
Heuristic
Multiple-Sequence-Alignment
Neptiale-S-qua-ce-Am--xnedt

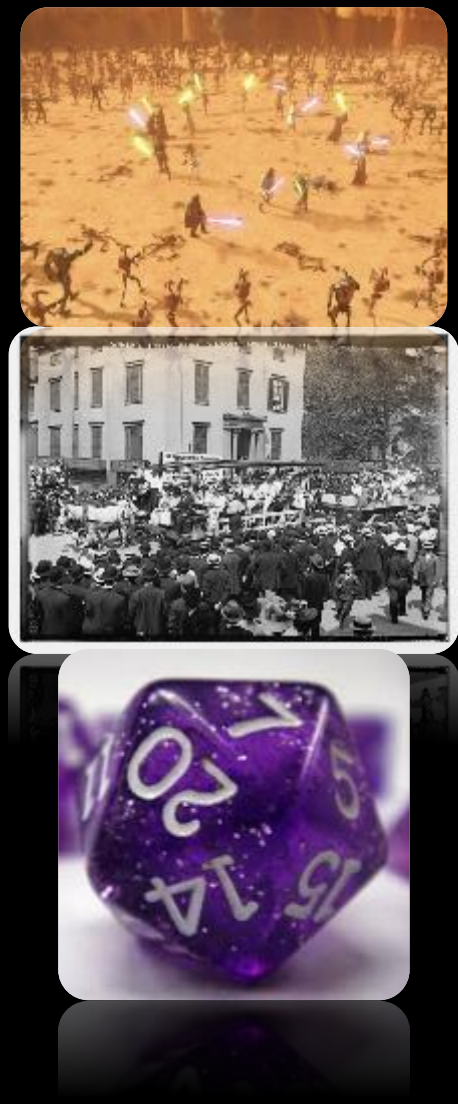
The story so far

- 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



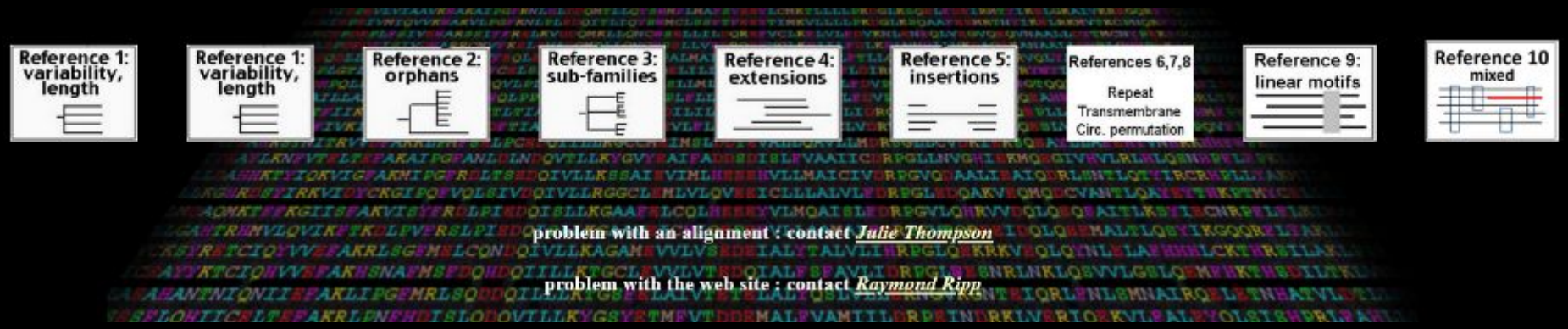
The magnitude of the problem

Progressive MSA (MUSCLE)

Other widely used methods

How do we validate MSAs?

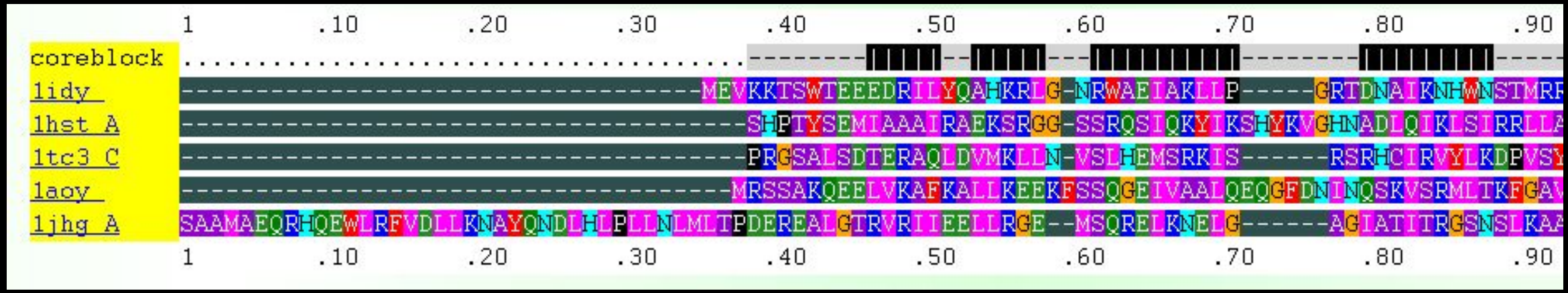
- Benchmark databases – sets of proteins whose alignment is based on known structures
- Proteins must be distantly related to make this an interesting challenge
- e.g. BAliBASE



Thompson et al. (2005) *Proteins*

The proving ground for MSAs

Example from BAliBASE:



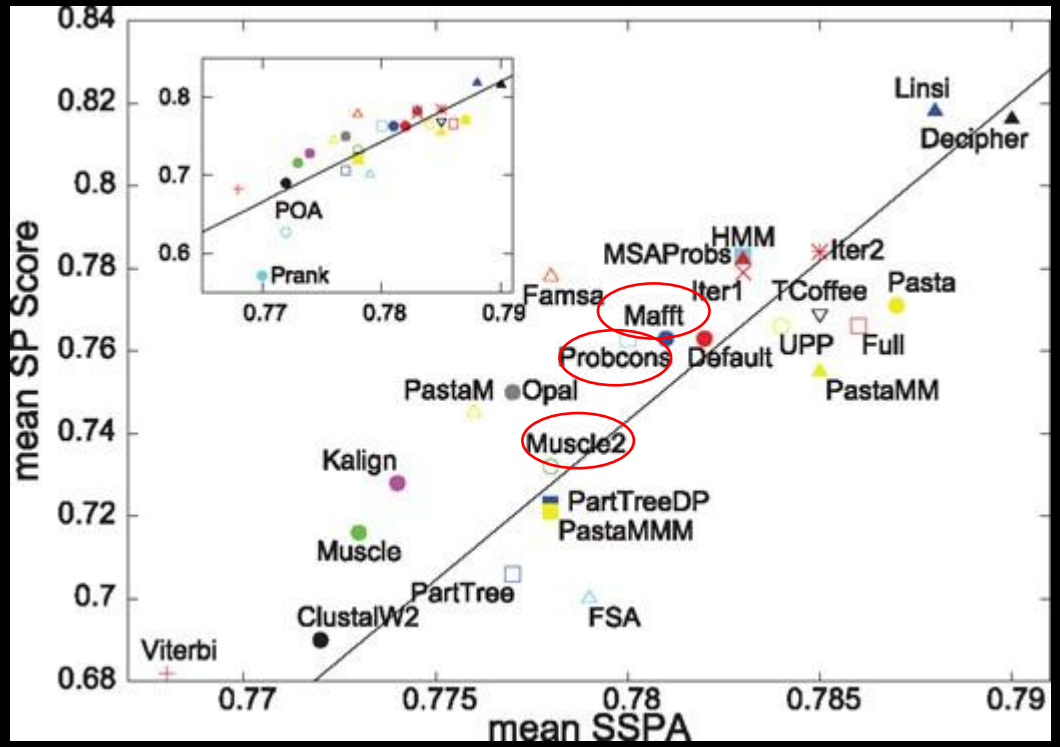
BALiBASE is actually horribly broken – lots of alignments of non-homologous "things"
- Edgar, C. (2010) Quality measures for protein alignment benchmarks.
Nucleic Acids Res **38:** 2145-2153

But the point remains – these are extremely difficult problems!

An Interesting Variation

QuanTest2 –validation with secondary structure prediction

Sum-of-pairs
(sequence based)



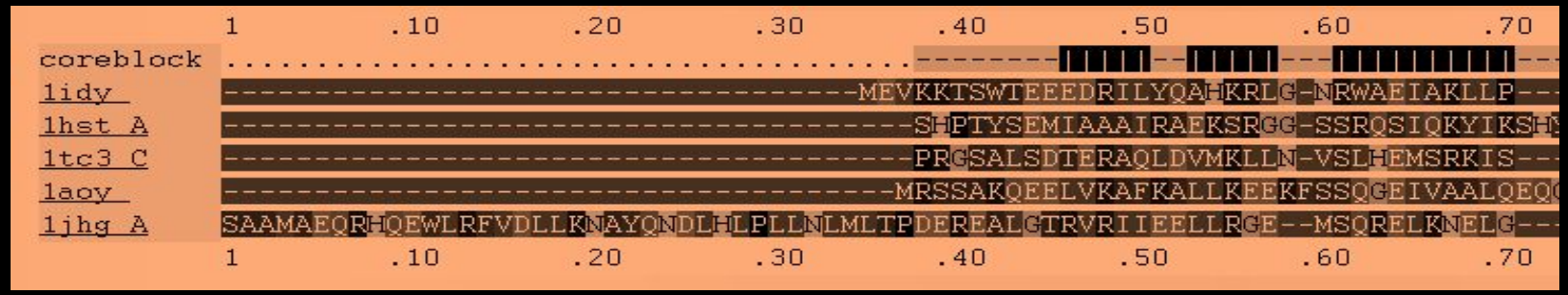
Look at those axis values!

***SSPA (structure based)

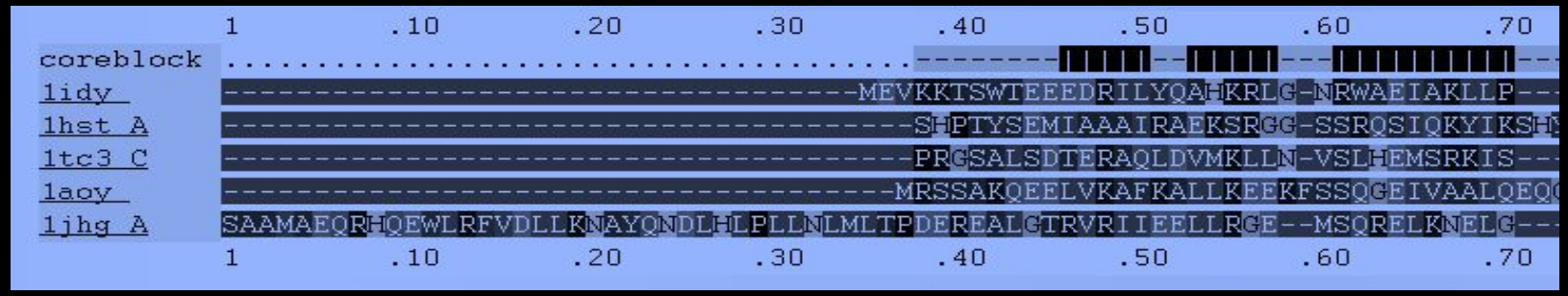
Sievers and Higgins (2020) *Bioinformatics*

Evaluating similarities

Reference alignment:



Your alignment:

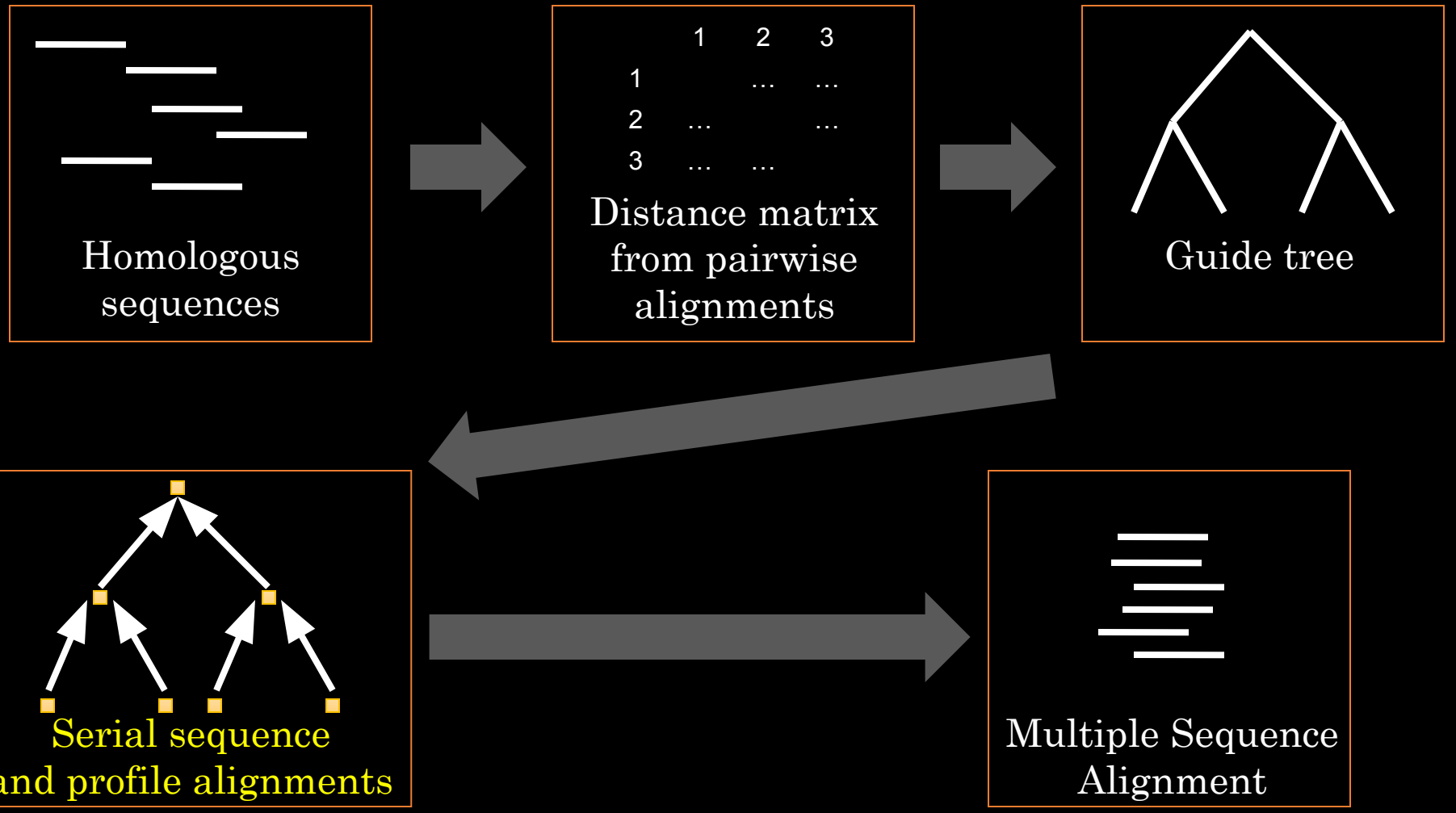


How many pairs of amino acids or entire alignment columns are consistent between the two?

