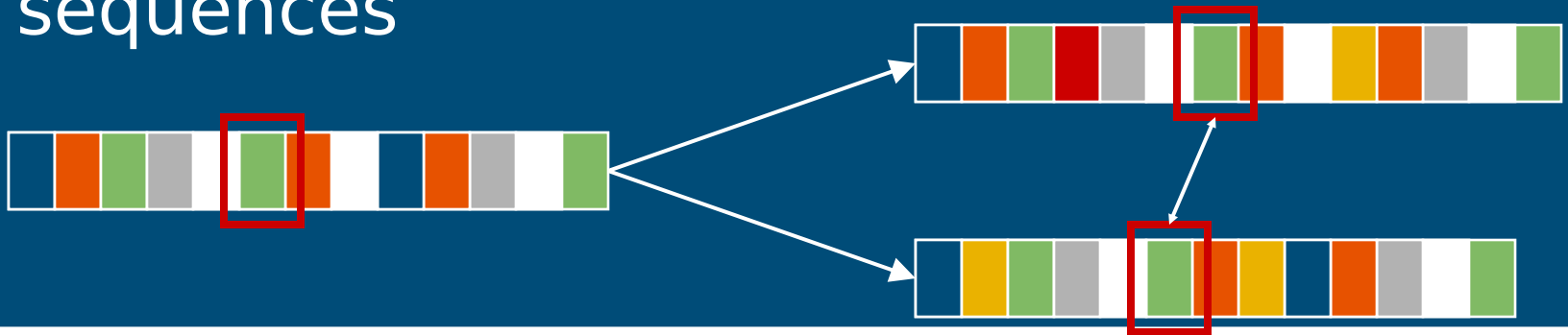


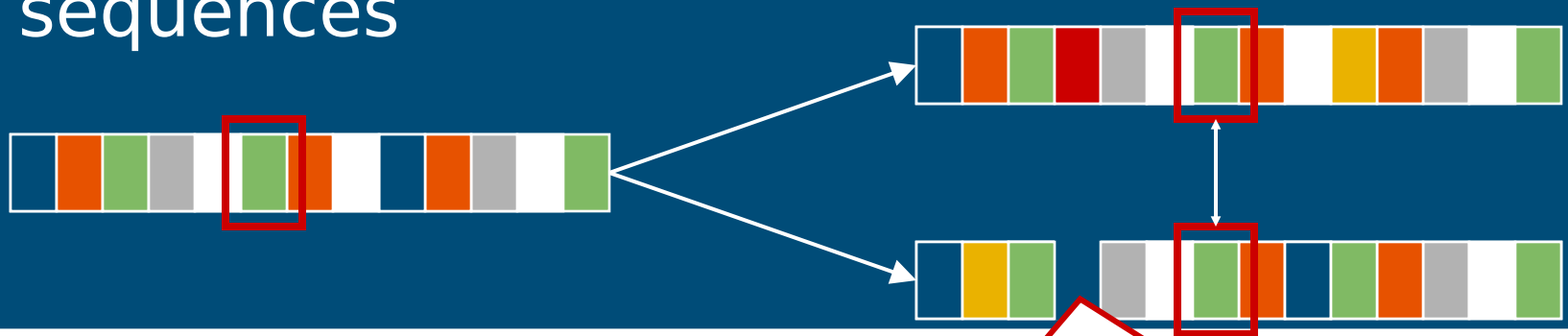
Multiple Alignment

- The purpose of a multiple alignment is to line up all residues that were derived from the same residue position in the ancestral gene or protein in any number of sequences



