Multiple-Sequence-Alignment

Muptiale-S-qua-ce-Am--xnedt

Heuristic

The story so far

- Multidimensional DP is not going to happen

- We have some efficient local alignment heuristics (BLAST, FASTA, etc.)

- But these are not directly extensible to larger sets of sequences

Efficient msa???

- As with database searching, we want to trade optimality for efficiency

- But fast pairwise methods will not scale well (because we still have that $%#&* multidimensional matrix)

- So, we need heuristics that are *tailored* to msa

Overview

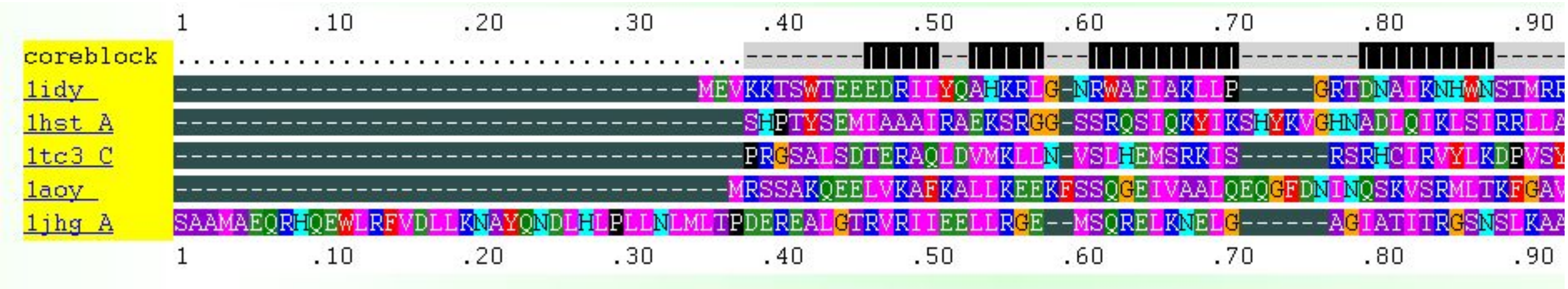
The magnitude of the problem

Progressive msa (MUSCLE)

Alternatives

The proving ground for MSAs

Example from BAliBASE:



```
          1        .10       .20       .30       .40       .50       .60       .70       .80       .90
coreblock ......................................--------|||||--|||||---||||||||||--------|||||||||-----
1idy      -----------------------------------MEVKKTSWTEEEDRILYQAHKRLG-NRWAEIAKLLP-----GRTDNAIKNHWNSTMRR
1hst A    --------------------------------------SHPTYSEMIAAAIRAEKSRGG-SSRQSIQKYIKSHYKVGHNADLQIKLSIRRLLA
1tc3 C    --------------------------------------PRGSALSDTERAQLDVMKLLN-VSLHEMSRKIS------RSRHCIRVYLKDPVSY
1aoy      -------------------------------------MRSSAKQEELVKAFKALLKEEKFSSQGEIVAALQEQGFDNINQSKVSRMLTKFGAV
1jhg A    SAAMAEQRHQEWLRFVDLLKNAYQNDLHLPLLNLMLTPDEREALGTRVRIIEELLRGE--MSQRELKNELG------AGIATITRGSNSLKAA
          1        .10       .20       .30       .40       .50       .60       .70       .80       .90
```

BALiBASE is actually horribly broken

- Edgar, C. (2010) Quality measures for protein alignment benchmarks. *Nucleic Acids Res* **38:** 2145-2153

But the point remains – these are extremely difficult problems!

What we need

- Algorithms that are better than exponential in their complexity

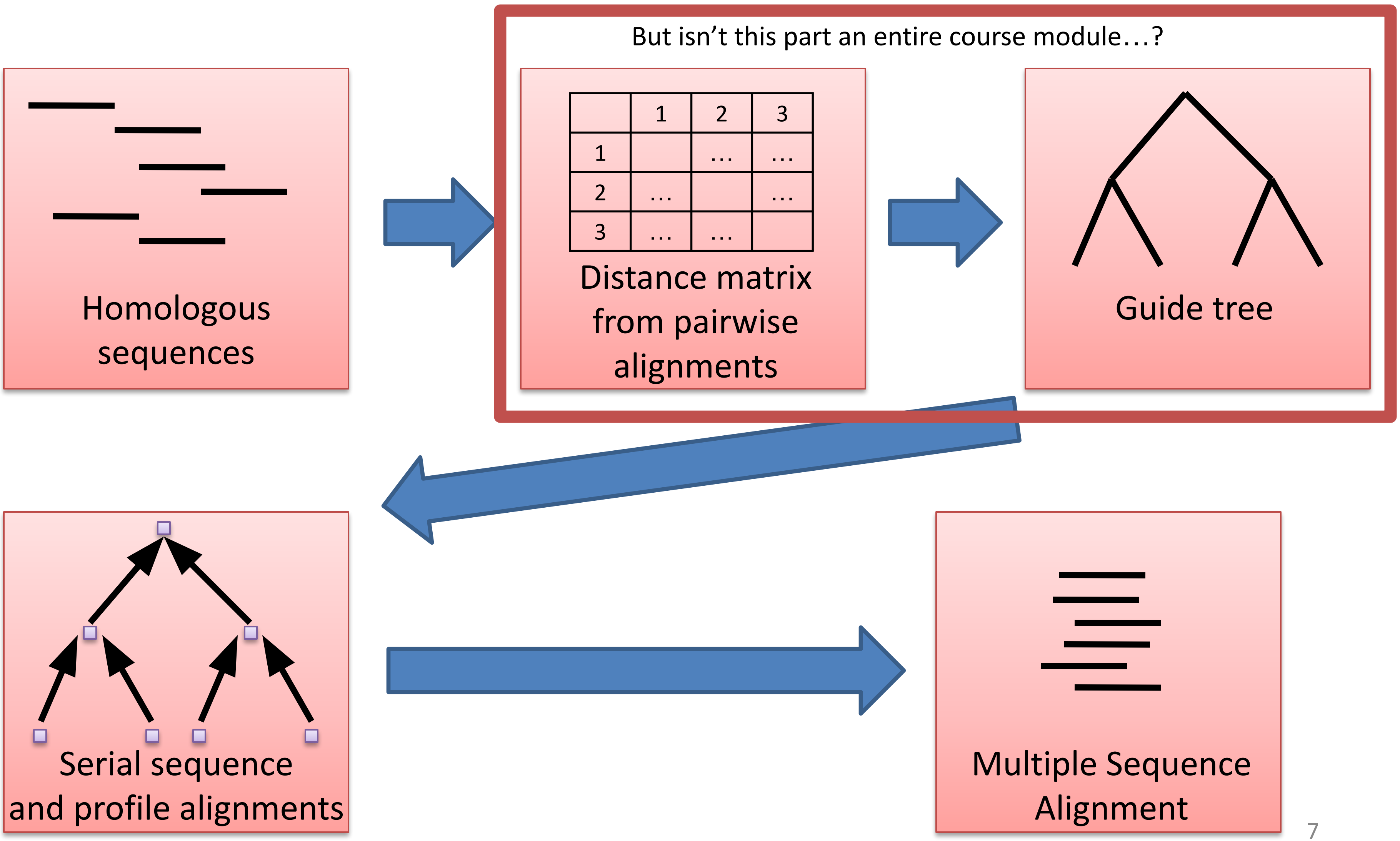
- (Pairwise DP is allowed – n^2 times a constant is not so bad)