MSA - What we need

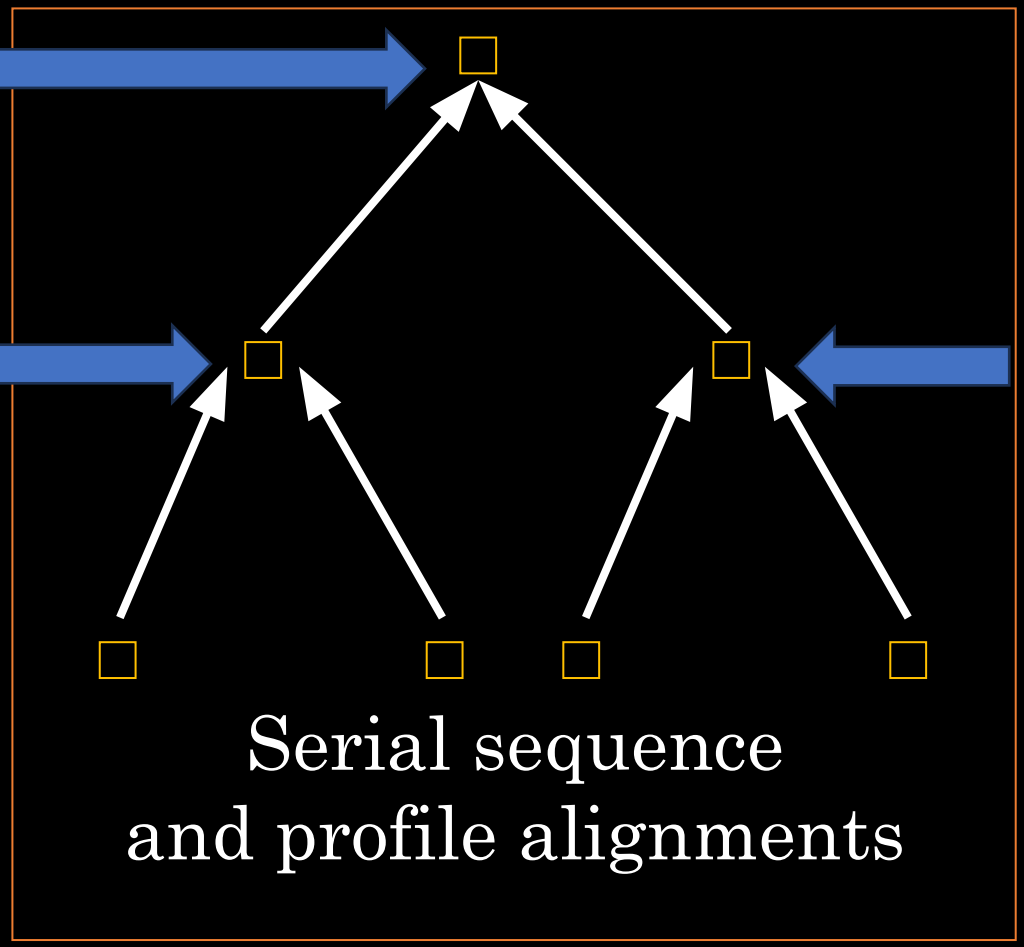
- Algorithms that are better than exponential in their complexity
- (Pairwise DP is allowed – n^2 times a constant is not so bad)
- Often an objective function (e.g., Sum of Pairs)

Progressive Alignment



The Crucial Idea of Progressive Alignment

Never align more than two things at once!



MUSCLE

Edgar (2004) *Bioinformatics*
Edgar (2022) *Nat Comms*

MUSCLE

- Three stages:

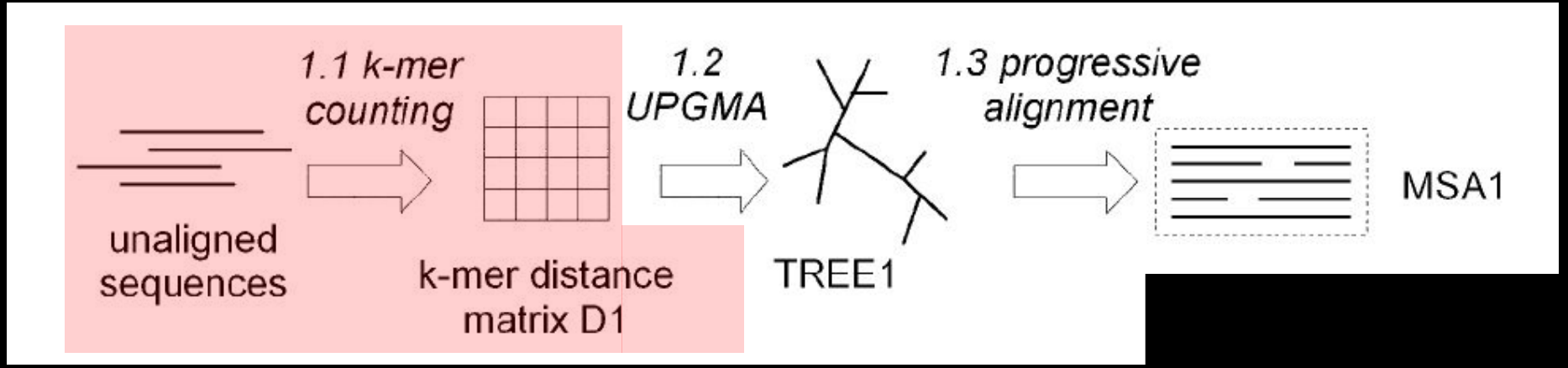
1. Draft progressive
2. Improved progressive
3. Iterative refinement

MUSCLE actually starts out with a compressed alphabet (kinda like DIAMOND-BLASTX)

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

Edgar (2004) *Nucleic Acids Res*

MUSCLE Step 1



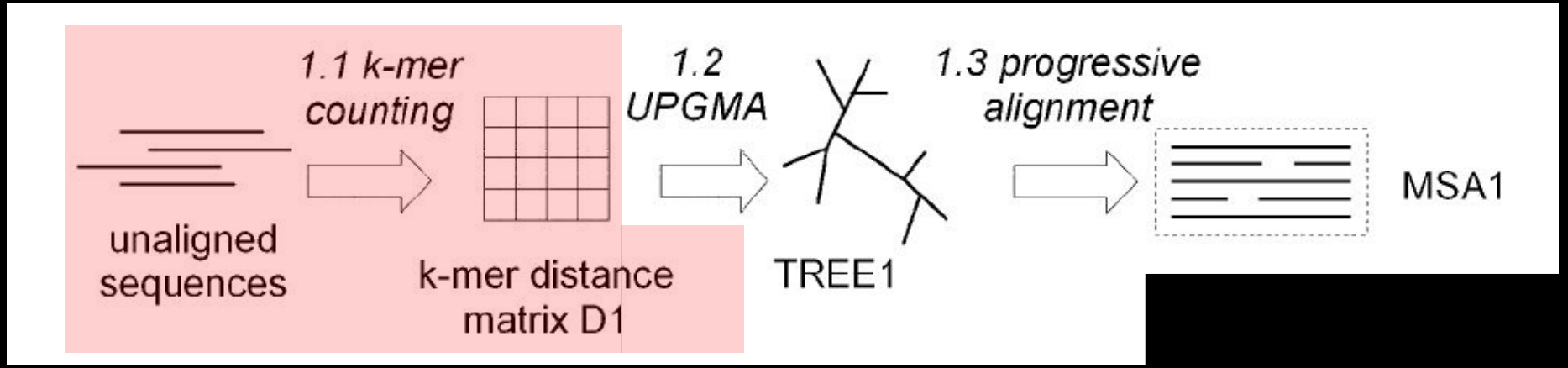
Unaligned sequences to *k*-mers
k-mer similarity for a pair of sequences:

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

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

Normalizing constant
(length of the shorter sequence)

MUSCLE Step 1

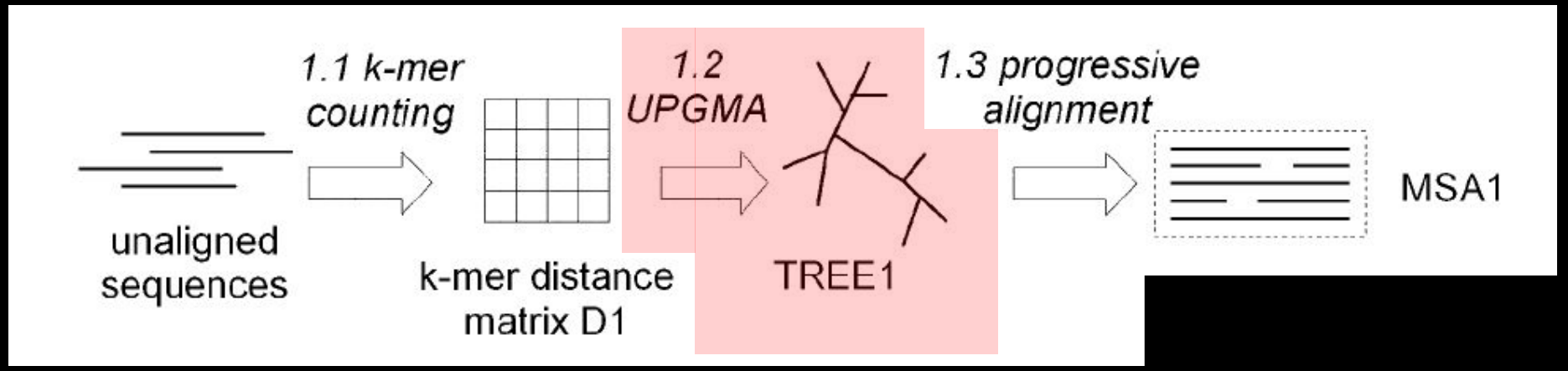


We convert F to a distance measure:

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

And populate a triangular distance matrix D1 with d_{kmer} values

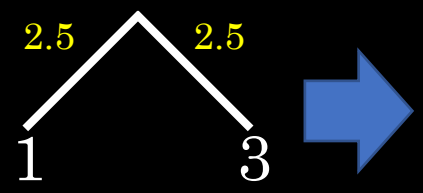
MUSCLE Step 1



UPGMA: Unweighted Pair Grouping with Arithmetic Mean

	1	2	3
1			
2	10		
3	5	12	

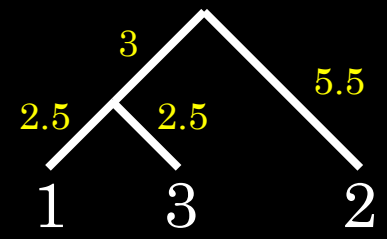
Identify smallest value $d_{i,j}$ in D matrix



Make i and j sisters in the tree

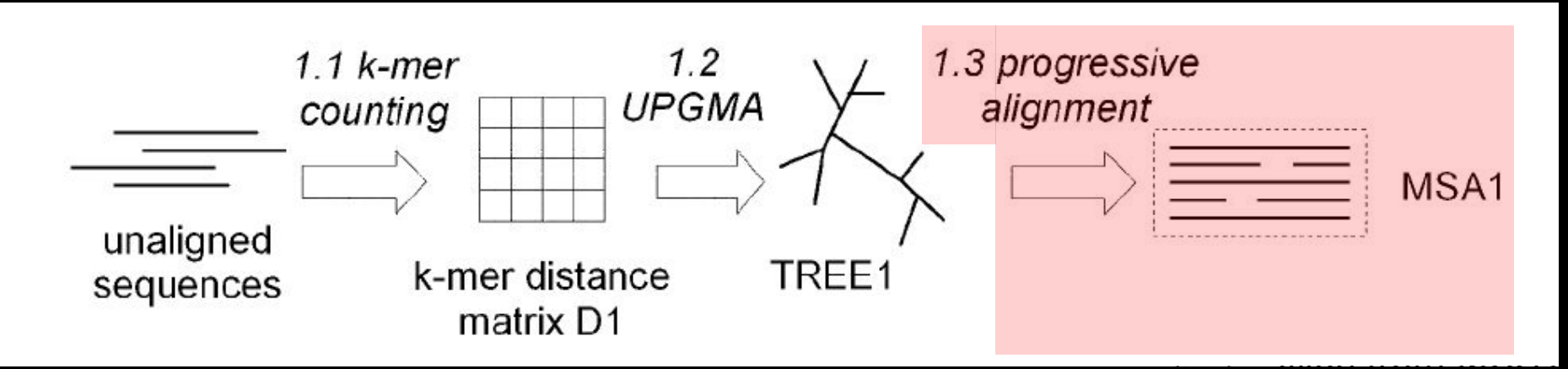
	1+3	2
1+3		
2	11	

Recalculate matrix, find smallest value

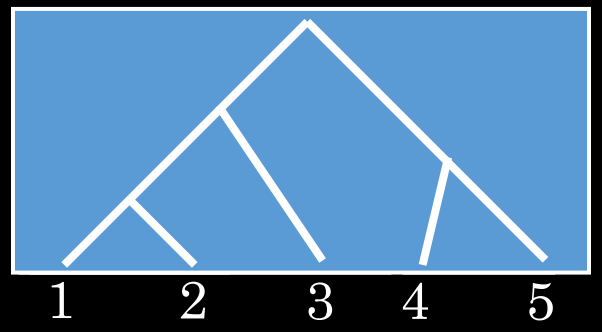


Final tree

MUSCLE Step 2



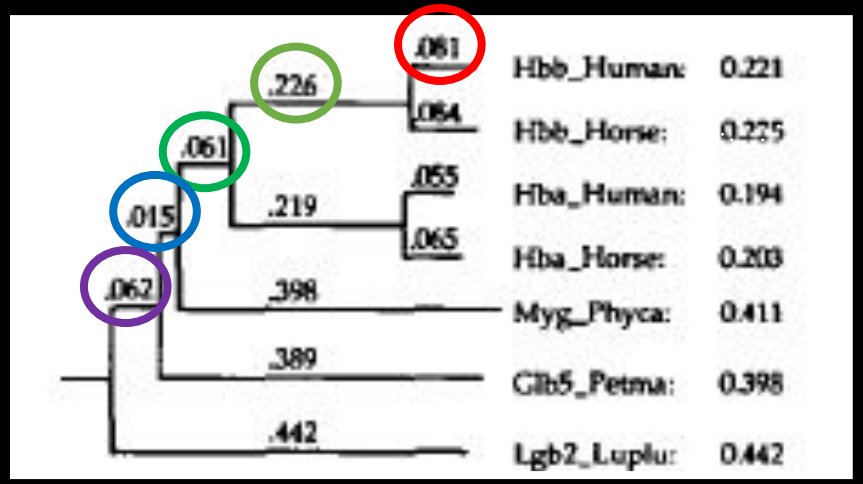
Progressive alignment based on the UPGMA 'guide' tree:
Align sequences and profiles in prefix order based on the tree



Each pairwise alignment is done with dynamic programming

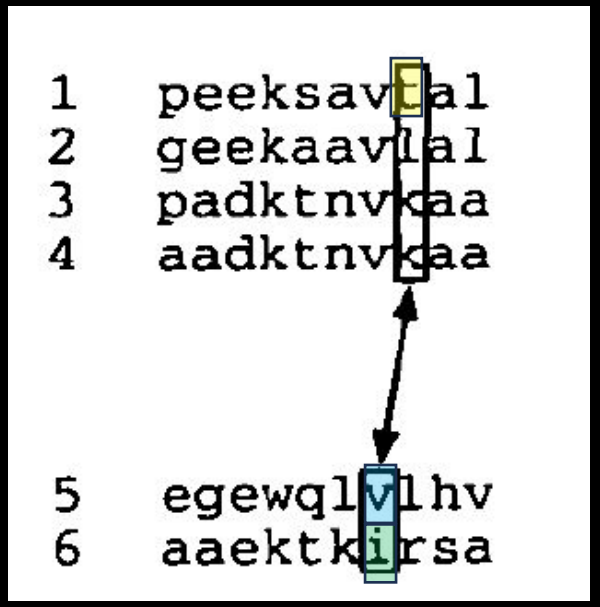
But we only need to do $4n^2$ operations instead of n^5