Multiple Alignment

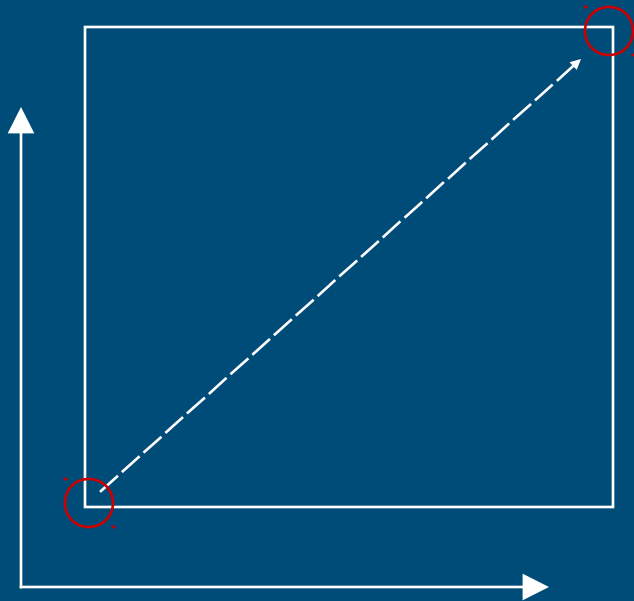
- The purpose of a multiple alignment is to line up all residues that were derived from the same residue position in the ancestral gene or protein in any number of sequences



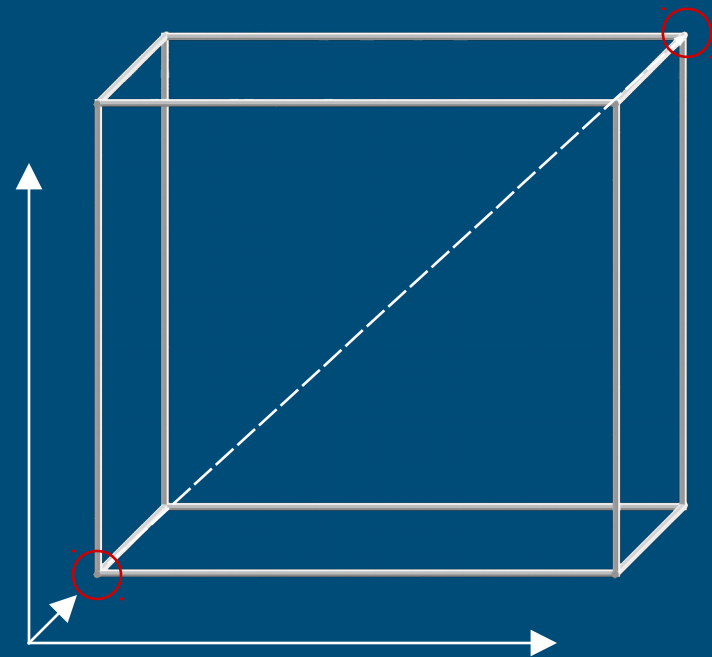
gap = insertion or deletion

From Pairwise To Multiple

Two sequences



Three sequences



And Beyond ...

- Assuming that it takes 1 kilobyte (1kb) to store one single sequence, then ...
- To do simultaneous alignment it takes for
 - 2 sequences : 1 megabyte of memory
 - 3 sequences : 1 gigabyte of memory
 - 4 sequences : 1 terabyte of memory
 - 5 sequences : 1 petabyte of memory
 - 6 sequences : 1 exabyte of memory

Iterative Approach

Or: the way we did multiple alignment in the middle ages

- First do all the easy alignments
- Then gradually add all the difficult ones
- But how do we know what alignments are easy or difficult?

Iterative Algorithm

as applied in ClustalW and similar programs

- Do a pairwise comparison of all sequences
- From this, calculate how sequences are related to each other (the more similar are easier to align)
- Perform multiple alignment in order; the most similar are aligned first, the others are saved for later

1: Pairwise Comparison

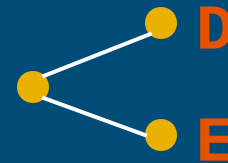
- Compare every single sequence to every other sequence, using pairwise sequence alignment
 - seq 1 \leftrightarrow seq 2 \Rightarrow 0.91
 - seq 1 \leftrightarrow seq 3 \Rightarrow 0.23
 - ...
 - seq 8 \leftrightarrow seq 9 \Rightarrow 0.87
- Record the resulting similarity scores
 - You can in fact use either similarities or differences between sequence

2: Calculate The Guide Tree

- Construct a guide tree from the matrix containing the pairwise comparison values, using a (relatively simple) clustering algorithm
 - UPGMA (PileUp & Clustal V)
 - Neighbor-Joining (Clustal W, Clustal X)

