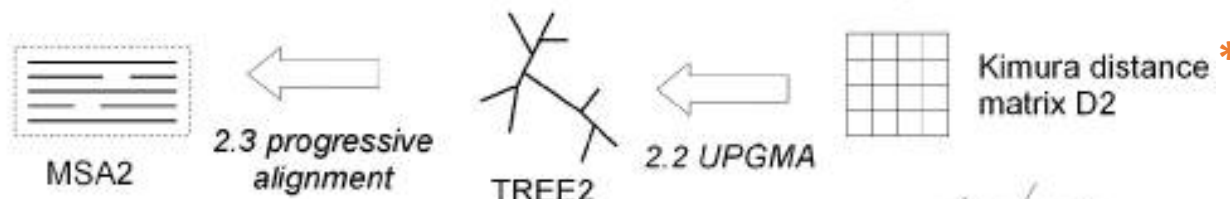
Muscle

How many k-mers match for each pair?

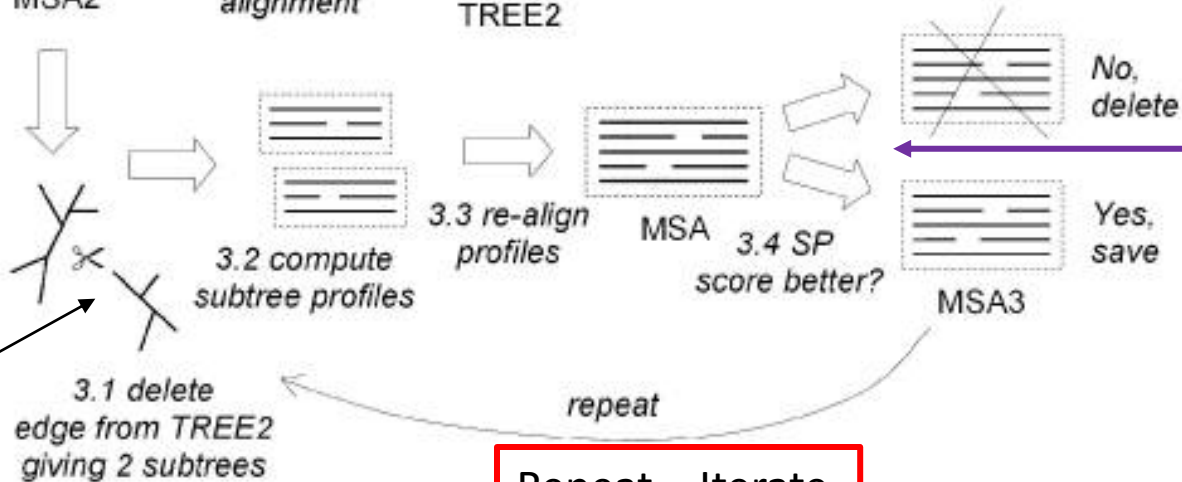
1. Draft progressive



2. Improved progressive



3. Refinement



Repeat = Iterate

Iterate

For every iteration, each edge is cut once, starting from the leaves and traversing up to the root.

Iterations continue until it reaches convergence or a user-specified number of iterations.

Score = average sum of all pairwise scores



*Kimura's protein distance matrix

All gapped sites are ignored

$$S = \frac{\text{\# exact matches}}{\text{Total \# of positions scored}}$$

$$D = 1 - S \quad (\text{uncorrected})$$

$$d = -\ln(1 - D - 0.2 D^2) \quad (\text{corrected})$$

Corrected here means corrected by the relationship between the number of observed amino acid substitutions and the actual amino acid substitutions.