How to align profiles



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

Weight sequences
by branch independence

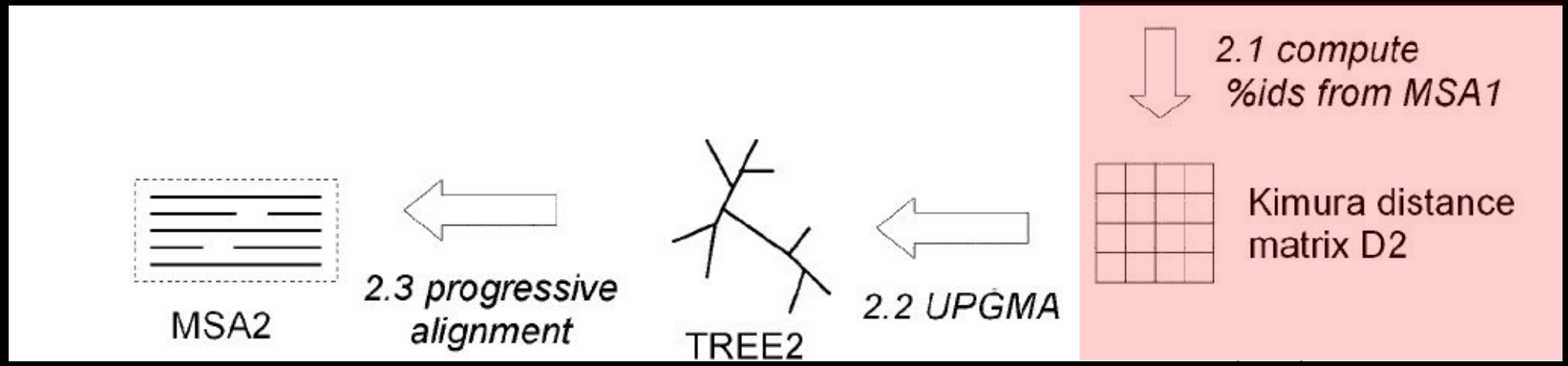


Score matches based on
weights and scoring matrix

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

Thompson et al. (1994) *Nucleic Acids Res*

MUSCLE Step 2



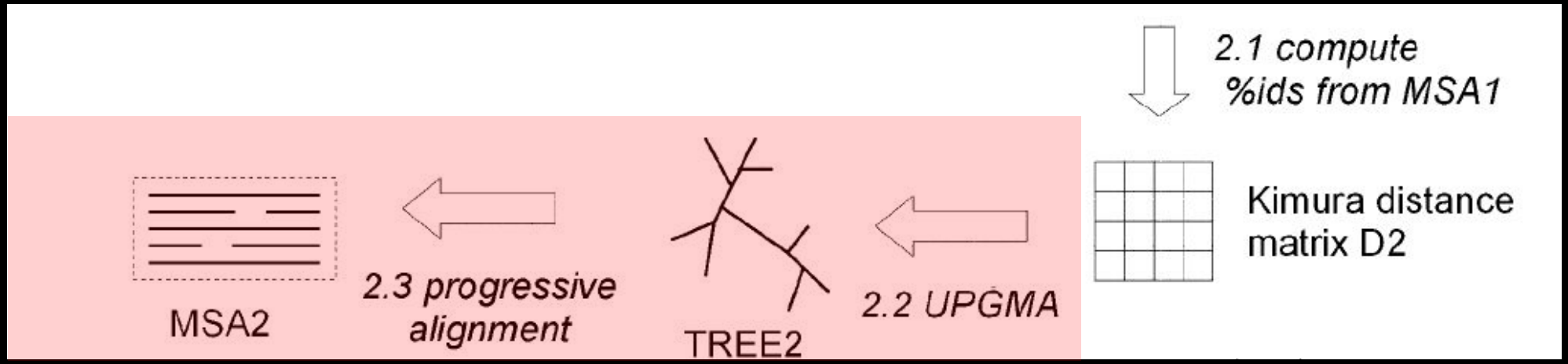
The distances used to build the initial guide tree were very crude

MUSCLE uses the first sequence alignment to compute Kimura distances:

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

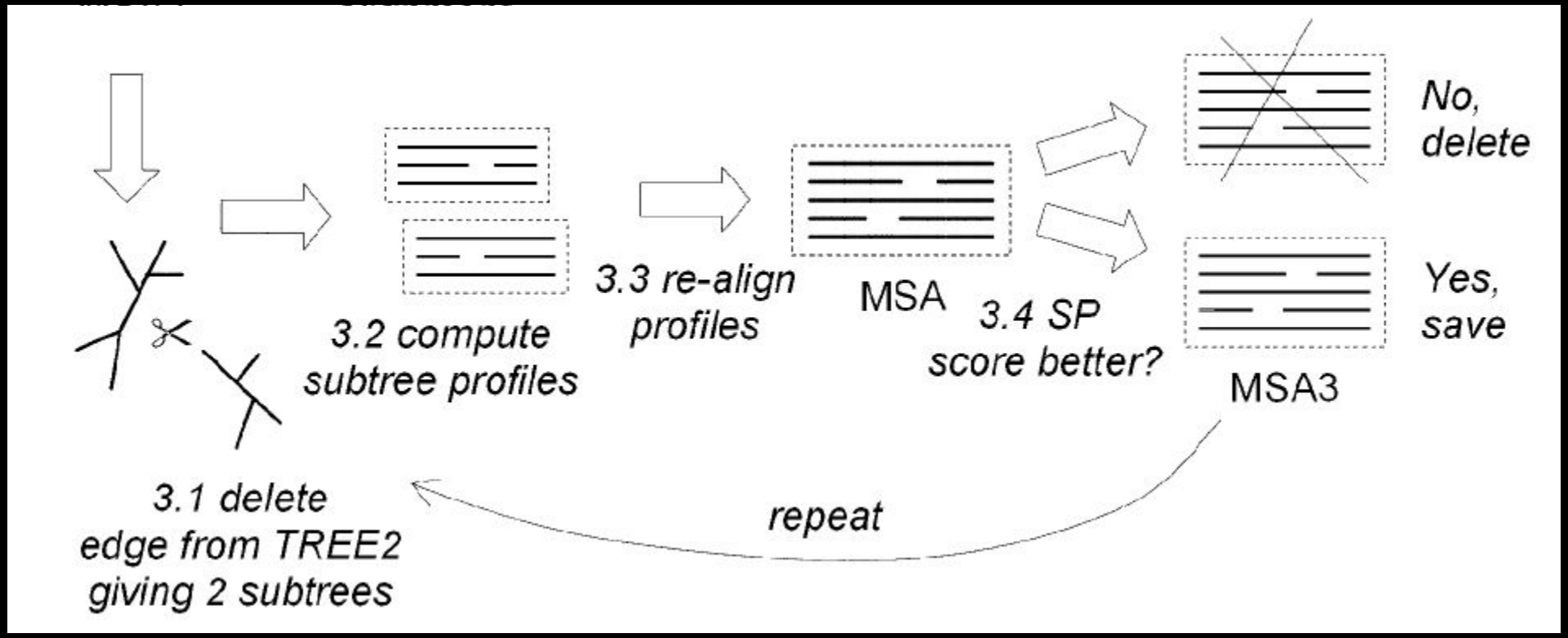
Multiple substitutions!

MUSCLE Step 2



(Better) tree and (better) progressive alignment than before

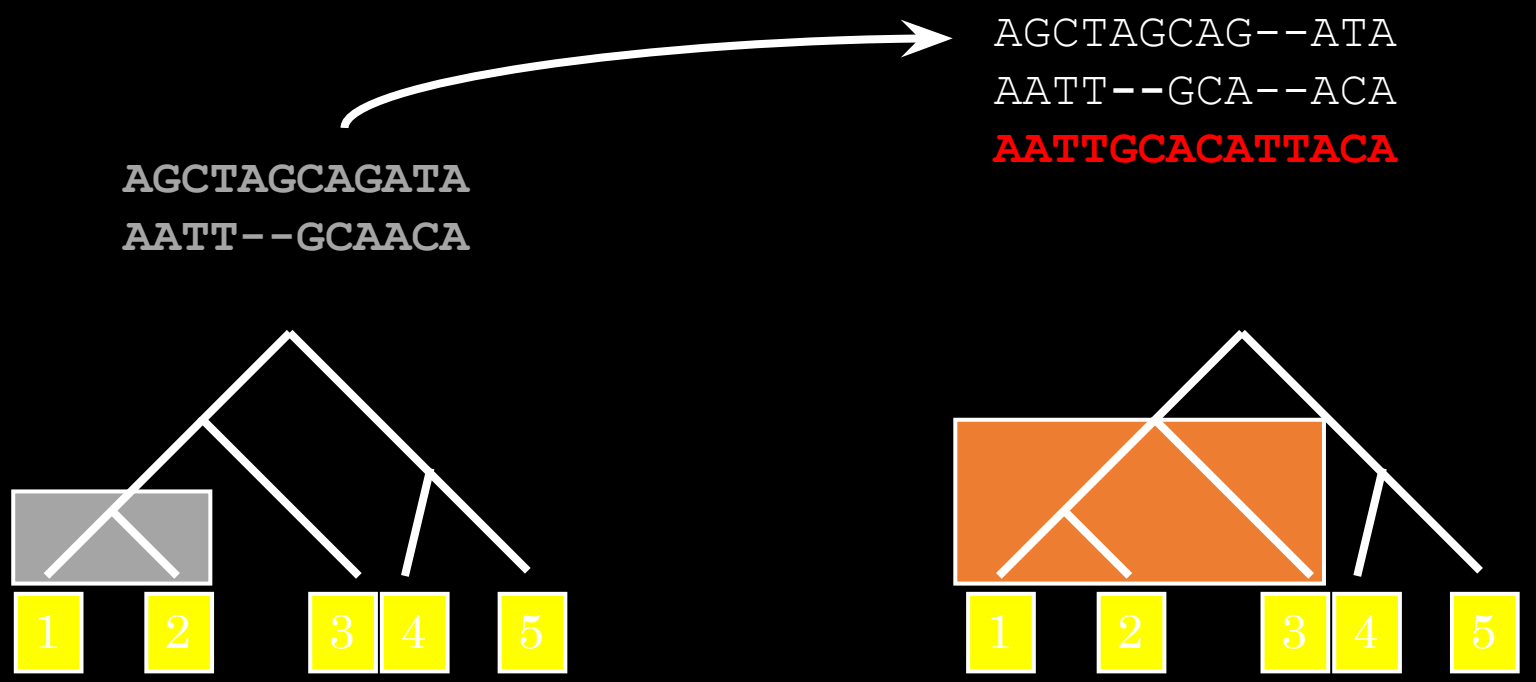
MUSCLE Step 3



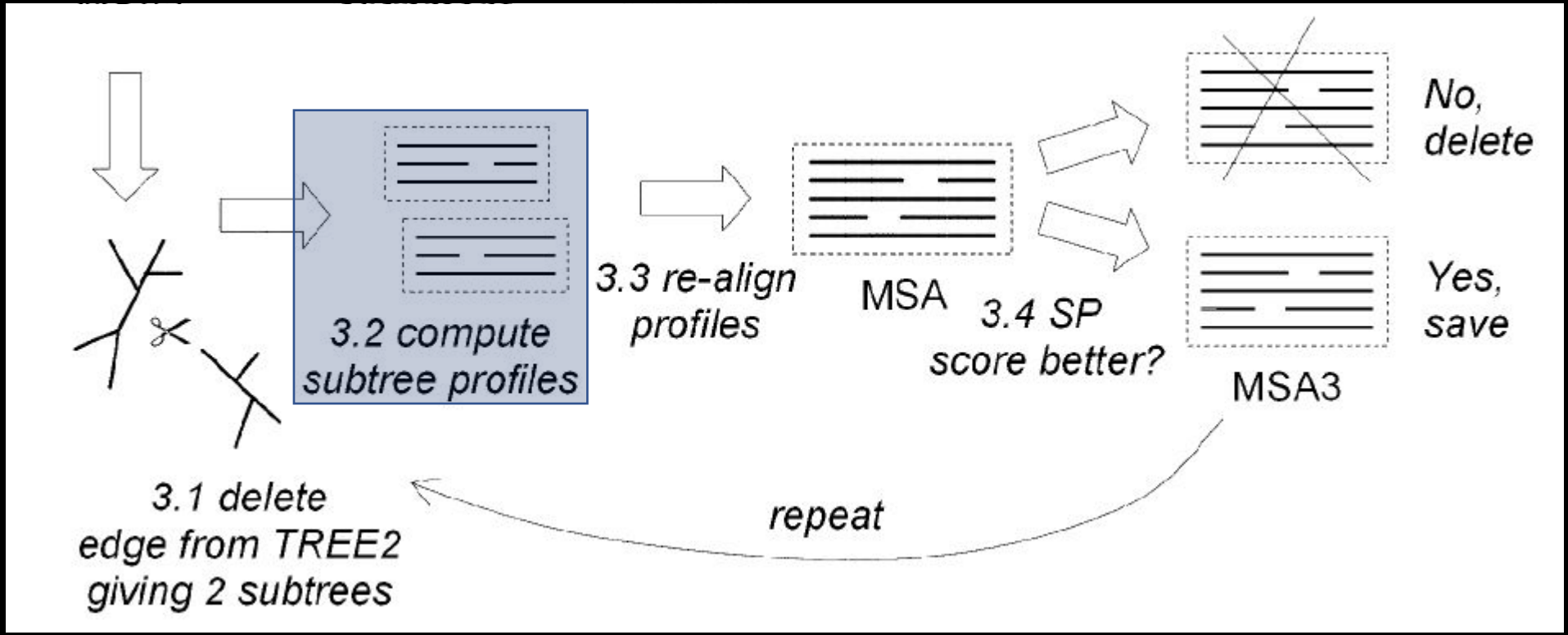
Why do we do this?