UPGMA - Step 1

	A	B	C	D	E
A	0	6	9	11	9
B	6	0	7	9	7
C	9	7	0	8	6
D	11	9	8	0	4
E	9	7	6	4	0



UPGMA - Step 2

	A	B	C	DE
A	0	6	9	10
B	6	0	7	8
C	9	7	0	7
DE	10	8	7	0



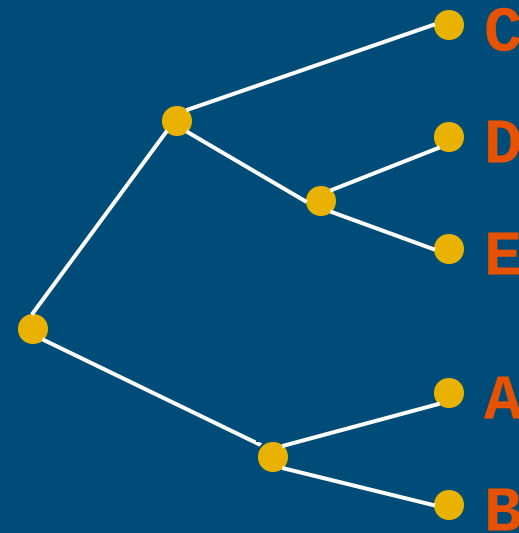
UPGMA - Step 3

	AB	C	DE
AB	0	8	9
C	8	0	7
DE	9	7	0



UPGMA - Step 4

	AB	CDE
AB	0	8.5
CDE	8.5	0



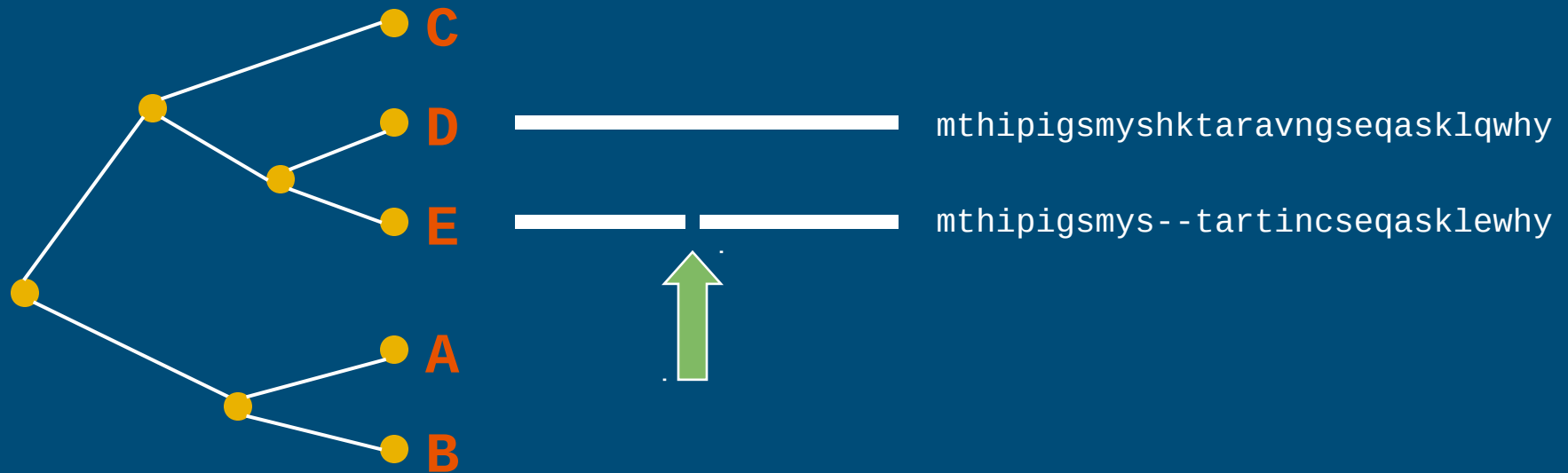
3: Multiple Alignment

- Using the guide tree, we start aligning groups of sequences
- The purpose of the guide tree is to know which sequences are most alike; so we can align the “easy” ones first, and postpone the tricky ones to later in the procedure!