- Often an OBJECTIVE FUNCTION (e.g., Sum of Pairs)

1	N
2	Q
3	Q
4	D

$$\begin{aligned} SP(N,Q,Q,D) = \; & 2 \times S(N,Q) \\ & + 2 \times S(D,Q) \\ & + S(Q,Q) \\ & + S(N,D) \end{aligned}$$

Progressive Alignment (1980s)



But isn't this part an entire course module…?

Homologous sequences

	1	2	3
1		…	…
2	…		…
3	…	…	

Distance matrix from pairwise alignments

Guide tree

Serial sequence and profile alignments

Multiple Sequence Alignment

MUSCLE - MUltiple Sequence Comparison by Log-Expectation

MUSCLE (Edgar, 2004)

- Three stages:

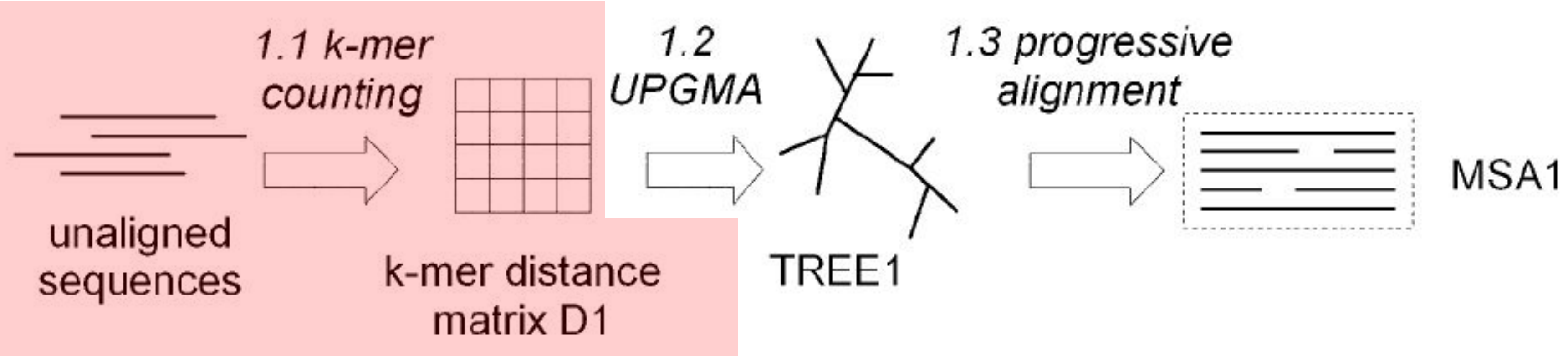
1. Draft progressive
2. Improved progressive
3. Iterative refinement

MUSCLE actually starts out with a **compressed alphabet**

There are many details and tweaks that I will not be talking about

Edgar (2004) *Nucleic Acids Res.*

MUSCLE Step 1



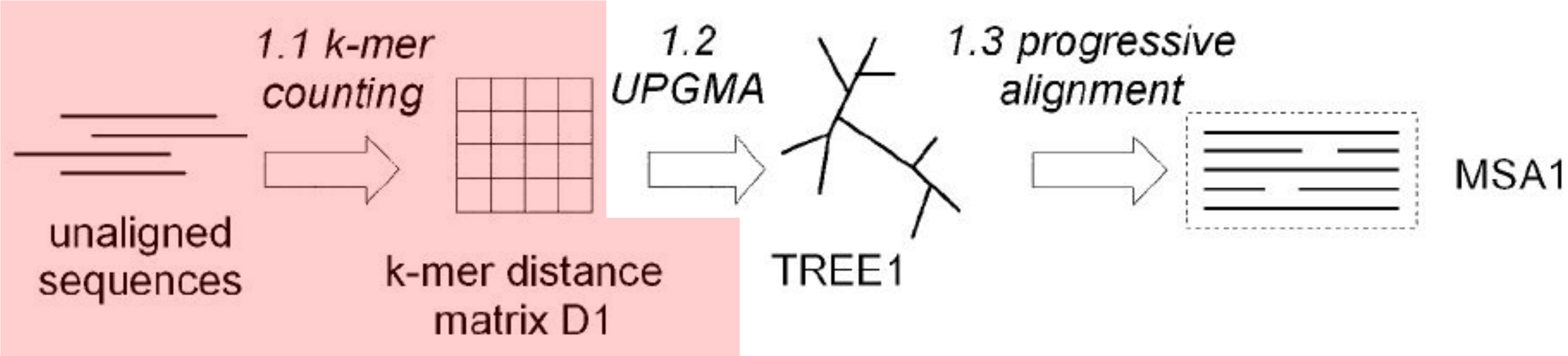
Unaligned sequences to *k*-mers
k-mer similarity for a pair of sequences:

$$F = \frac{\sum_{all\_kmers} \delta_{XY}(kmer)}{\min(L_X, L_Y) - k + 1}$$

δ_{XY} = 1 if k-mer is present in both
0 otherwise

Normalizing constant
(length of the shorter sequence)

MUSCLE Step 1

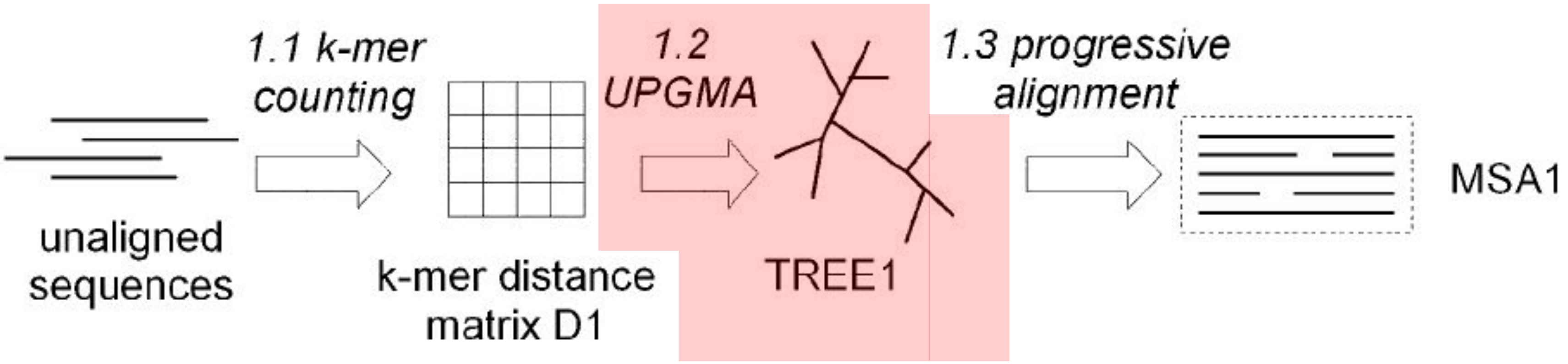


We convert F to a distance measure:

$$d_{kmer} = 1 - F$$

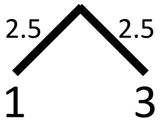
And populate a triangular distance matrix with *d* values

MUSCLE Step 1

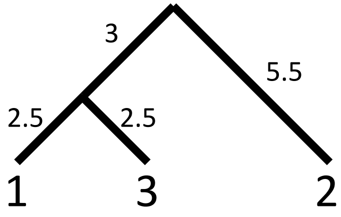


UPGMA: Unweighted Pair Grouping with Arithmetic Mean

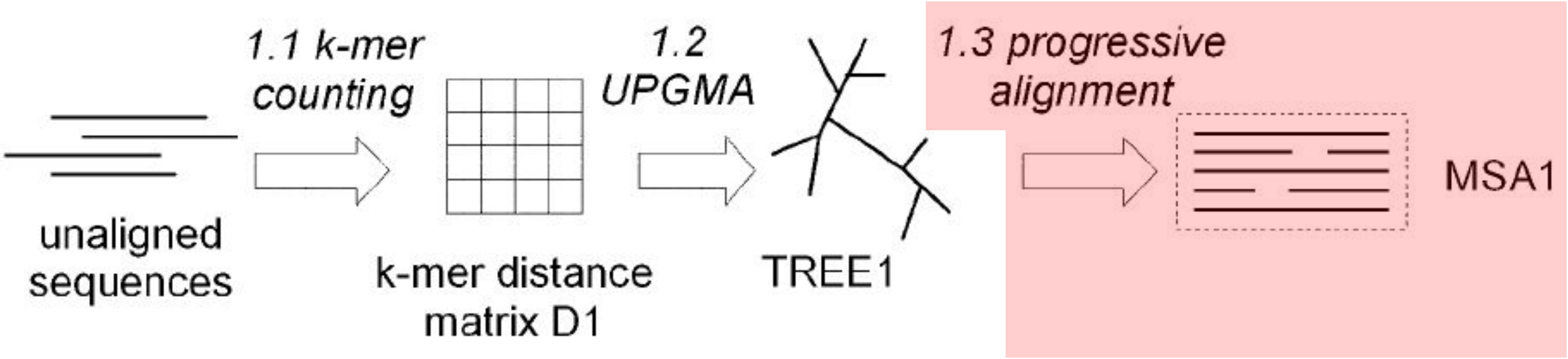
	1	2	3
1			
2	10		
3	5	12	



	1+3	2
1+3		
2	11	

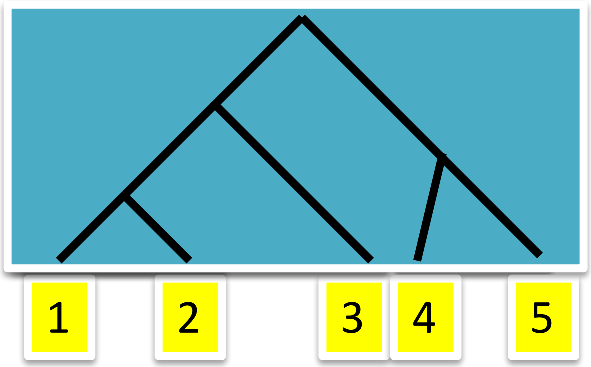


MUSCLE Step 1



Progressive alignment based on the UPGMA ‘guide’ tree:

Convert each sequence to a profile
Align profiles in prefix order based on the tree

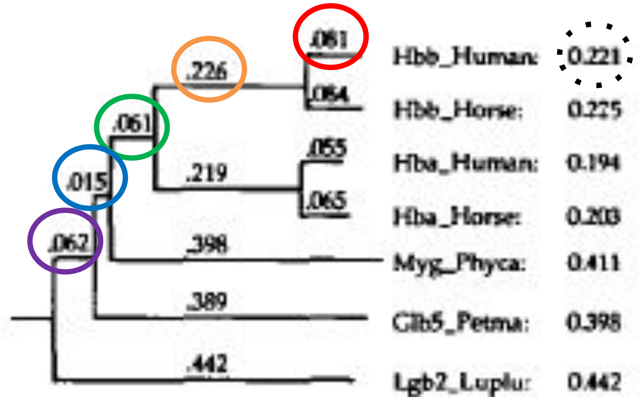


Each pairwise alignment is done with dynamic programming

But we only need to do $4n^2$ operations instead of n^5

How to align profiles

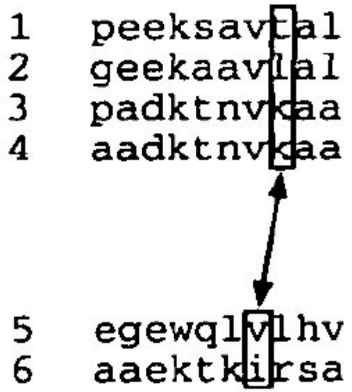
- First of all, sequences are *weighted* to reflect non-independent contributions



Weighting sequences by branch independence

(Thompson et al., 1994)