The classic limitation of progressive alignment

- "once a gap, always a gap"



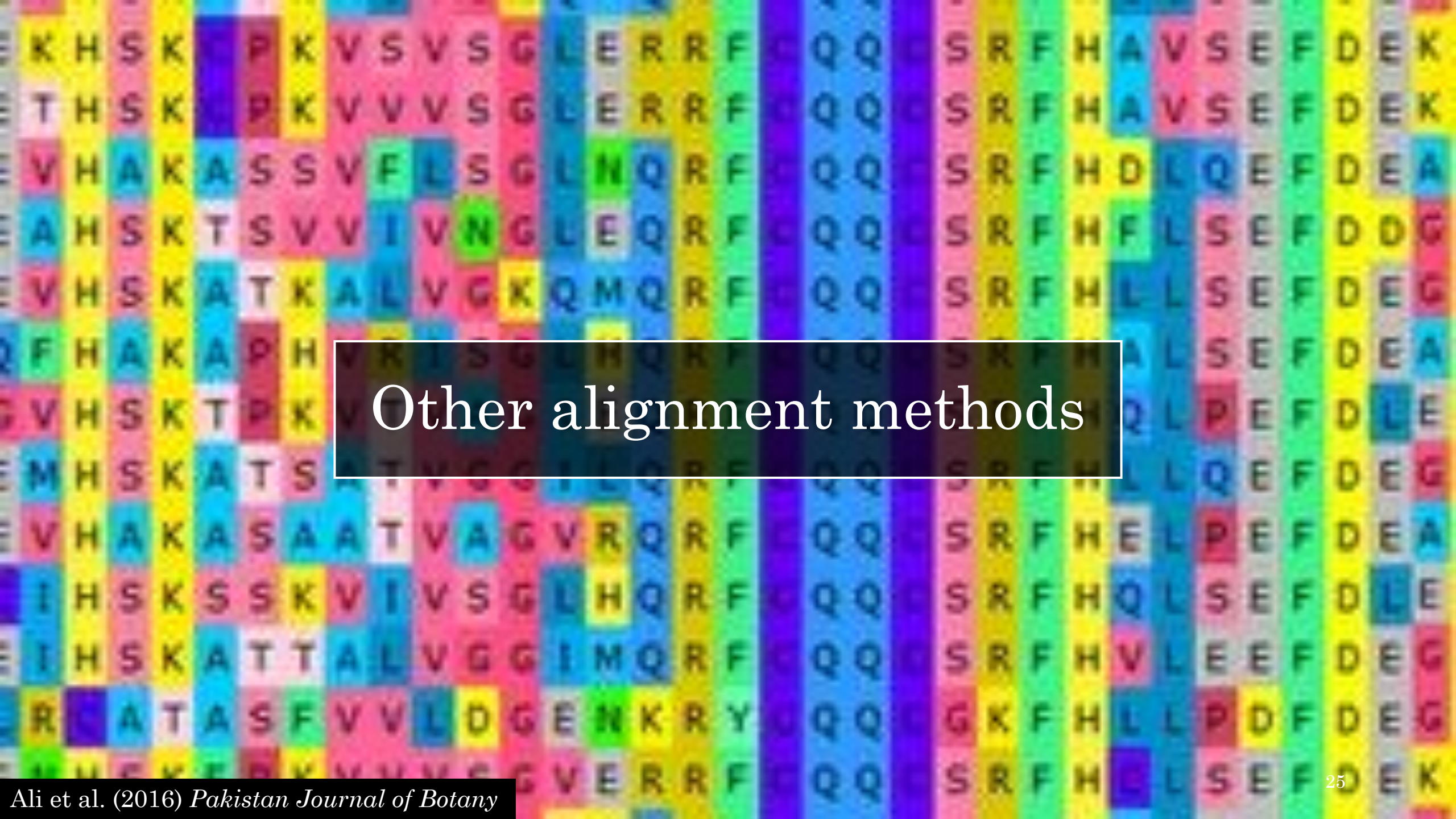
By breaking a branch of the guide tree, removing all gap-only columns and realigning the two profiles, we may find a better alignment



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*

Key Ideas

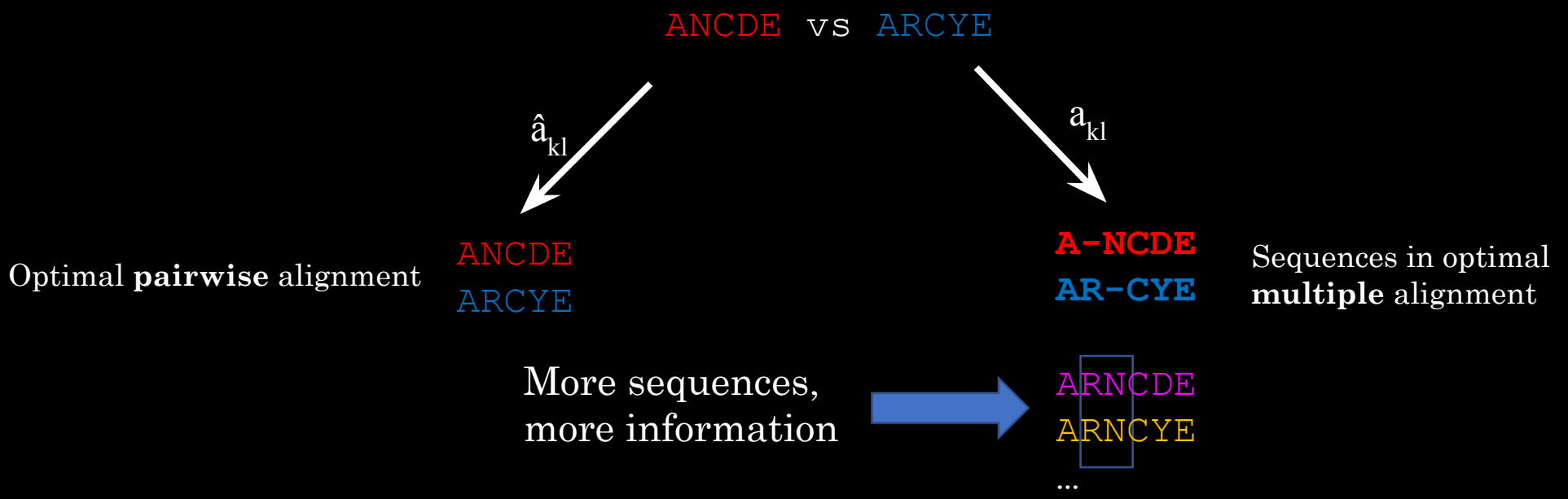
- Progressive alignment
- Probabilities
- Consensus between different alignments of the same sequences
- Efficient sequence comparisons / representations

Consensus Alignment: Combining information from multiple sources

The best alignment between a pair of sequences may not appear in the optimal multiple alignment

- AND -

Different algorithms may produce different alignments of the same sequences



T-COFFEE

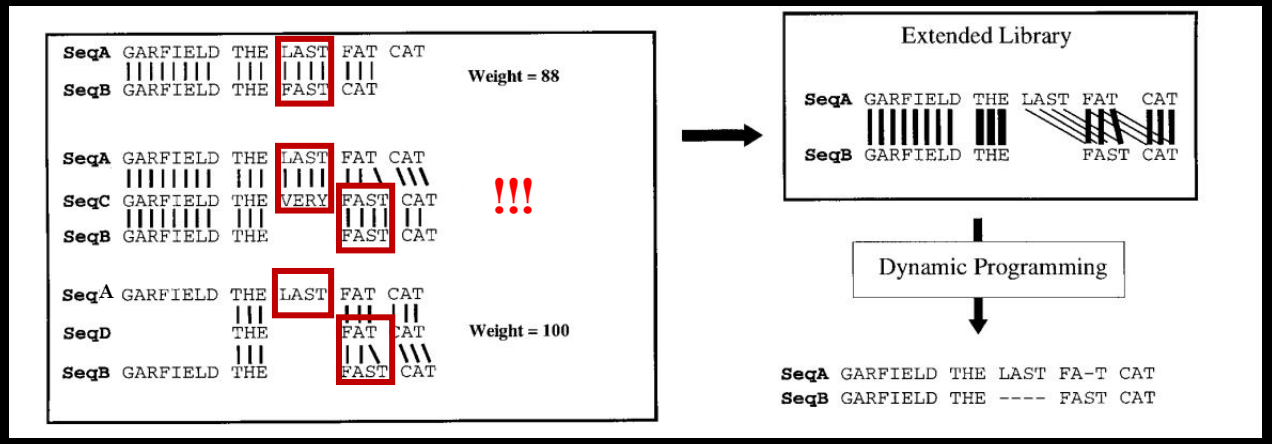
Tree-based Consistency Objective Function for alignment Evaluation

and other consensus-based methods

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

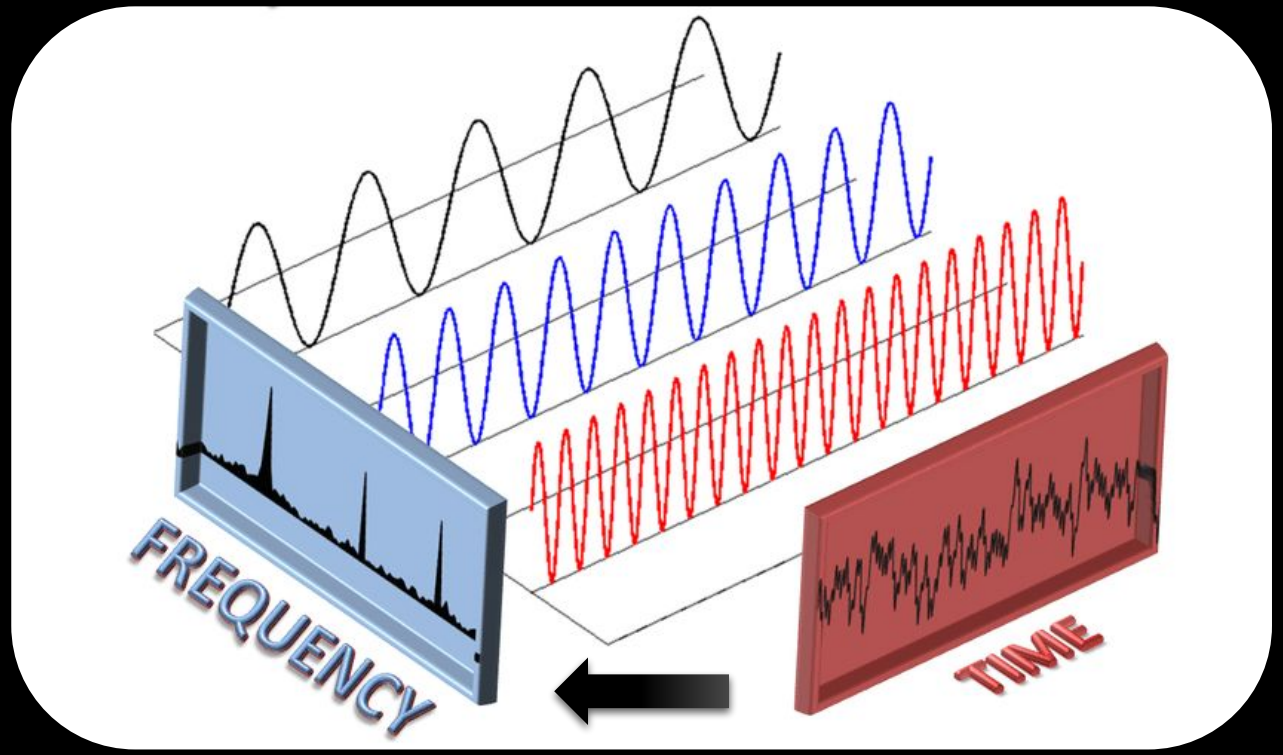
All pairwise alignments (LAST, FAST highlighted)

Consensus information



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

MAFFT – Multiple Alignment using the Fast Fourier Transform



Kalhara et al. (2017) *SKIMA*

Discrete Fourier Transform: decompose a time signal into its frequency spectrum

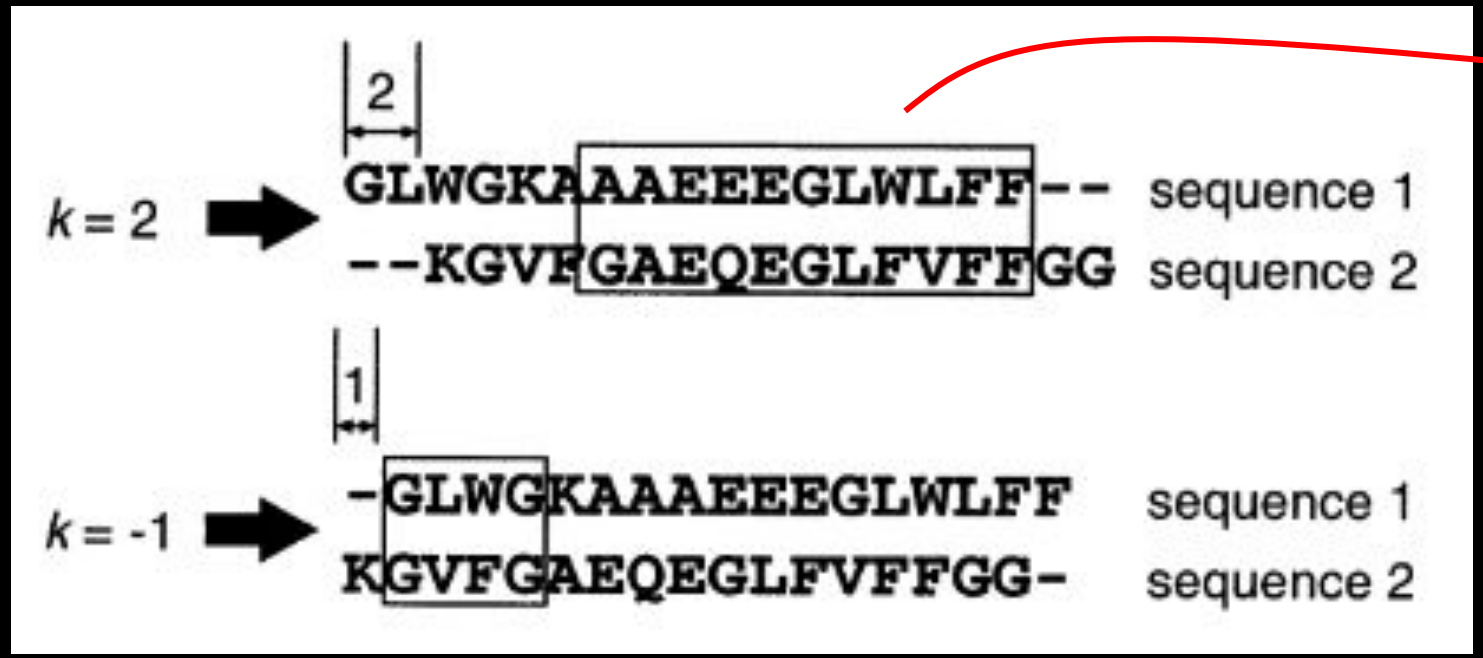
Fast Fourier Transform: Do this quickly ($n \log n$ vs n^2)

Katoh et al. (2002) *Nucleic Acids Res*
List of MAFFT papers: https://mafft.cbrc.jp/alignment/software/

MAFFT

Multiple alignment using fast Fourier transform

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



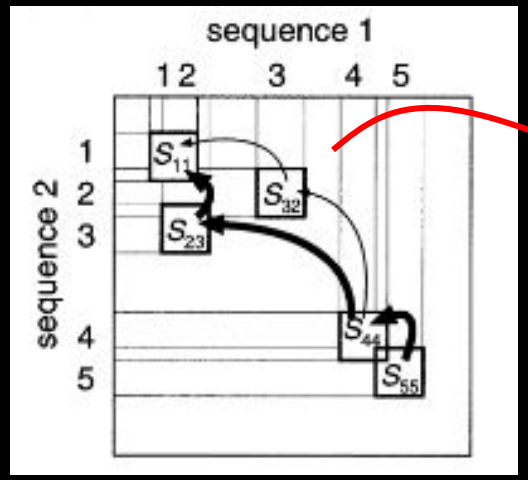
High size/polarity score: Candidate homologous region with offset 2

FFT: Do this for all possible offsets quickly

MAFFT

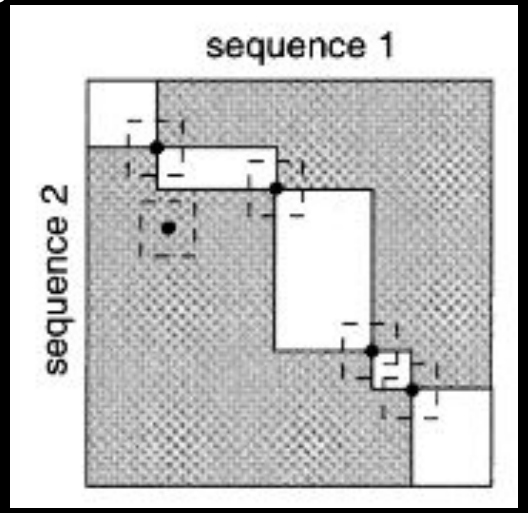
Multiple alignment using fast Fourier transform

3. Use candidate homologous regions as anchor points for DP



Five homologous regions in DP matrix

4. Progressive alignment



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 best alignment:

```
...NQK...
...IDL...
```