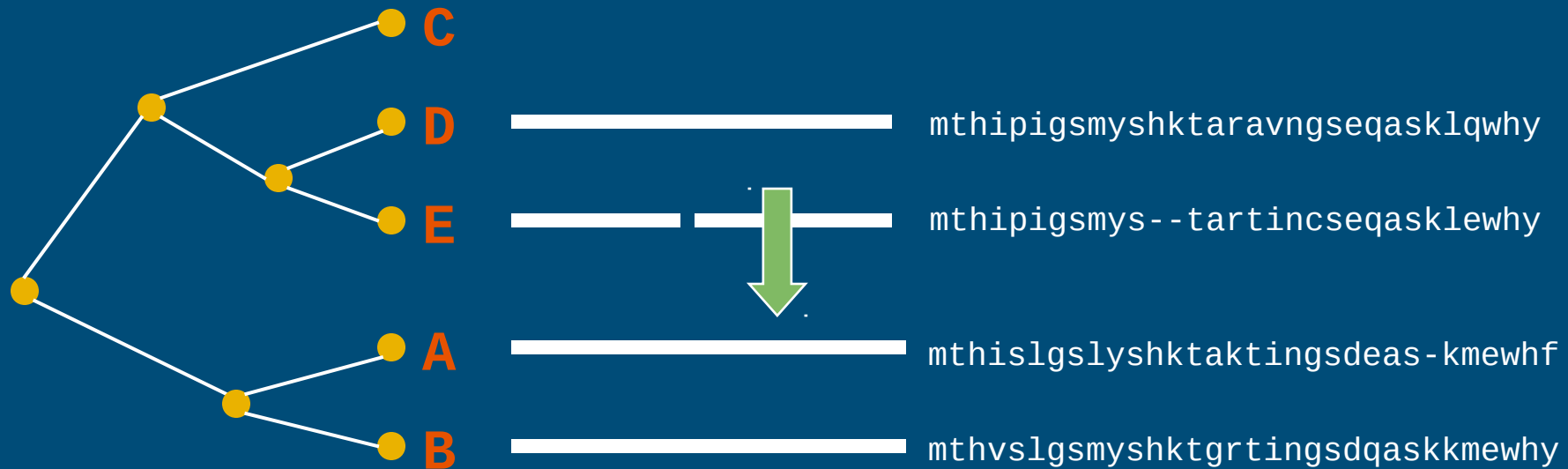
Input: Unaligned Sequences

A mthislgslyshktaktingsdeaskmewhf
B mthvslgsmyshtgrtingsdqaskkmewhy
C mshisitmyshktartidgseqaskmewhy
D mthipigsmyshtaravngseqasklqwhy
E mthipigsmystartincseqasklewhy