= .081
+ .226 /2
+ .061 /4
+ .015 /5
+ 0.062 /6



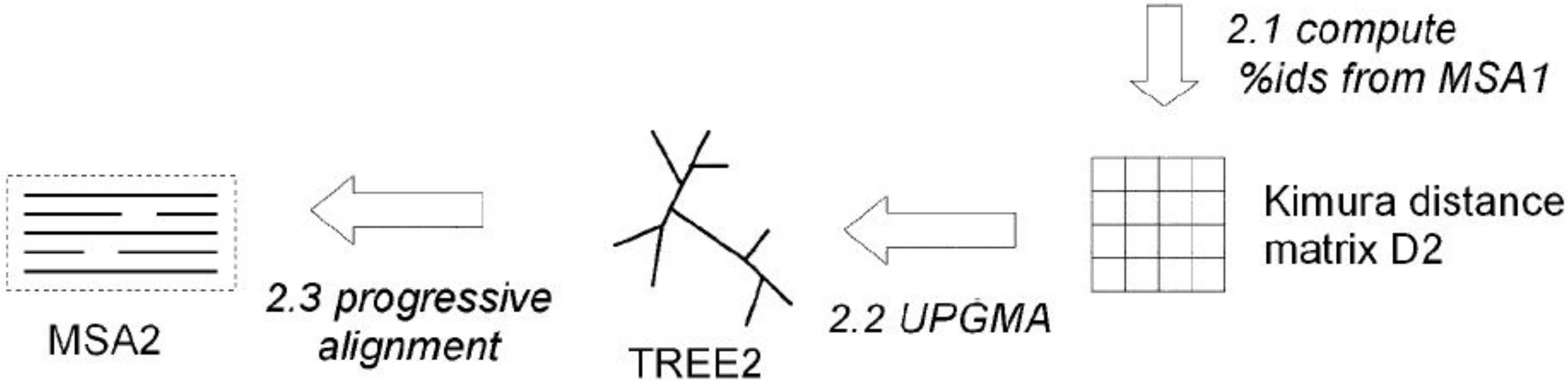
```
1   peeksavtal
2   geekaavlal
3   padktnvkaa
4   aadktnvkaa

5   egewqlvlhv
6   aaektkirsa
```

Scoring matches based on weights and scoring matrix

PAM250(T,V) * (w1 + w5)
+PAM250(T,I) * (w1 + w6)
…

Muscle Step 2



What is different here?

The distances used to build the initial guide tree were very crude

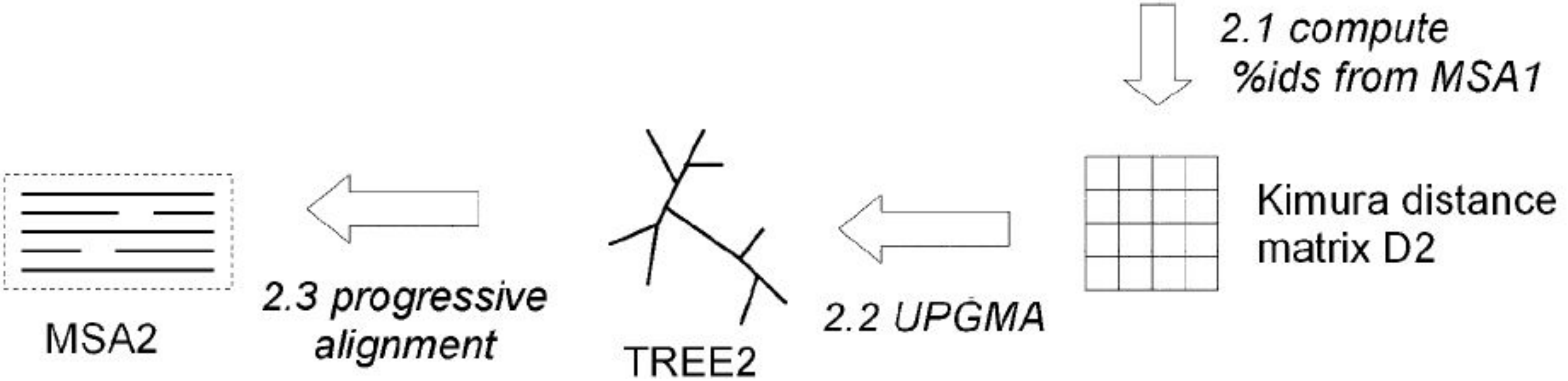
MUSCLE uses the first sequence alignment to compute Kimura distances:

$$d_{Kimura} = -\ln(1 - I - \boxed{I^2/5})$$

I = % identical

Multiple substitutions!