A bunch of alignments that are almost as good:

```
...NQK-...
...-IDL...
```

Do et al. (2005) *Genome Res*

ProbCons

probabilistic consistency-based alignment

- Calculate these probabilities using a hidden Markov model (coming soon)

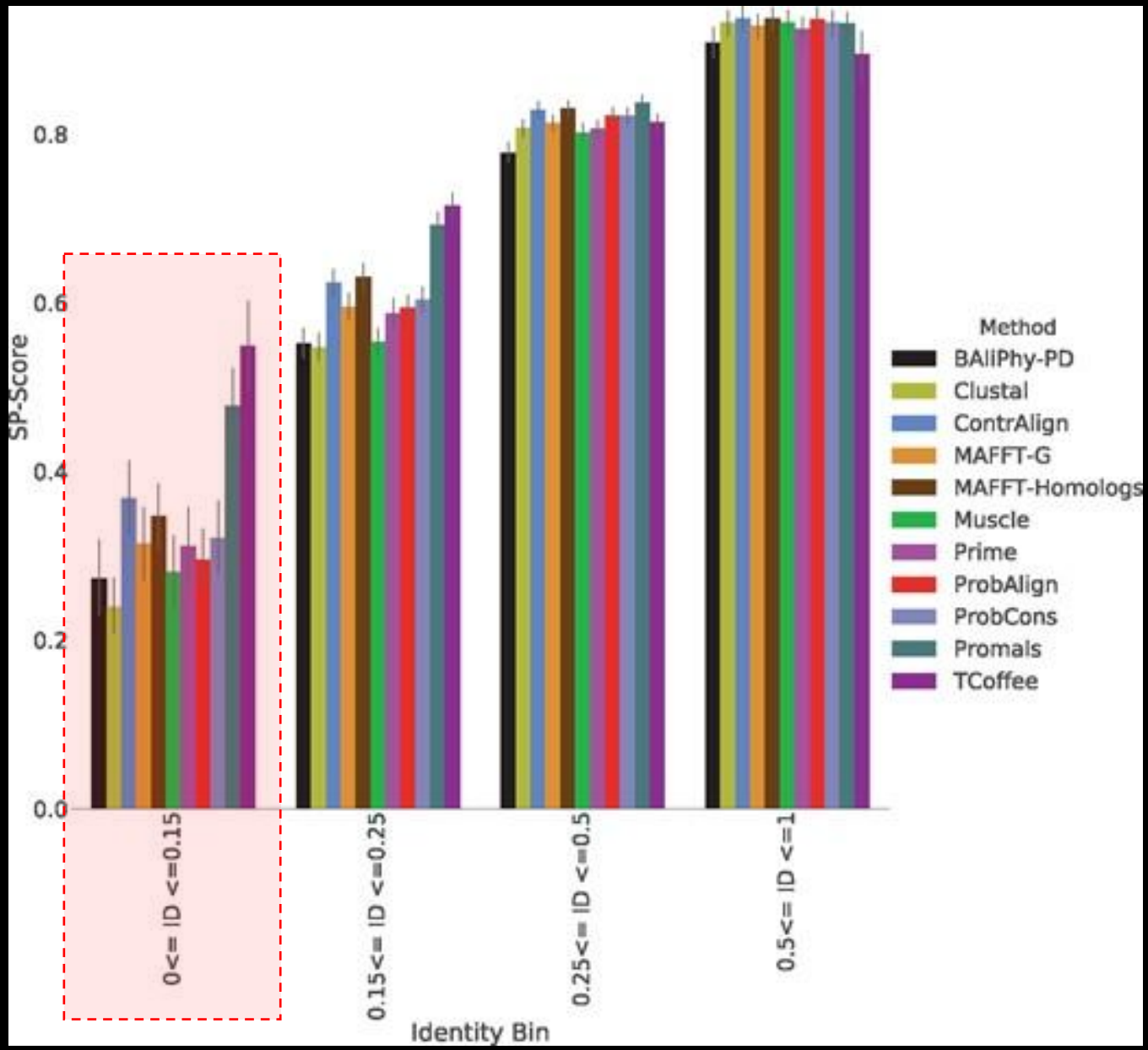
- Replace the PAM matrix score for a pair of amino acids with their cumulative probability across all alignments, then do dynamic programming!

Do et al. (2005) *Genome Res*

Comparison

Compare against benchmark databases

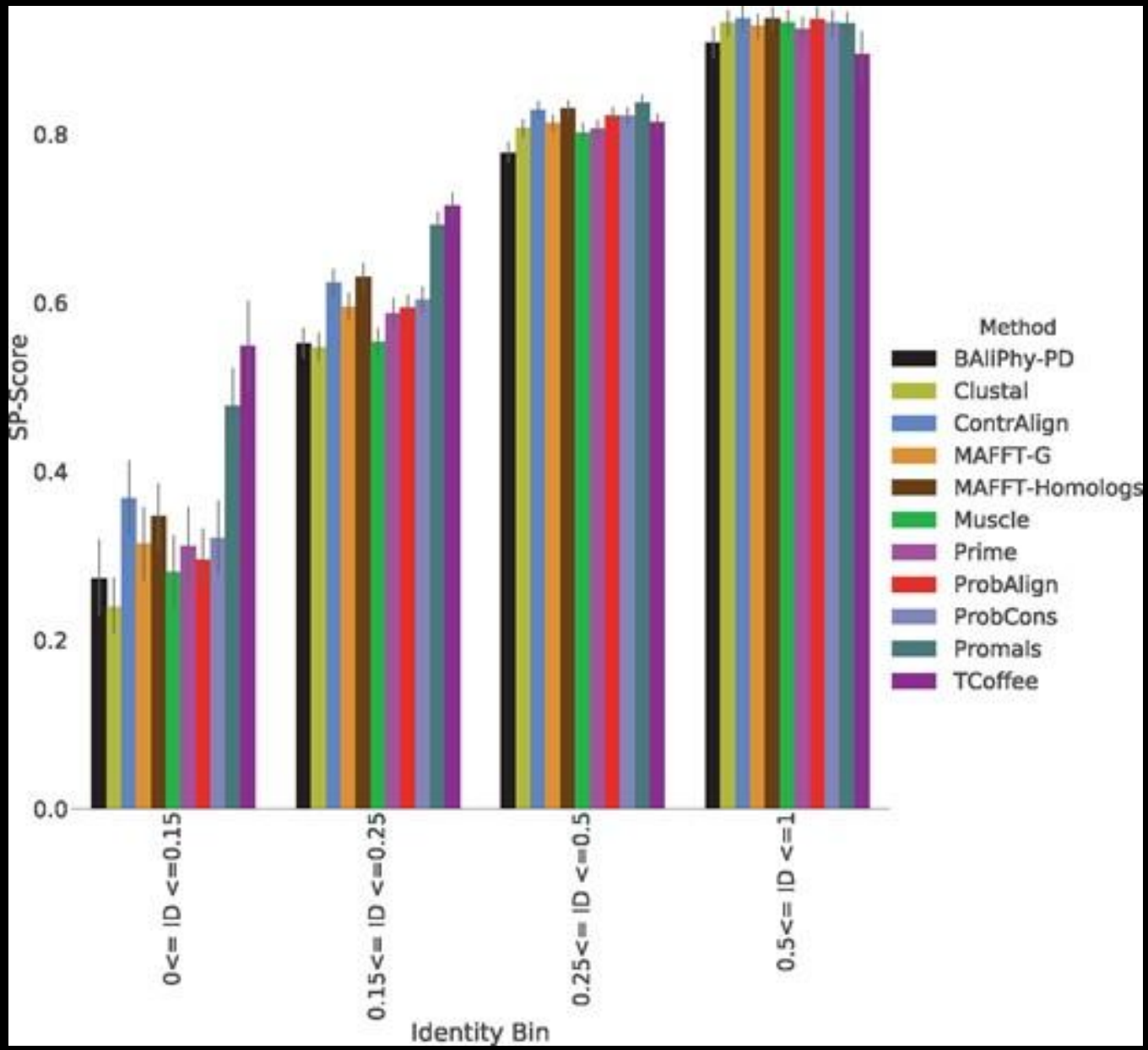
Identity bins: lower %ID = most difficult



Nute et al. (2019) *Syst Biol*

Comparison

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
Clustal	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



Nute et al. (2019) *Syst Biol*

So

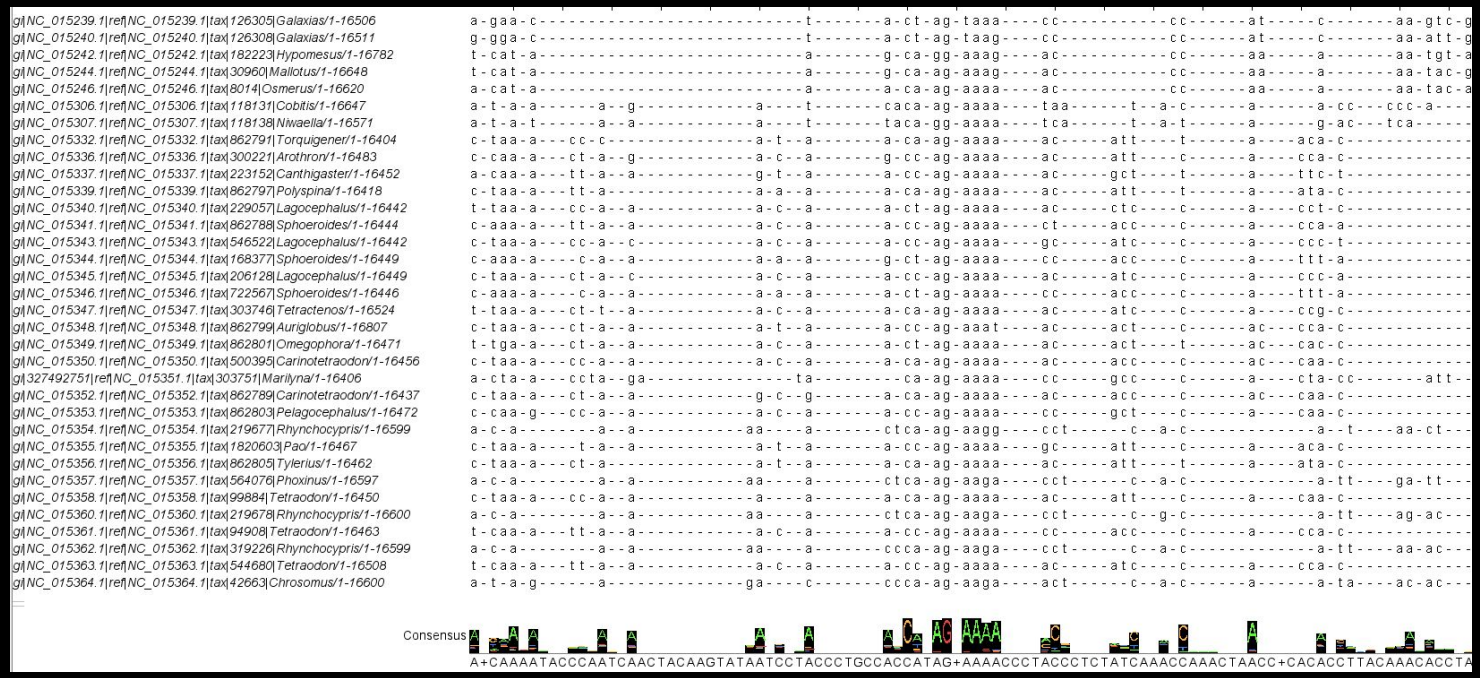
- Which alignment program should you use?

- Considerations:
  - What other people use
  - Accuracy / speed tradeoff
  - Working implementation on your system
  - Anything else?

Also

- Be very careful with sequence alignments, especially if they are large. Inspect!

4213 sequences
IMMENSE number of gaps



Mohamed Fares (last week)

People Are Still Trying

[PDF] A Hybrid Approach using Progressive and Genetic Algorithms for Improvements in Multiple Sequence Alignments. [PDF] scitepress.org
GFD Zafalon, VZ Gomes, AR Amorim, CR Valêncio - ICEIS (2), 2021 - scitepress.org
... Our results show that our method is able to improve the quality of the alignments of all families from **BAliBase**. Considering Q and TC quality measures from **BaliBase**, we have obtained ...
☆ Save 99 Cite Cited by 1 Related articles All 5 versions »

PC_ali: a tool for improved multiple alignments and evolutionary inference based on a hybrid protein sequence and structure similarity score [PDF] oup.com FindIt @ Dalhousie
U Bastolla, D Abia, O Piette - Bioinformatics, 2023 - academic.oup.com
... We assess these MSAs against eight multiple alignment programs on the **Balibase** set of ... between the **Balibase** and the PDB sequences, either because **Balibase** omits residues whose ...
☆ Save 99 Cite Related articles All 7 versions

Particle swarm optimization with tabu search algorithm (PSO-TS) applied to multiple sequence alignment problem
L Chaabane, A Khelassi, A Terziev... - ... , Modeling and Findings ..., 2021 - Springer
... Numerical experimental outcomes on some **BaliBASE** benchmark instances confirmed the capability of the PSO-TS approach to producing better by-products while comparing it to other ...
☆ Save 99 Cite Cited by 12 Related articles All 3 versions

A Novel Population-based Optimization for Multiple Sequence Alignment in Protein Sequencing [PDF] espublisher.com
A Goswami, KK Dubey - Engineered Science, 2022 - espublisher.com
... from the **Bali base** benchmark database in order to ensure that it was effective over a broad range of datasets from the **Bali base** benchmark database. The **Bali base** version 1.0 has a ...
☆ Save 99 Cite Cited by 4 Related articles »

[PDF] Characterization of Protein Sequences Aligned with MUSCLE using Guide Trees from SARELI [PDF] academia.edu
A Chavoya, R Ortega - academia.edu
... The **BAliBASE** database was manually designed as an evaluation resource for addressing ... sequence sets, followed by SABRE, and with **BAliBASE** having the smallest number of files. ...
☆ Save 99 Cite Related articles »

An effective cooperative aligner to resolve multiple-sequence alignment problem
L Chaabane - International Journal of Cloud Computing, 2021 - inderscienceonline.com
In this research work, we propose a new cooperative aligner based on metaheuristics to find an approximate solution to the multiple-sequence alignment (MSA) problem. The ...
☆ Save 99 Cite Cited by 6 Related articles All 2 versions