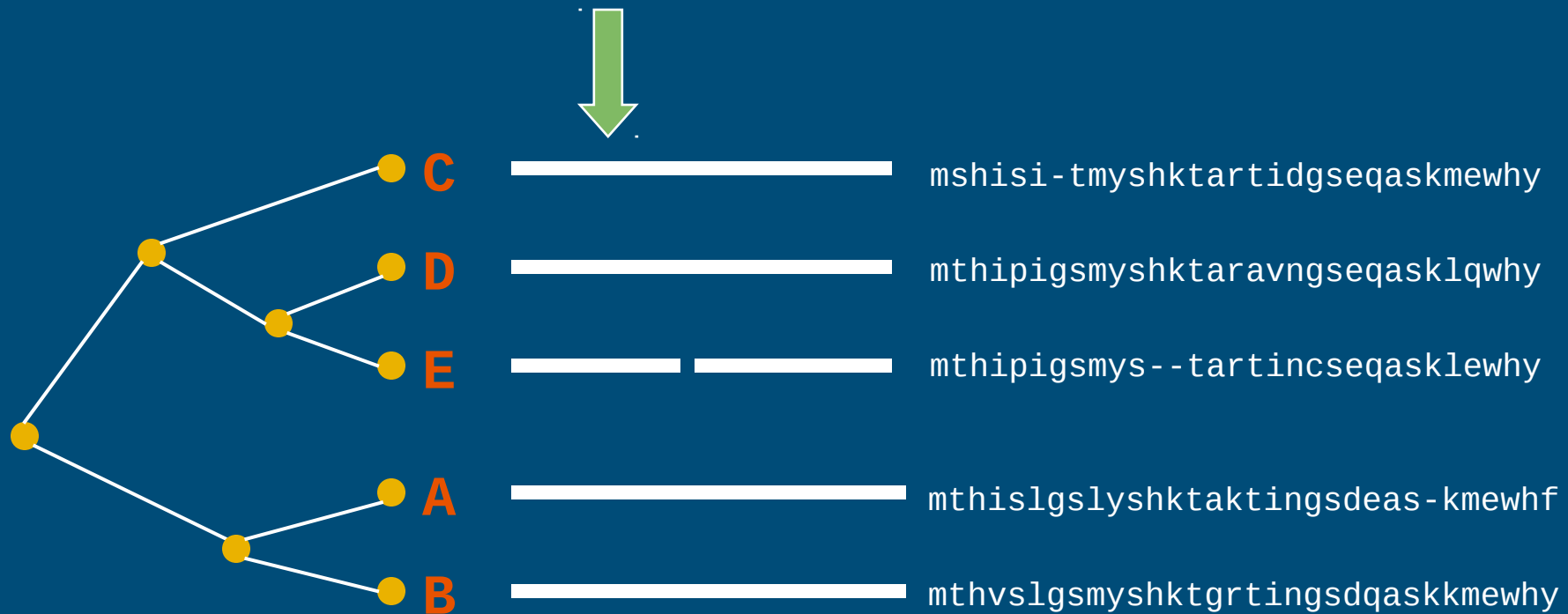
Multiple Alignment



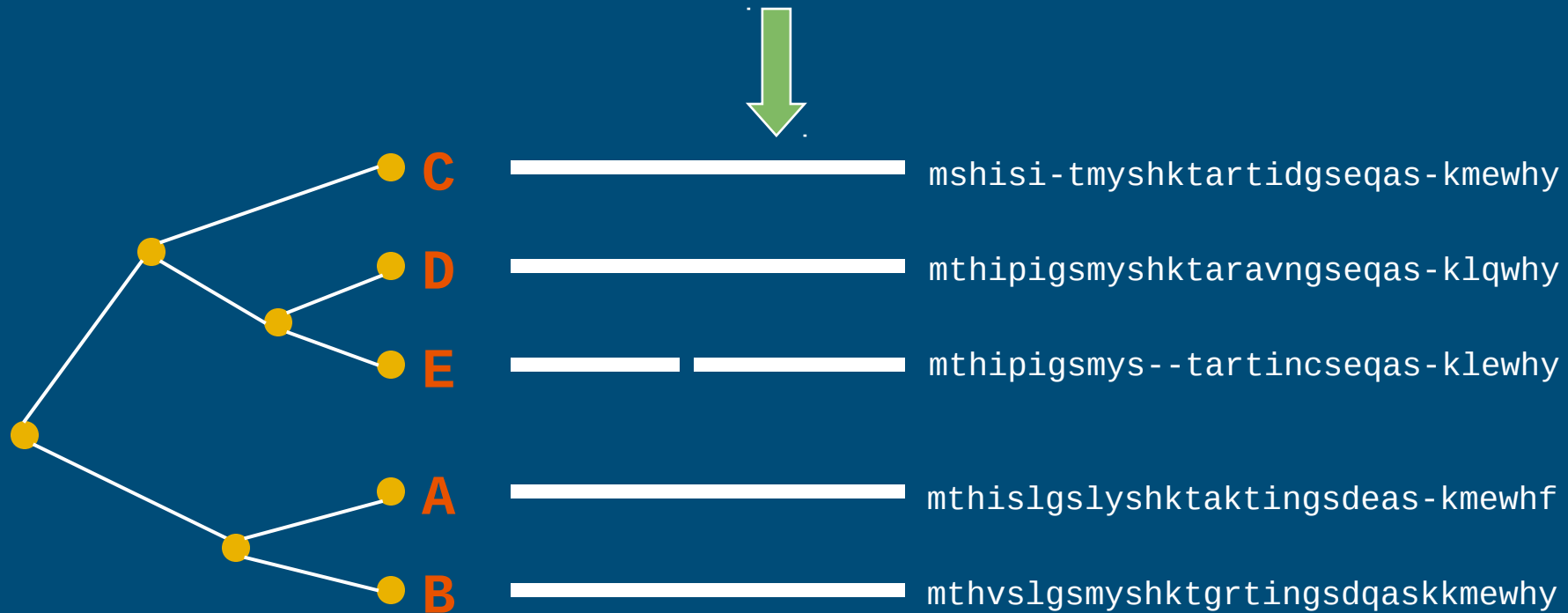
Multiple Alignment



Multiple Alignment



Multiple Alignment



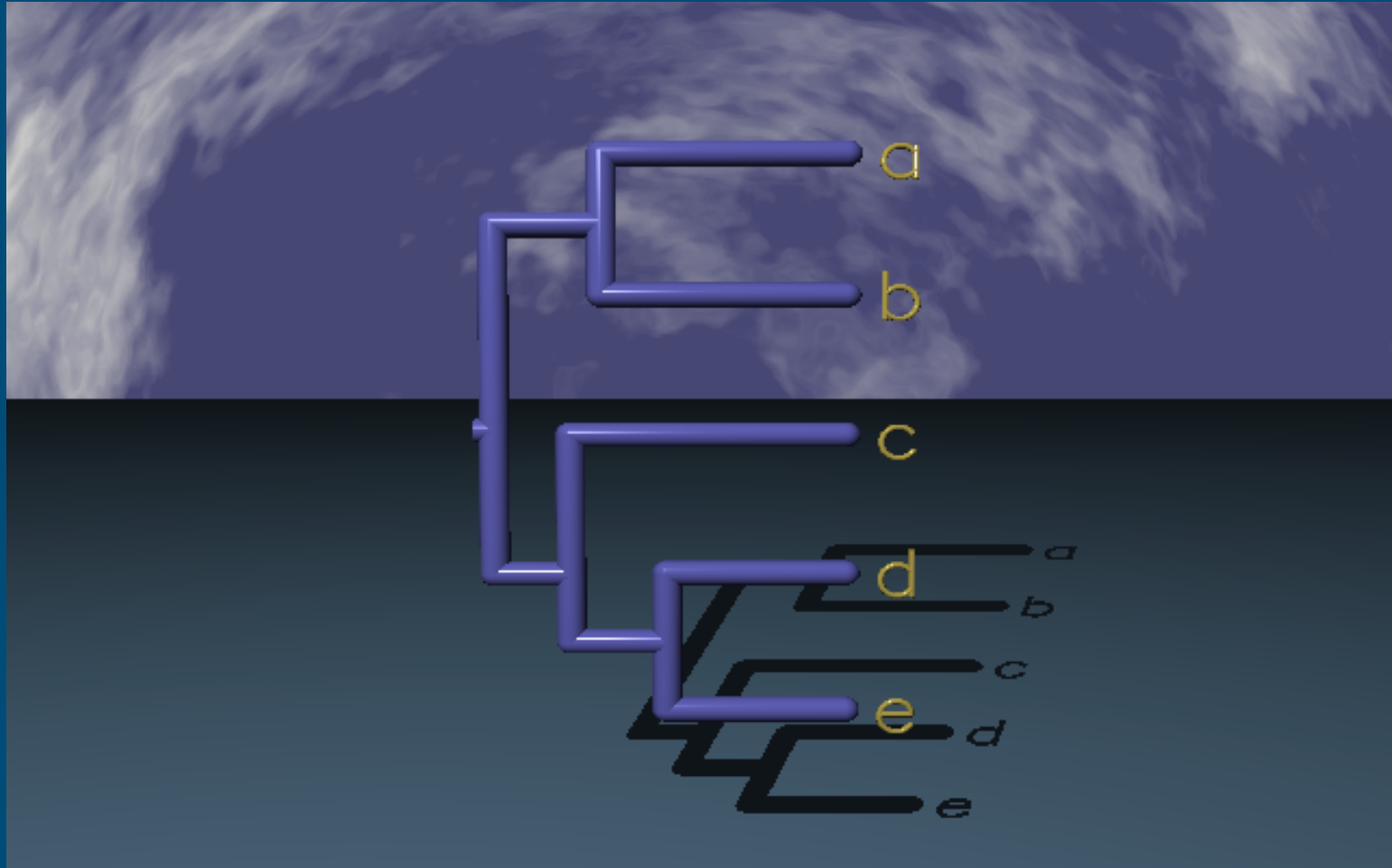
Output: Aligned Sequences

A mthislgslyshktaktingsdeas-kmewhf
B mthvslgsmyshtgrtingsdqaskkmewhy
C mshisi-tmyshktartidgseqas-kmewhy
D mthipigsmyshtaravngseqas-klqwhy
E mthipigsmys--tartincseqas-klewhy

Things To Remember ...