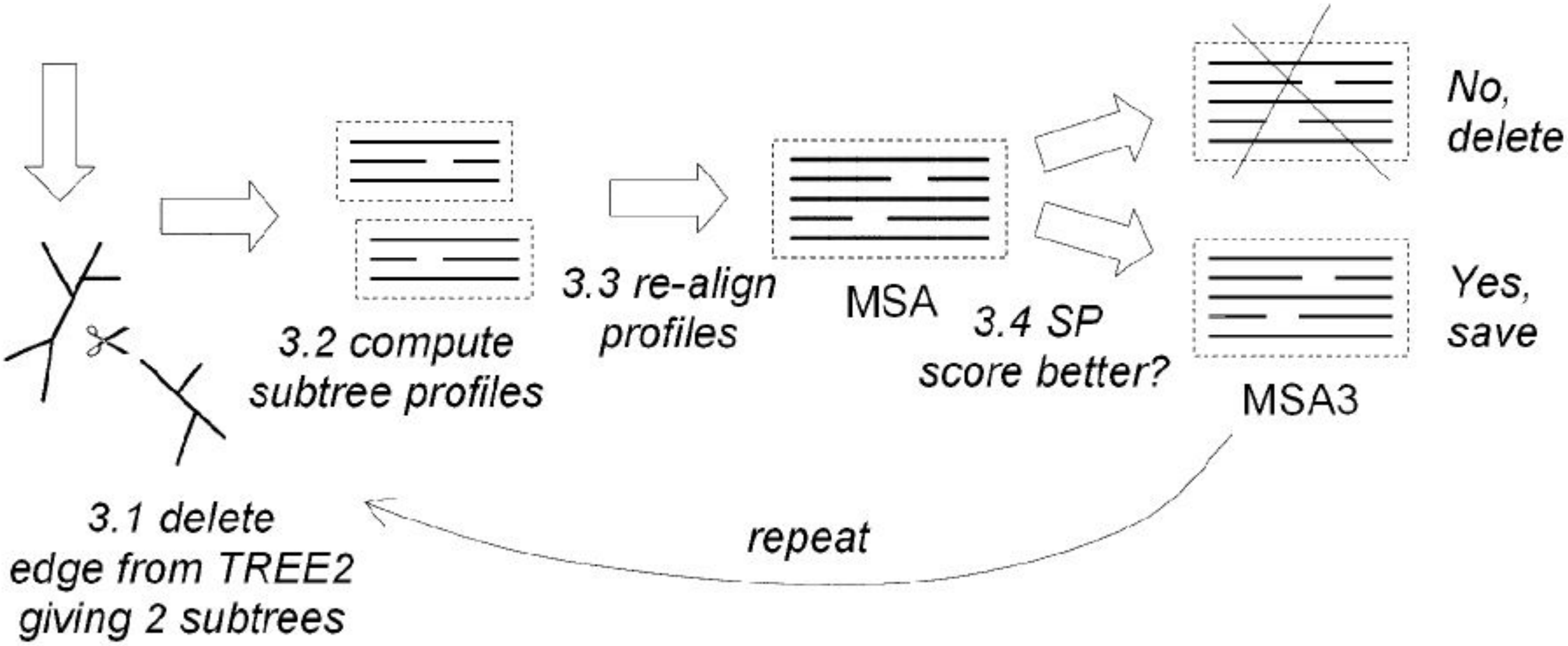
Muscle Step 2



With our more-accurate distances, build a new matrix, and a new UPGMA tree

Then build the multiple sequence alignment as before

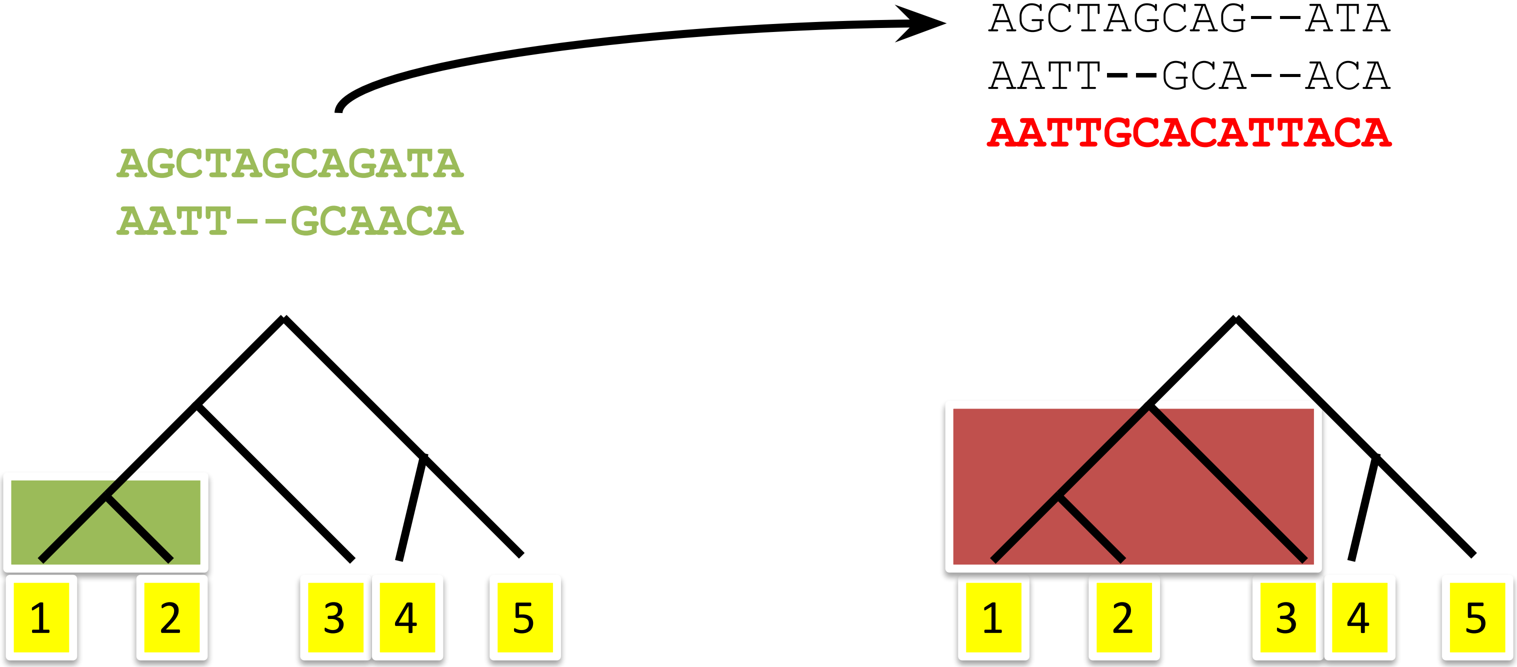
MUSCLE Step 3



Why do we do this?

The classic limitation of progressive alignment

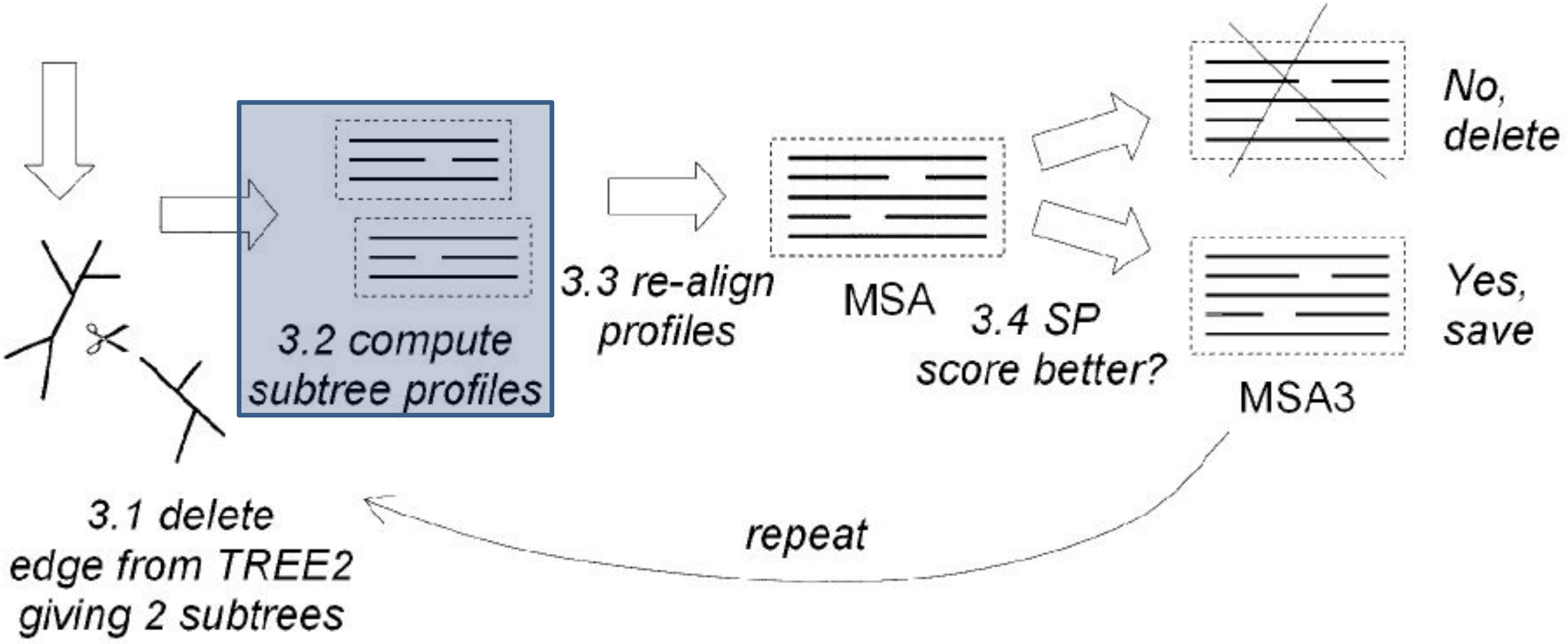
- “once a gap, always a gap”



```
AGCTAGCAGATA
AATT--GCAACA
```

```
AGCTAGCAG--ATA
AATT--GCA--ACA
AATTGCACATTACA
```