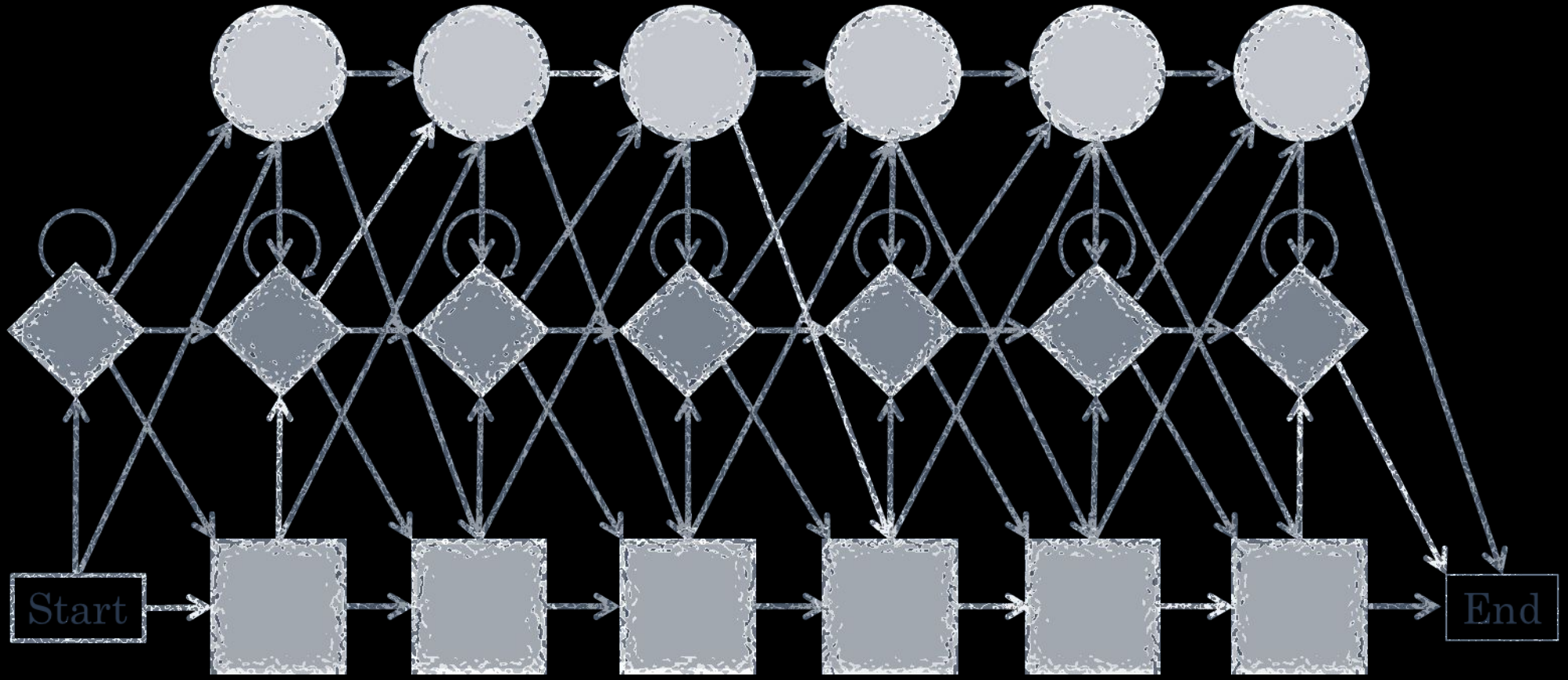
An enhanced cooperative method to solve multiple-sequence alignment problem
L Chaabane - ... Journal of Data Mining, Modelling and ..., 2021 - inderscienceonline.com

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

Hidden Markov Models and Gene Prediction



Overview

- Sequence profiles
- How hidden Markov models work
- Training HMMs
- HMMer
- Other applications of HMMs

K-ELQRAASLTIEV

KDEGQK--SLVIDV

If we have an alignment…

…what can we do with it?

For many questions, we would like to know the distribution of residues (and gaps) in a block of sequences

Position-specific scoring matrices (PSSM)

PAM, BLOSUM, etc. are position-independent scoring matrices

A PSSM is a log-odds matrix of column frequencies

CGGCCT
CGAGCT
GATGCA
AAAGCA
ATAGCA
TCTACT
AACATC
TACGCC
AACGAG
AGCTGT

Background frequencies:
A = 19/60 = 0.317
C = 17/60 = 0.283
G = 12/60 = 0.2
T = 12/60 = 0.2

Frequency Matrix

	1	2	3	4	5	6
A	0.5	0.5	0.3	0.2	0.1	0.3
C	0.2	0.1	0.4	0.1	0.7	0.2
G	0.1	0.3	0.1	0.5	0.1	0.1
T	0.2	0.1	0.2	0.2	0.1	0.4

Background frequencies:
A = 19/60 = 0.317
C = 17/60 = 0.283
G = 12/60 = 0.2
T = 12/60 = 0.2

Frequency Matrix

	1	2	3	4	5	6
A	0.5	0.5	0.3	0.2	0.1	0.3
C	0.2	0.1	0.4	0.1	0.7	0.2
G	0.1	0.3	0.1	0.5	0.1	0.1
T	0.2	0.1	0.2	0.2	0.1	0.4

$\log_n$-odds matrix ($n = e$)

	1	2	3	4	5	6
A	0.66	0.66	-0.08	-0.66	-1.66	-0.08
C	-0.5	-1.5	0.5	-1.5	1.31	-0.5
G	-1	0.58	-1	1.32	-1	-1
T	0	-1	0	0	-1	1

Background frequencies:
A = 19/60 = 0.317
C = 17/60 = 0.283
G = 12/60 = 0.2
T = 12/60 = 0.2

Frequency Matrix

	1	2	3	4	5	6
A	0.5	0.5	0.3	0.2	0.1	0.3
C	0.2	0.1	0.4	0.1	0.7	0.2
G	0.1	0.3	0.1	0.5	0.1	0.1
T	0.2	0.1	0.2	0.2	0.1	0.4

$\log_n$-odds matrix ($n = e$)

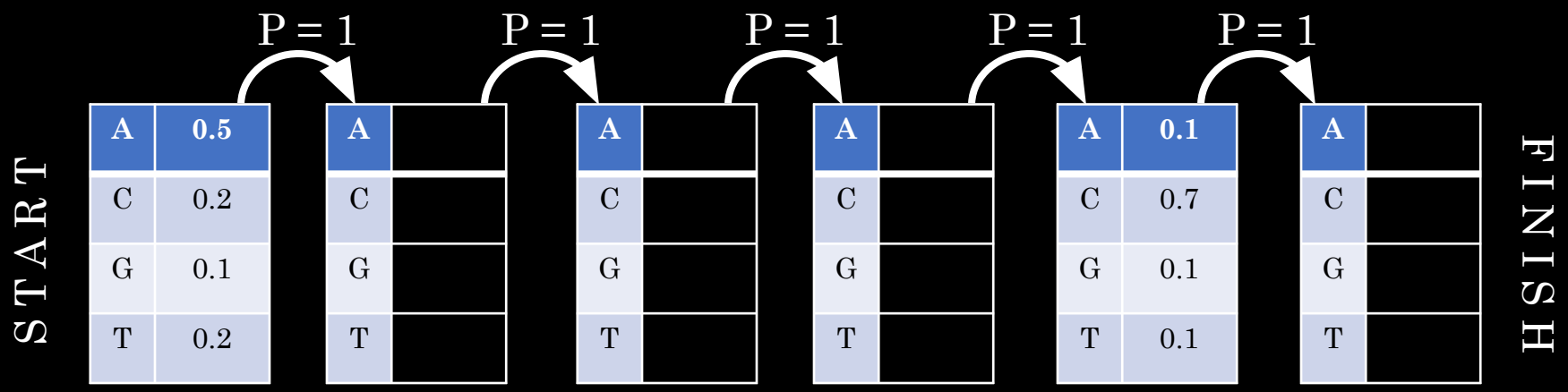
	1	2	3	4	5	6
A	0.66	0.66	-0.08	-0.66	-1.66	-0.08
C	-0.5	-1.5	0.5	-1.5	1.31	-0.5
G	-1	0.58	-1	1.32	-1	-1
T	0	-1	0	0	-1	1

Aligning a sequence against log-odds matrix:
Add scores for residue at each position, then take n^{sum}

AGAGGT

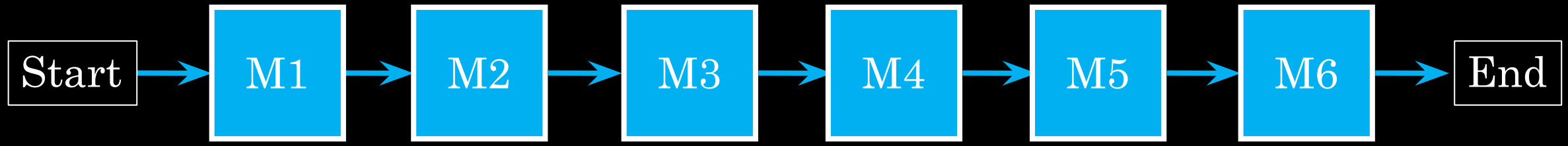
How do we represent insertions and deletions?

Transitions in a Probability Matrix

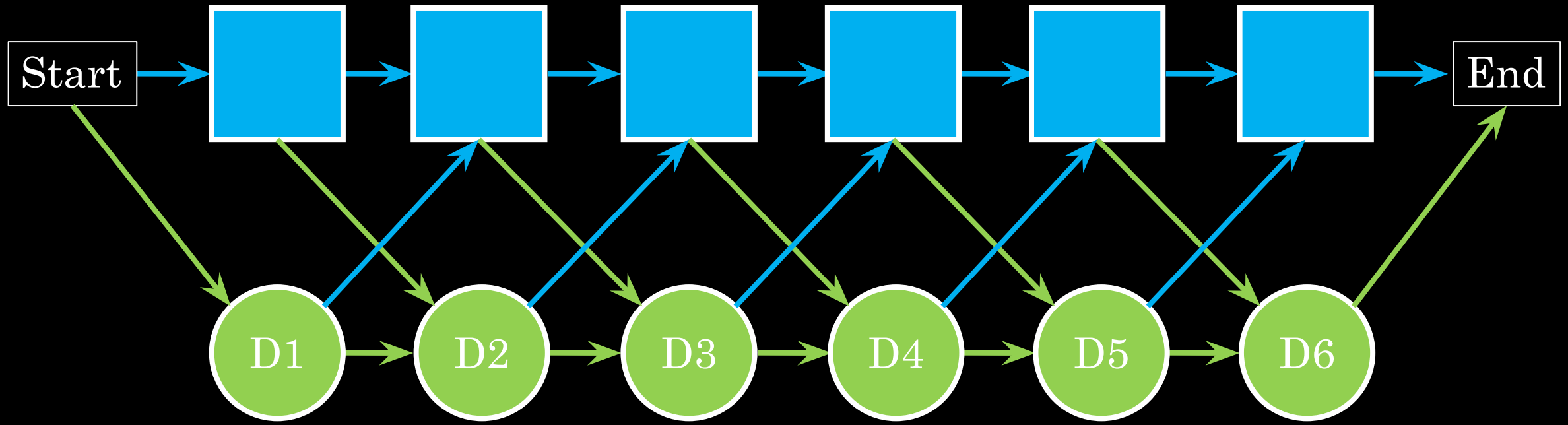


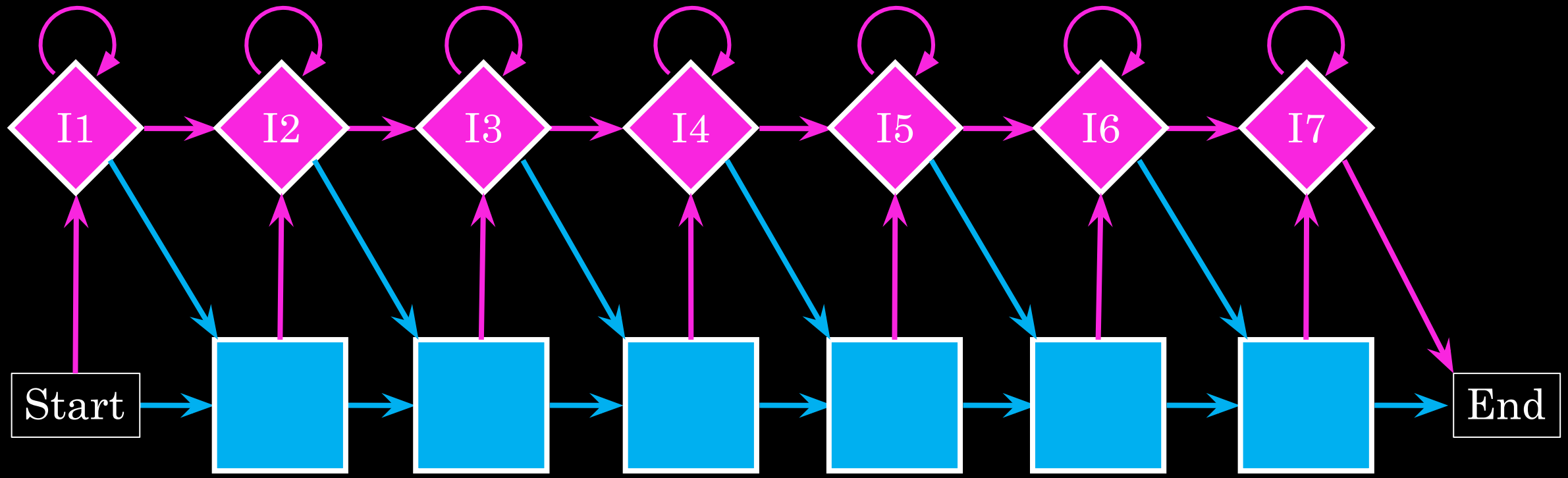
Transition from match state k to match state $k + 1$ with probability 1.0