- Most multiple alignment programs are GLOBAL alignment programs
- The guide tree is NOT the phylogenetic tree

... no matter how beautiful it looks!



Things To Remember ...

- Most multiple alignment programs are GLOBAL alignment programs
- The guide tree is NOT the phylogenetic tree
- A multiple alignment program is the *starting point*, not the end point of producing a good, meaningful alignment

MUSCLE Algorithm

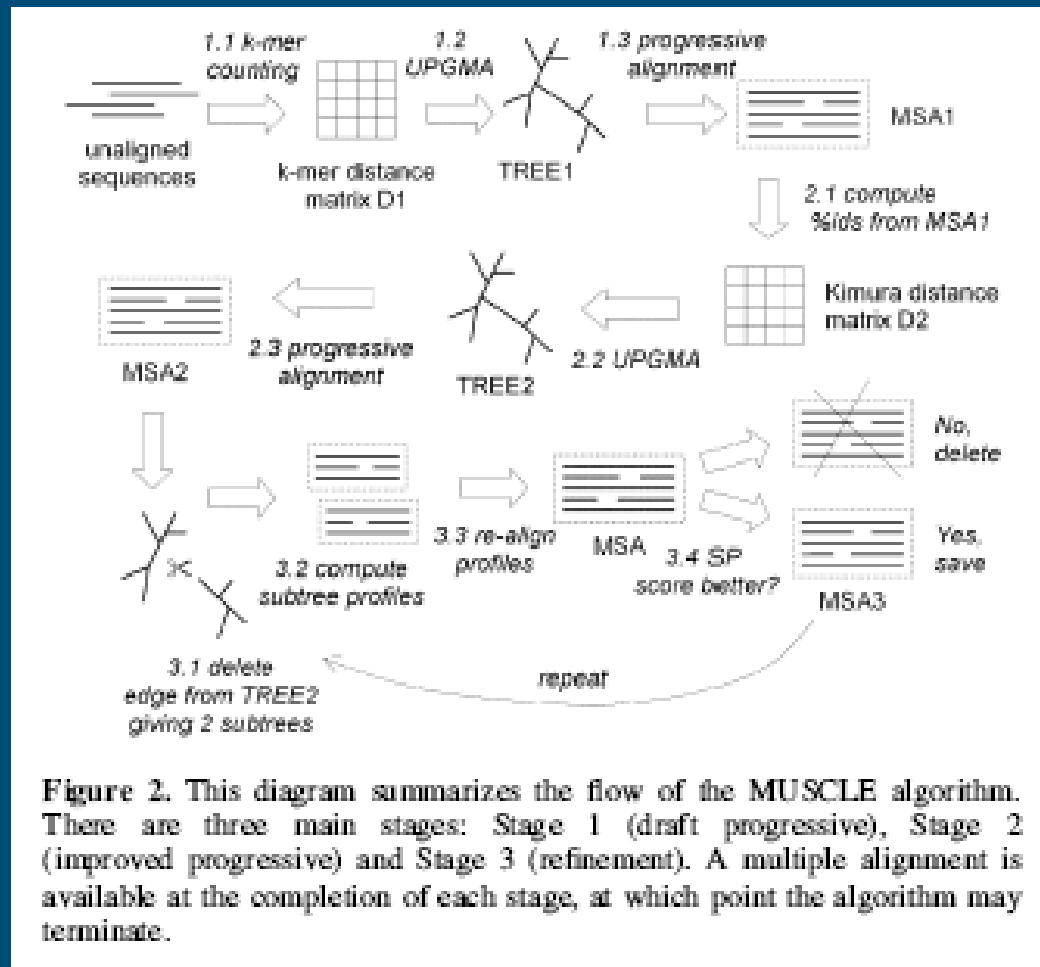


Figure 2. This diagram summarizes the flow of the MUSCLE algorithm. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). A multiple alignment is available at the completion of each stage, at which point the algorithm may terminate.