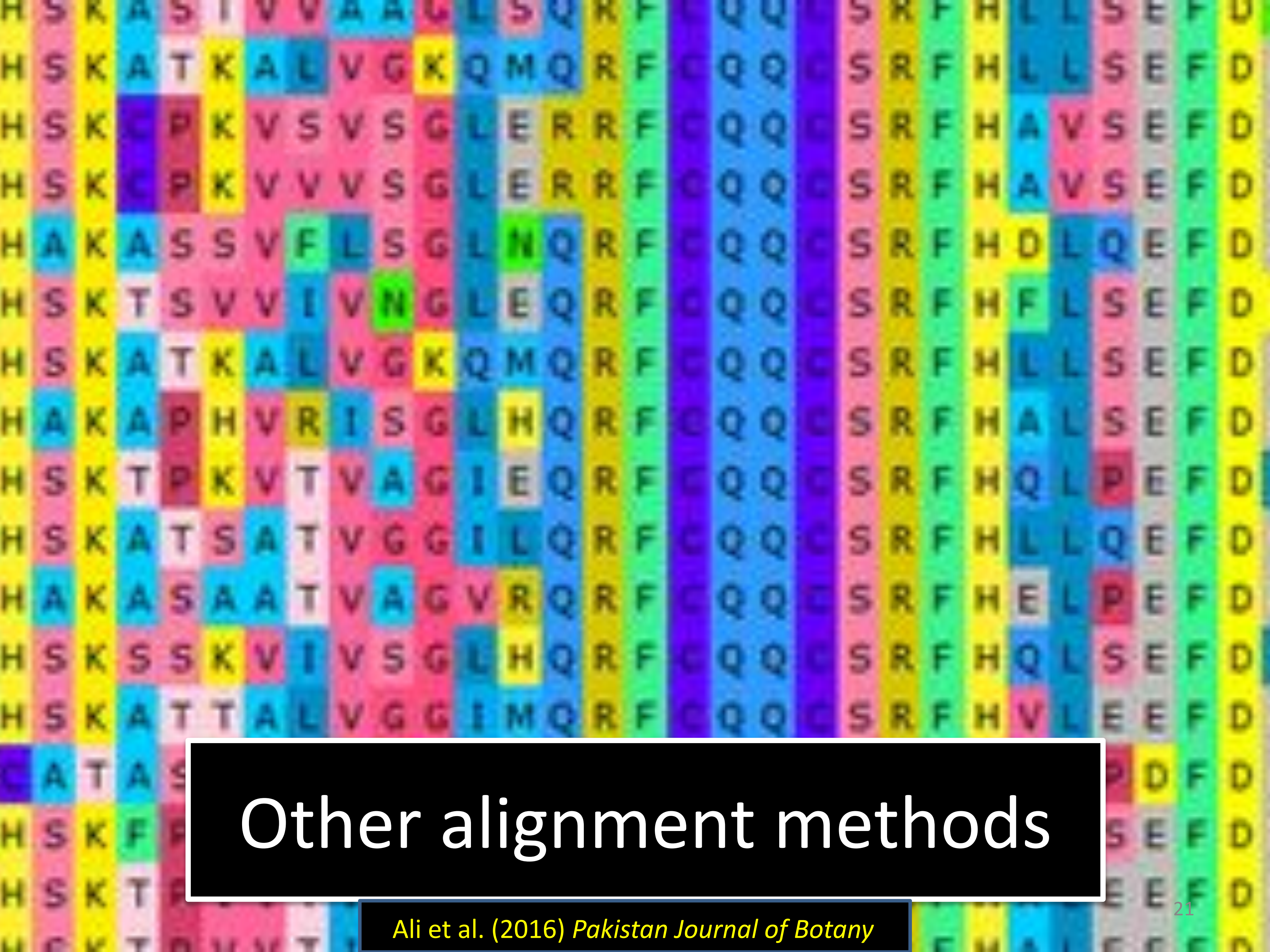
By breaking a branch of the guide tree, removing all gap-only columns and realigning the two profiles, we may find a better alignment



Advantages of MUSCLE

- It is ridiculously FAST – where quick n’ dirty is appropriate, it makes extensive use of the fastest available methods

- Phase 3 (iterative refinement) is very effective in overcoming the limitations of ‘traditional’ progressive methods



Other alignment methods

Ali et al. (2016) *Pakistan Journal of Botany*

MAFFT

Multiple alignment using fast Fourier transform

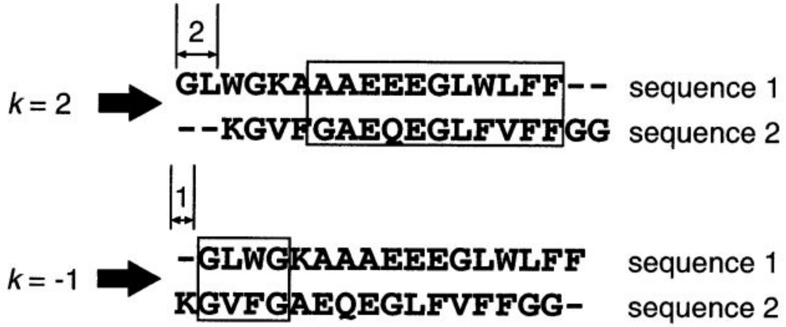
1. Represent amino acid sequences as vectors of *size* and *polarity*

Katoh et al. (2002) *Nucleic Acids Res.*

MAFFT

Multiple alignment using fast Fourier transform

1. Represent amino acid sequences as vectors of *size* and *polarity*
2. Look at correlation of these properties at different offsets


2
k = 2
GLWGKAAAEEEGLWLFF-- sequence 1
--KGVFGAEQEGLFVFFGG sequence 2
1
k = -1
-GLWGKAAAEEEGLWLFF sequence 1
KGVFGAEQEGLFVFFGG- sequence 2

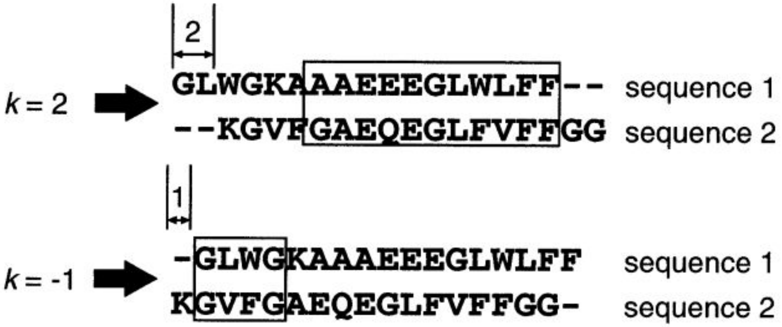
Regular correlation: O(n^2)
Fast Fourier Transform: O(nlogn)