Abstraction – Match States

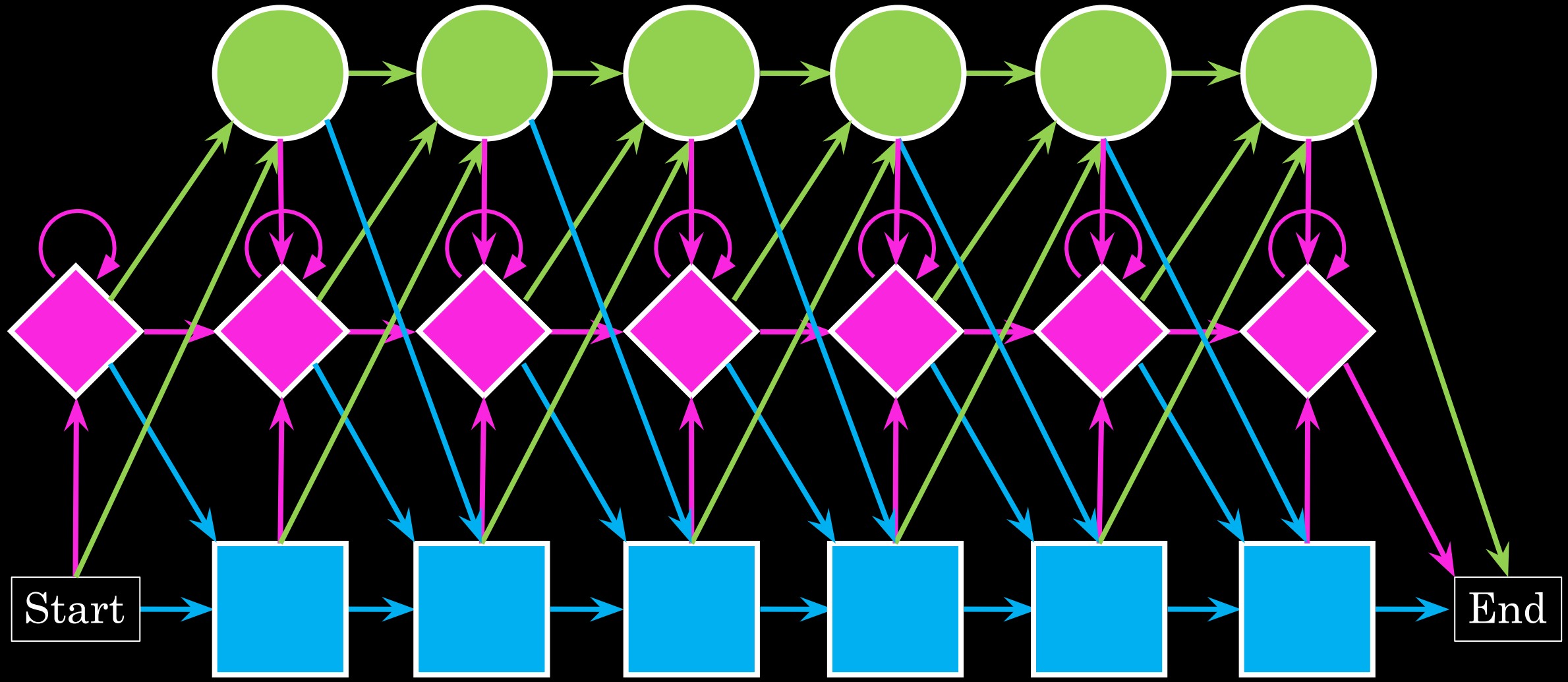


Deletion States



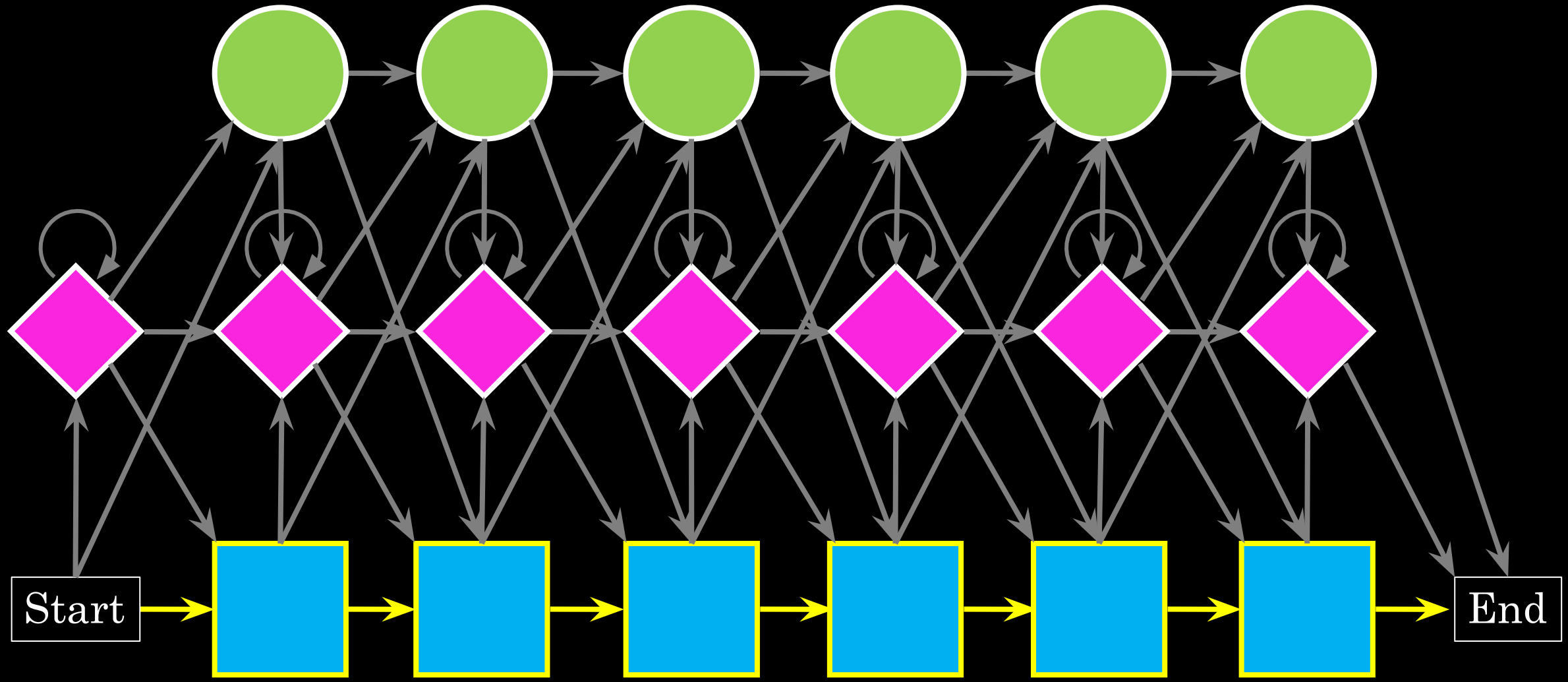


Insertion States



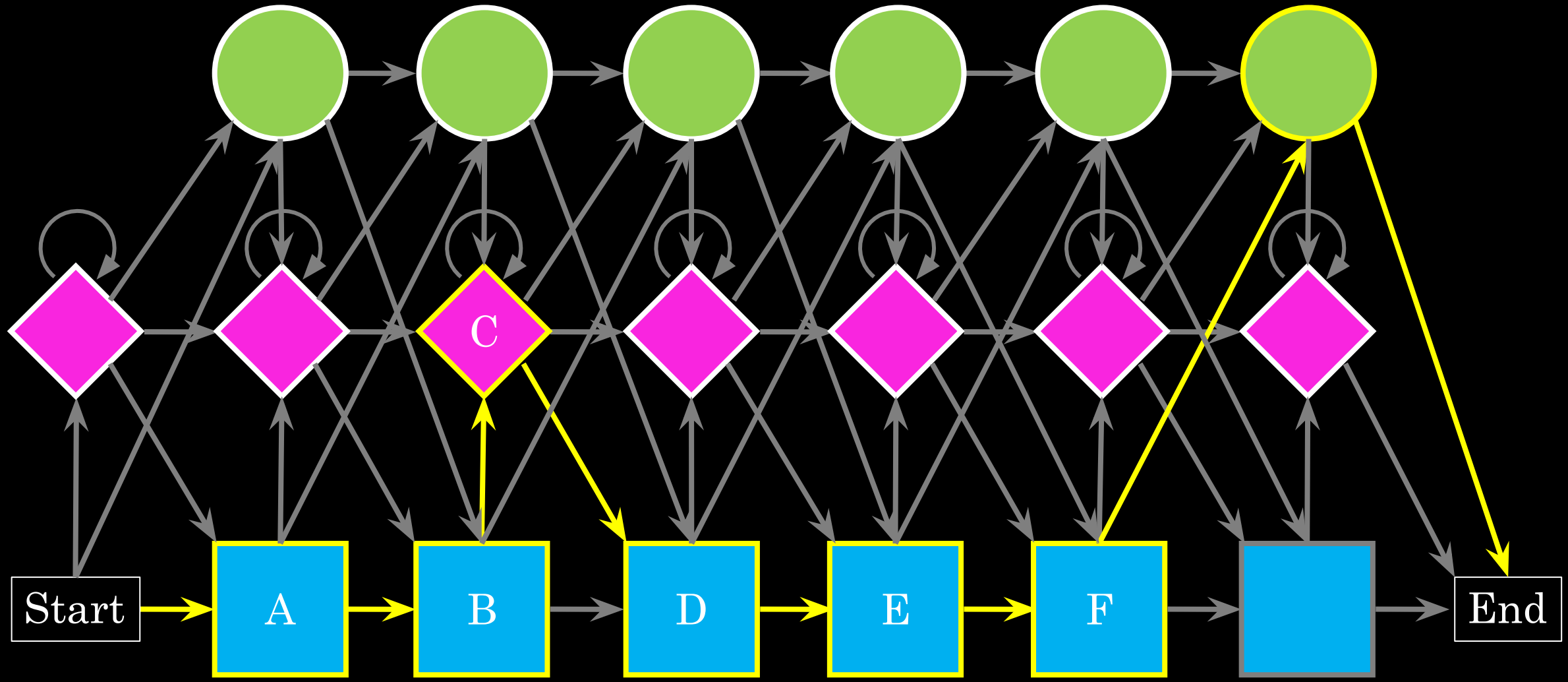
The Full Model

Aligning sequences to the model

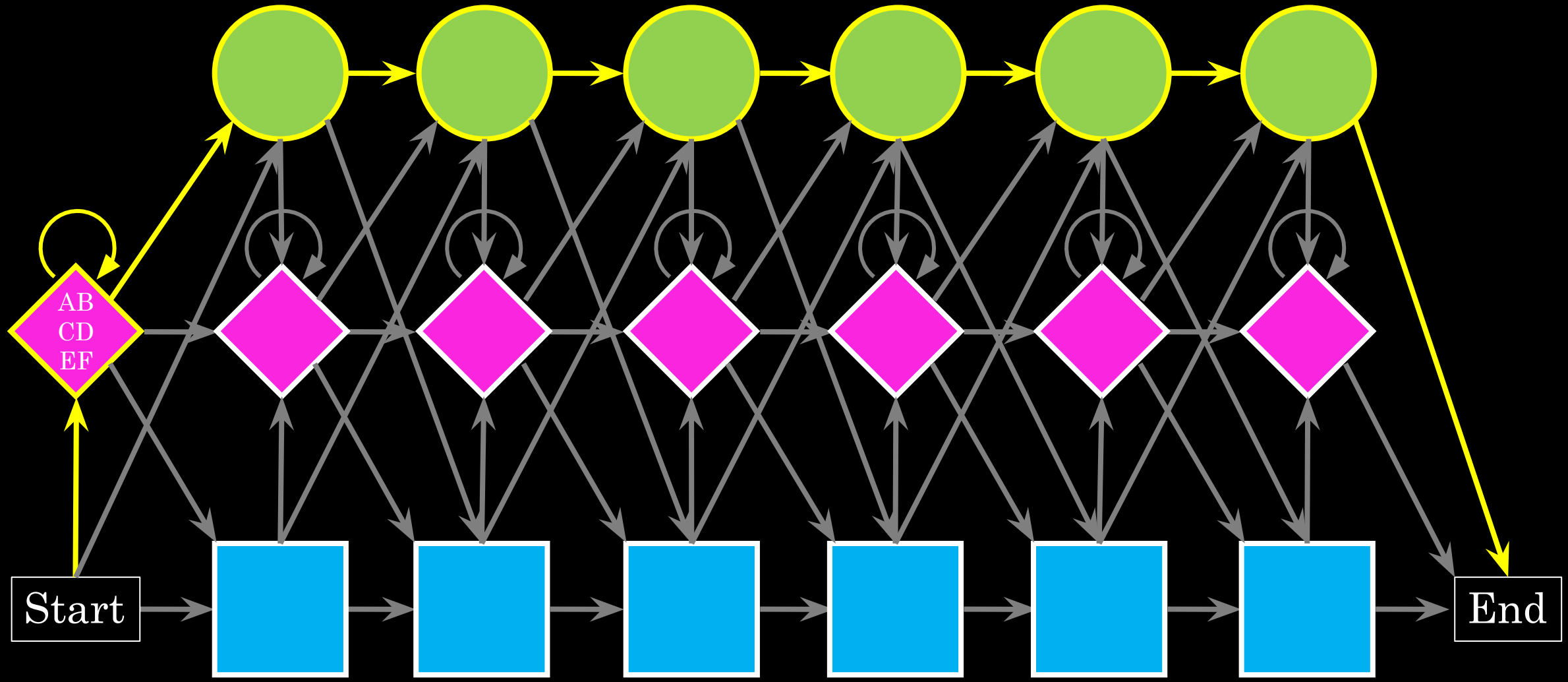


ABCDEF (similar to PSSM)

Aligning sequences to the model



ABCDEF (C is inserted,
no character homologous to match state 6)



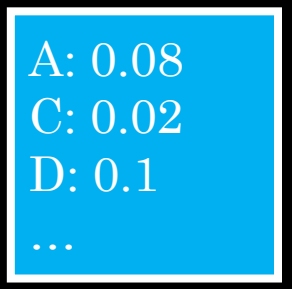
ABCDEF (nothing is homologous)

This is a hidden Markov model

- Hidden: We don't know the correct alignment of a sequence to the model
- Markov: Future probabilities are not dependent on previous ones
- Model: It's … a model

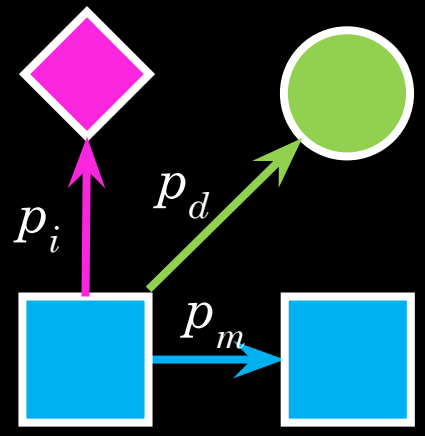
Key components of an HMM

- EMISSIONS: A character (nucleotide or amino acid) produced by a given insertion or match state



emission probabilities
(equivalent to values in a PSSM)

- TRANSITIONS: The probability of going from state *i* to state *j* (sum of all transitions from a given state = 1)



$p_i + p_d + p_m = 1$

The product of the emission probabilities e and the transition probabilities a through the model

=

The joint probability of the *sequence* x and the *path* π

The sum of the log(emission probabilities) e and the log(transition probabilities) a through the model

=

The joint probability of the *sequence* x and the *path* π

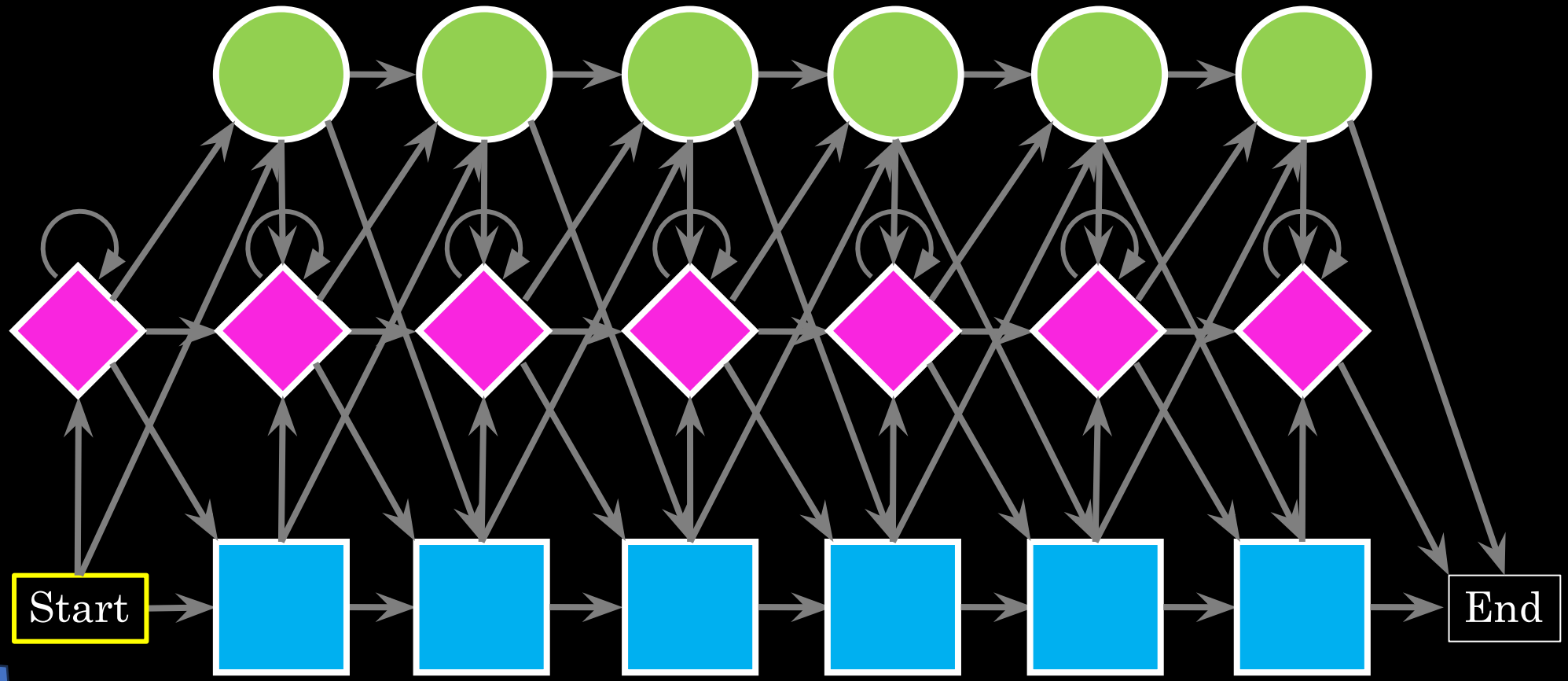
Avoids tiny numbers!