Katoh et al. (2002) *Nucleic Acids Res.*

MAFFT

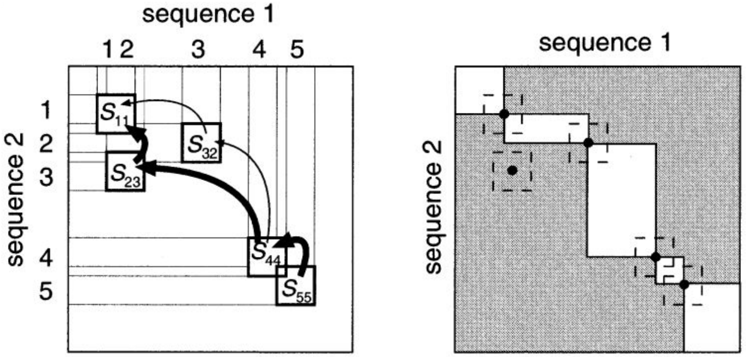
Multiple alignment using fast Fourier transform

1. Represent amino acid sequences as vectors of *size* and *polarity*
2. Look at correlation of these properties at different offsets



Regular correlation: O(N^2)
Fast Fourier Transform: O(N log N)

3. Use these as anchor points for DP

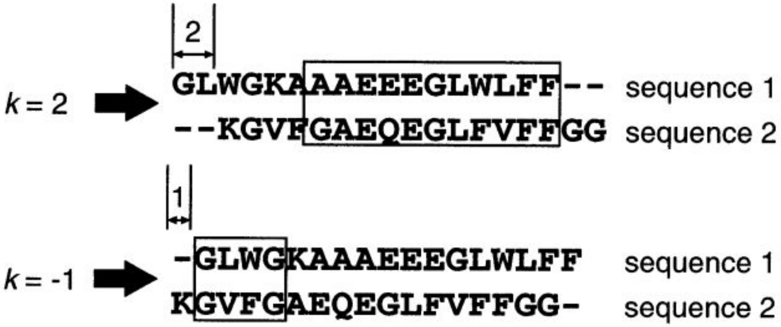


Katoh et al. (2002) *Nucleic Acids Res.*

MAFFT

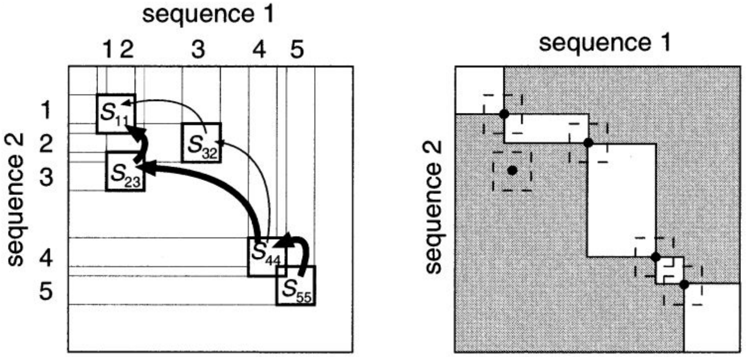
Multiple alignment using fast Fourier transform

1. Represent amino acid sequences as vectors of *size* and *polarity*
2. Look at correlation of these properties at different offsets



Regular correlation: $O(n^2)$
Fast Fourier Transform: O(nlogn)

3. Use these as anchor points for DP



4. Progressive alignment

Katoh et al. (2002) *Nucleic Acids Res.*

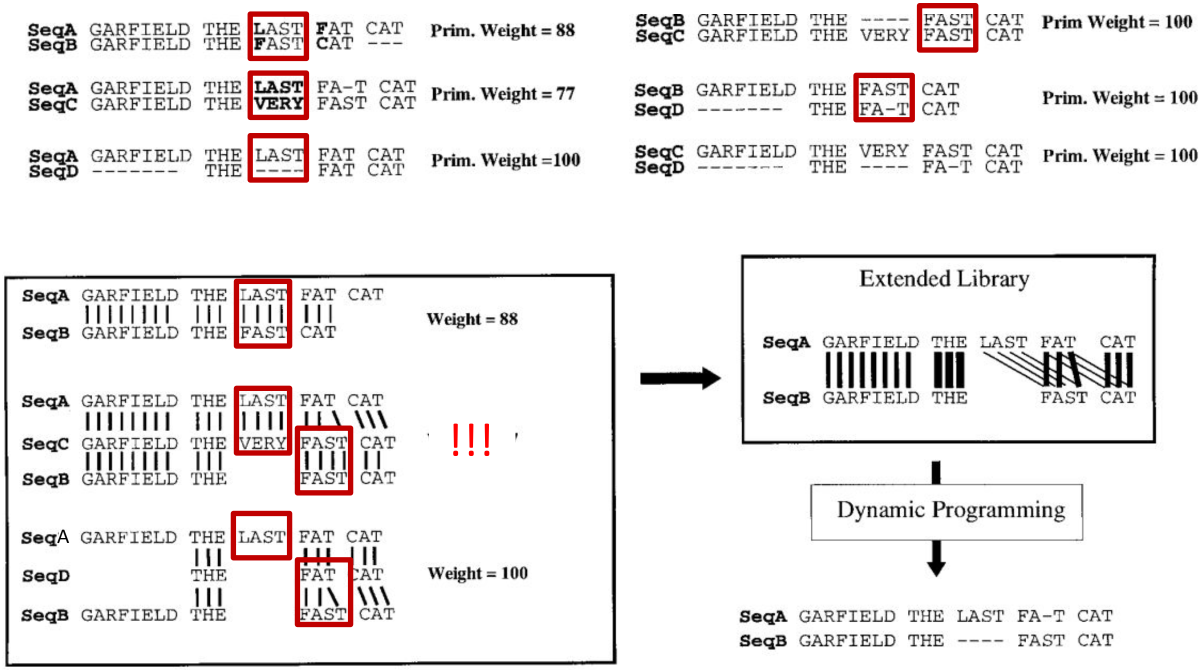
T-COFFEE

Tree-based Consistency Objective Function for alignment Evaluation

(and other consensus-based methods)

- Input sets of alignments of the same sequences (generated e.g. using different other programs, other parameter settings)

Pairwise alignments!



```
SeqA GARFIELD THE LAST FAT CAT    Prim. Weight = 88
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT   Prim. Weight = 77
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT    Prim. Weight =100
SeqD -------- THE ---- FAT CAT

SeqB GARFIELD THE ---- FAST CAT   Prim Weight = 100
SeqC GARFIELD THE VERY FAST CAT

SeqB GARFIELD THE FAST CAT        Prim. Weight = 100
SeqD -------- THE FA-T CAT

SeqC GARFIELD THE VERY FAST CAT   Prim. Weight = 100
SeqD -------- THE ---- FA-T CAT
```

```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT          Weight = 88

SeqA GARFIELD THE LAST FAT CAT
SeqC GARFIELD THE VERY FAST CAT     !!!
SeqB GARFIELD THE FAST CAT

SeqA GARFIELD THE LAST FAT CAT
SeqD          THE      FAT CAT      Weight = 100
SeqB GARFIELD THE      FAST CAT
```

```
SeqA GARFIELD THE LAST FAT  CAT
SeqB GARFIELD THE      FAST CAT
```

```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE ---- FAST CAT
```

Notredame et al. (2000) *J. Mol. Biol.*

ProbCons

probabilistic consistency-based alignment

- Key idea: *best* alignment vs the set of *good* alignments (expressed as a probability: see next lecture)
- The pairing of amino acids in the *best* alignment might not be the pairing we see across a greater cumulative probability of *good* alignments
- The point: replace the PAM matrix score for a pair of amino acids with their cumulative probability across all alignments, then do dynamic programming!

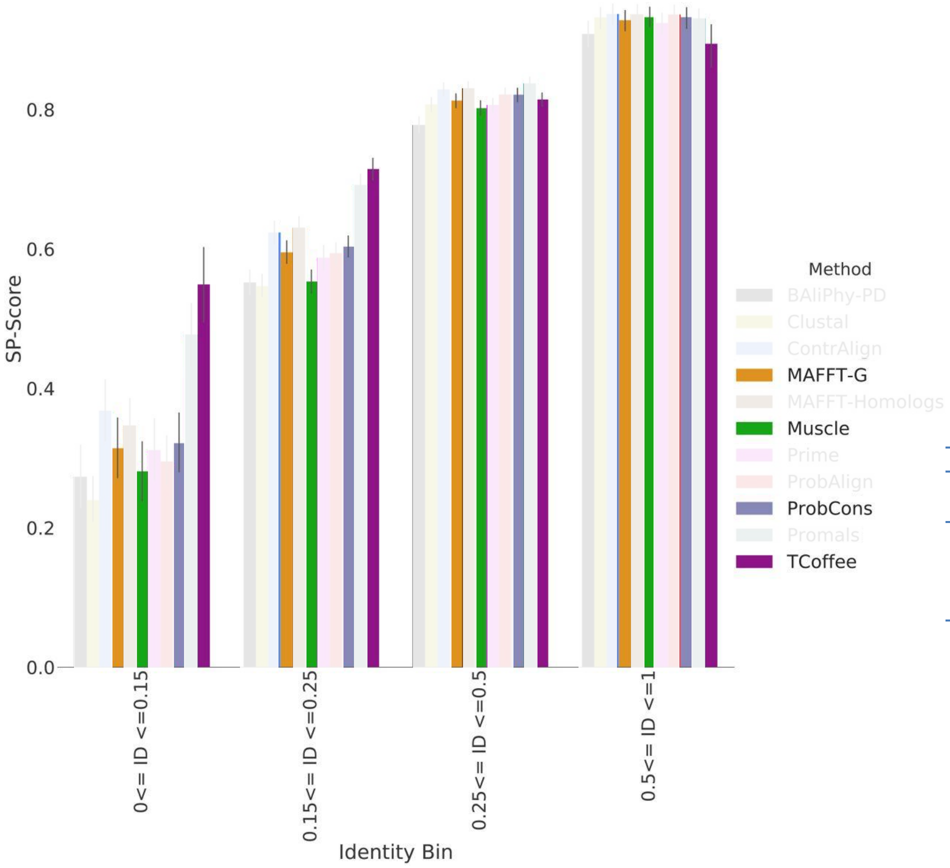
(see bonus slides at end of deck)

Do et al. (2005) *Genome Res.*

BAli-Phy

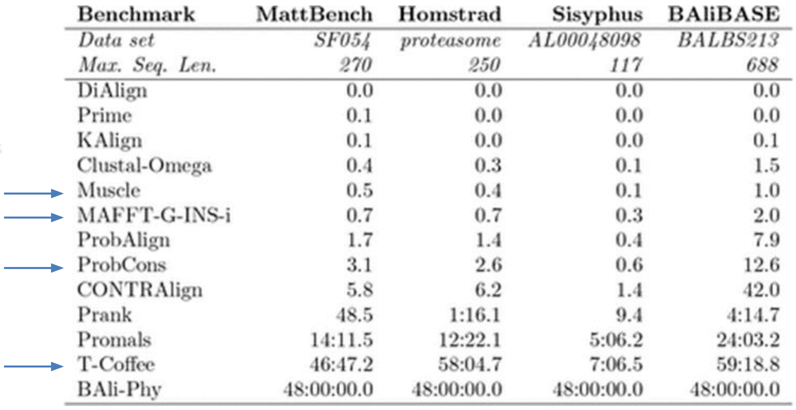
- Joint-inference of alignment and guide-tree
- Hand-wavey Bayesian approaches we will talk about during phylogenetics
- Most principled approach.
- INCREDIBLY and IMPRACTICALLY slow.

Comparison



More dissimilar (hard)

More similar (easy)



Benchmark	MattBench	Homstrad	Sisyphus	BAliBASE
Data set	*SF054*	*proteasome*	*AL00048098*	*BALBS213*
Max. Seq. Len.	*270*	*250*	*117*	*688*
DiAlign	0.0	0.0	0.0	0.0
Prime	0.1	0.0	0.0	0.0
KAlign	0.1	0.0	0.0	0.1
Clustal-Omega	0.4	0.3	0.1	1.5
Muscle	0.5	0.4	0.1	1.0
MAFFT-G-INS-i	0.7	0.7	0.3	2.0
ProbAlign	1.7	1.4	0.4	7.9
ProbCons	3.1	2.6	0.6	12.6
CONTRAlign	5.8	6.2	1.4	42.0
Prank	48.5	1:16.1	9.4	4:14.7
Promals	14:11.5	12:22.1	5:06.2	24:03.2
T-Coffee	46:47.2	58:04.7	7:06.5	59:18.8
BAli-Phy	48:00:00.0	48:00:00.0	48:00:00.0	48:00:00.0

Runtime (s)

Nute et al. (2018) *BioRXiv.*

Conclusions

- Lots of different ways to approach the problem
  - Progressive
  - Consensus
  - Iterative

- Usually (but not always) pairwise DP is an important component of the method