Best path

- There are many paths π through the model for any given sequence x
- We can use the Viterbi algorithm to find the path π^* with the highest probability

The Viterbi Algorithm

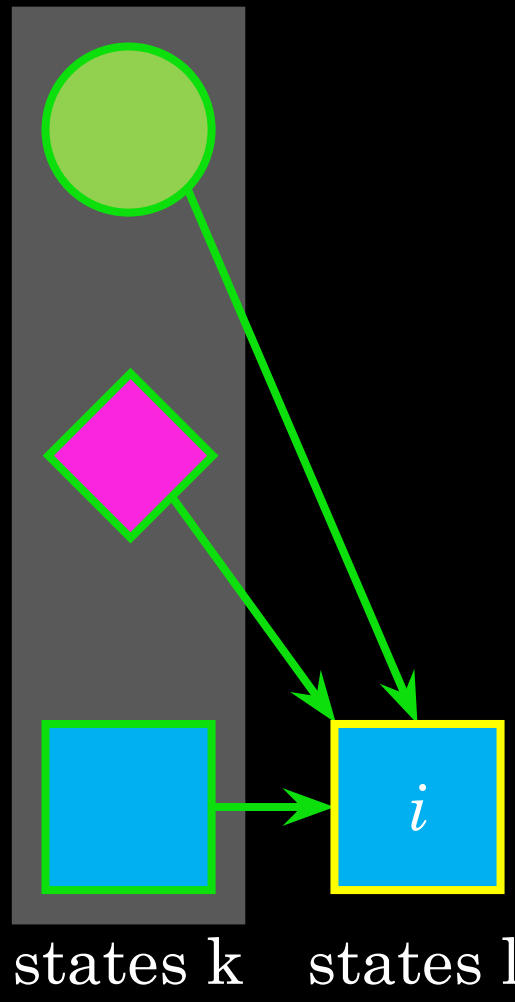
- As with multiple sequence alignment, we cannot be greedy in our choice of path
- But we only need to consider the best path to every possible state in the model
- Dynamic programming!



v(Start) = 1

$$v_l(i) = e_i(x_i)\max_k(v_k(i-1)a_{kl})$$

Huh?



X = ABCDEF

$i = \{A,B,C,D,E,F\}$

$$v_l(i) = e_i(x_i) \max_k (v_k(i-1)a_{kl})$$

Viterbi score of sequence position I at state l

Emission probability of x_i

max over all 3 possibilities

Viterbi score at previous state, times the transition probability

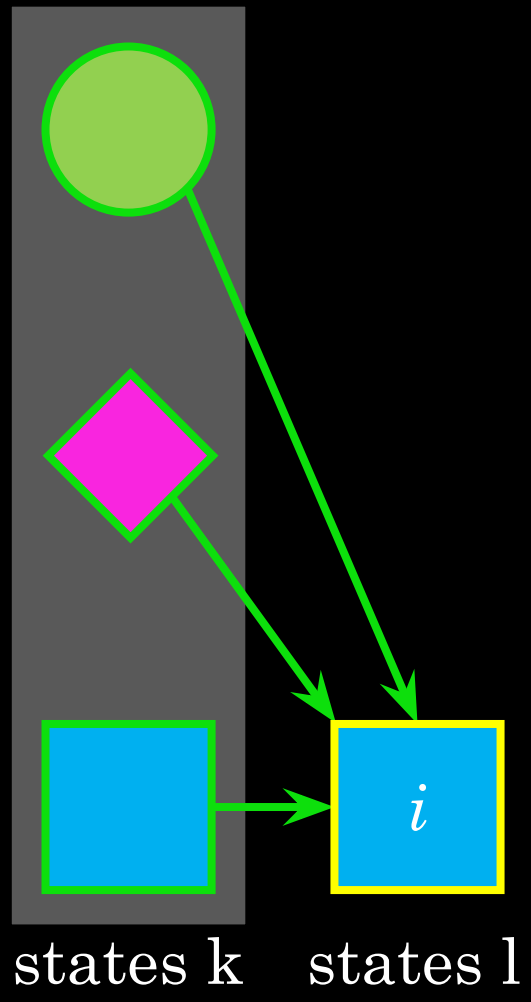
1. Save the best path for each character { A,B,C,D,E,F } at each state in the HMM

2. When we choose our best incoming path, we save a pointer as before and backtrace

3. The Viterbi alignment of each member of a set of sequences *X* to a trained HMM yields a multiple alignment of these sequences

Complexity = O(LS) (# of characters x # of states in the HMM structure) – kinda like n^2

All paths

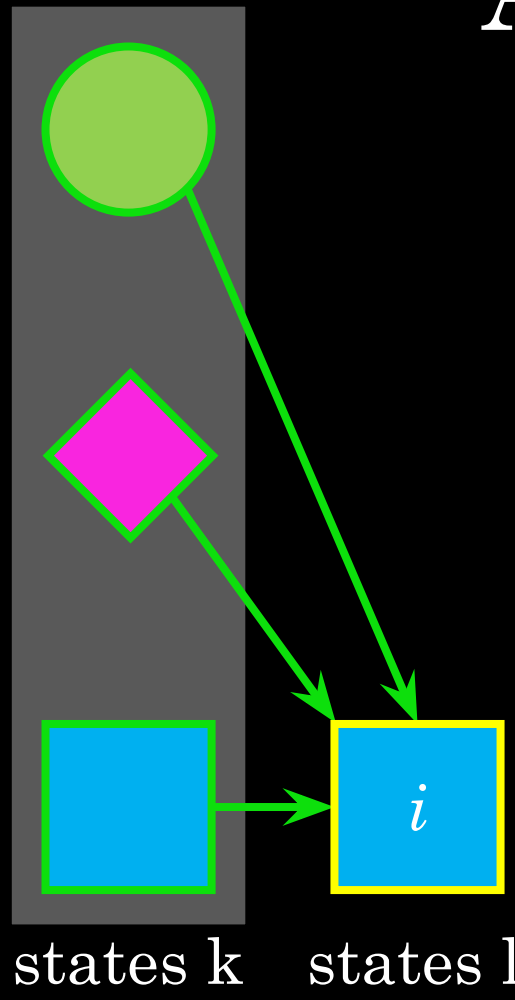


Forward algorithm sums over incoming paths instead of taking max

$$f_l(i) = e_i(x_i) \sum_k (f_k(i-1) a_{kl})$$

Sum (not max)

All paths, reverse order

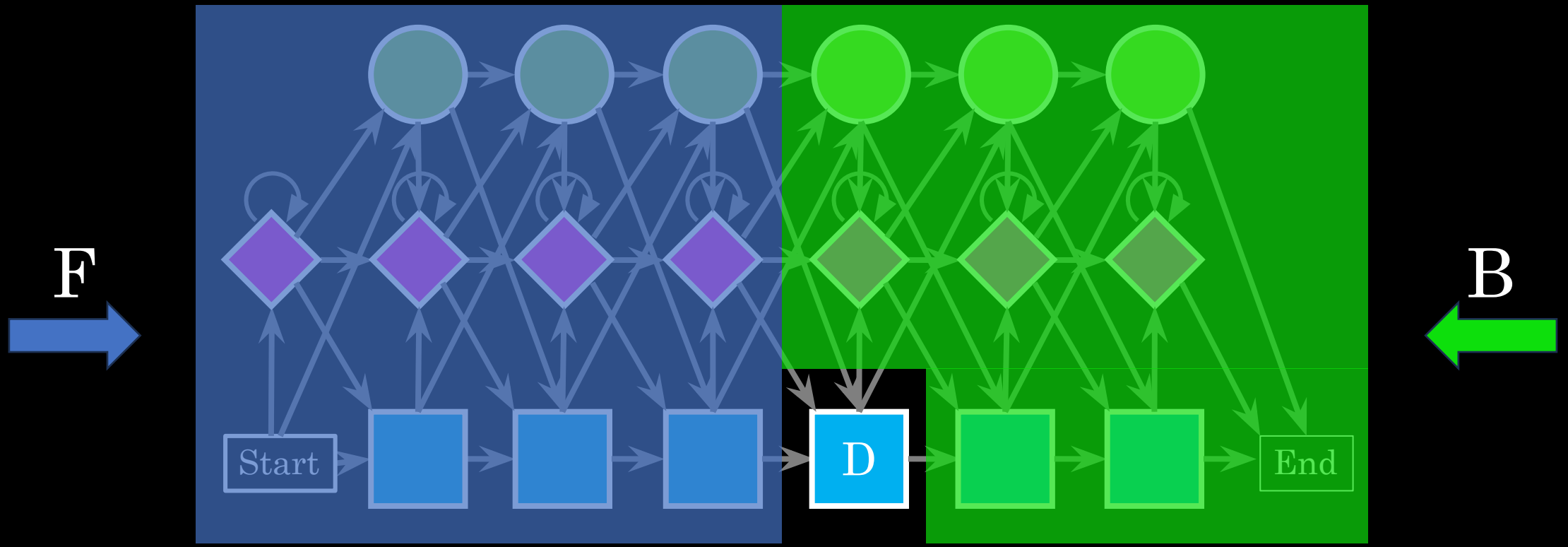


Backward algorithm sums over incoming paths instead of taking max

$$b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$$

Sum (not max)

Combining forward and backward

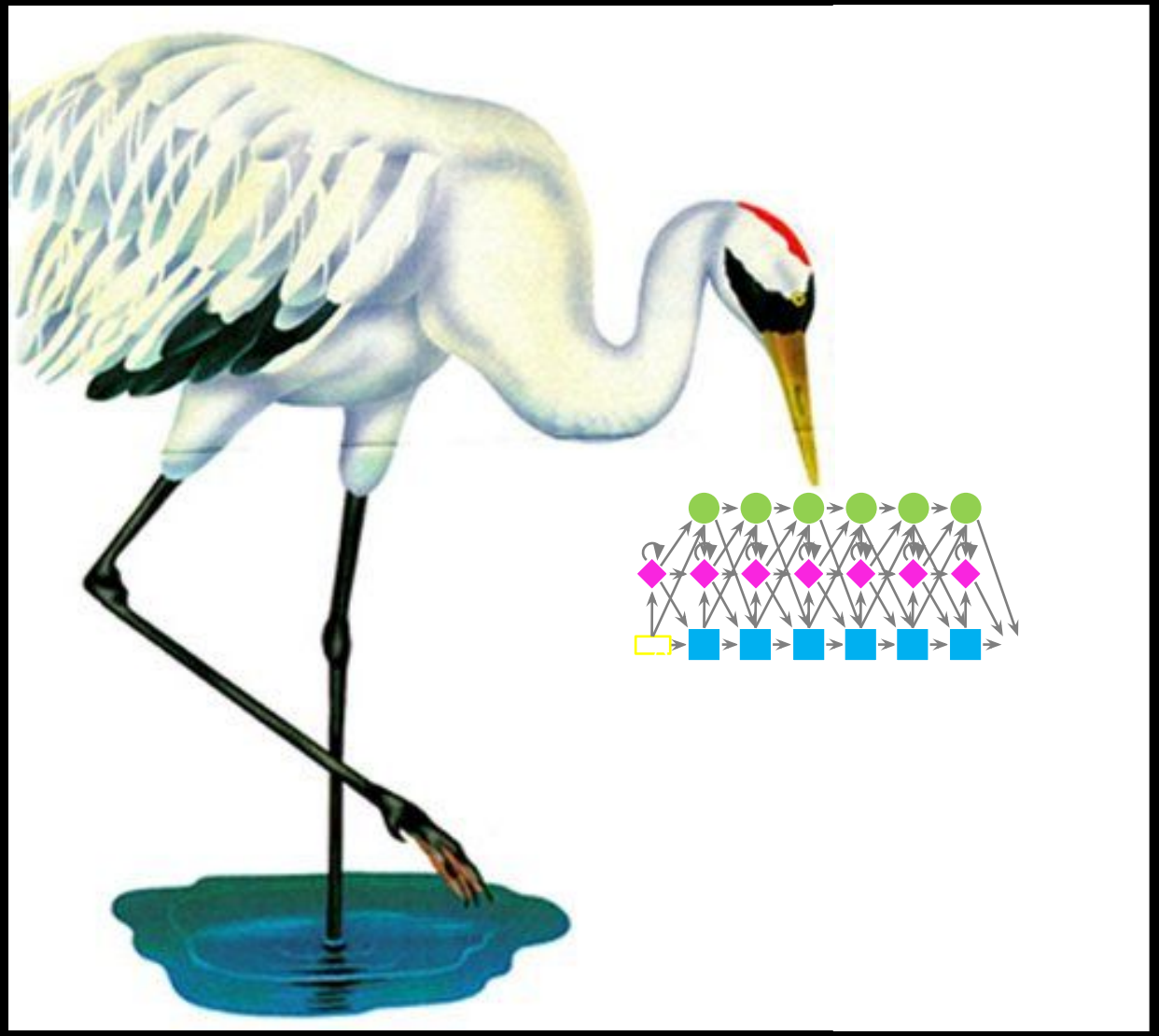


By running the forward and backward algorithms together for a given sequence, we can compute the probability that character i in sequence x maps to state k

$$P(x, k = D)?$$

This is similar to what ProbCons does

Training HMMs



Two components of training

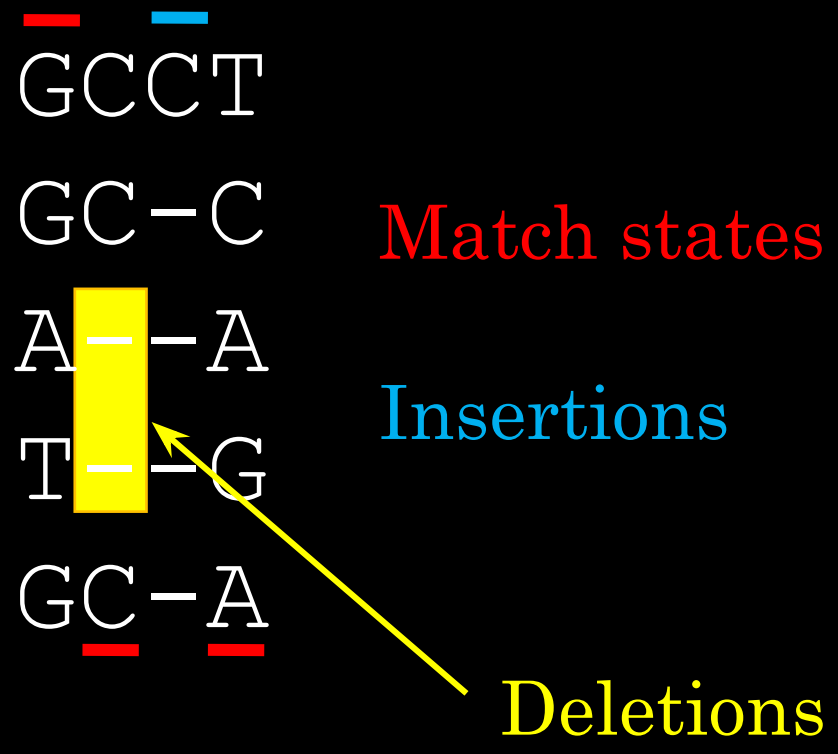
1. Build the HMM structure or ‘skeleton’
   - Custom-tailored with exquisite knowledge of the problem to be modelled
   - In ignorance, whatever, build a complete model

2. Assign transition and emission probabilities

Training (supervised)

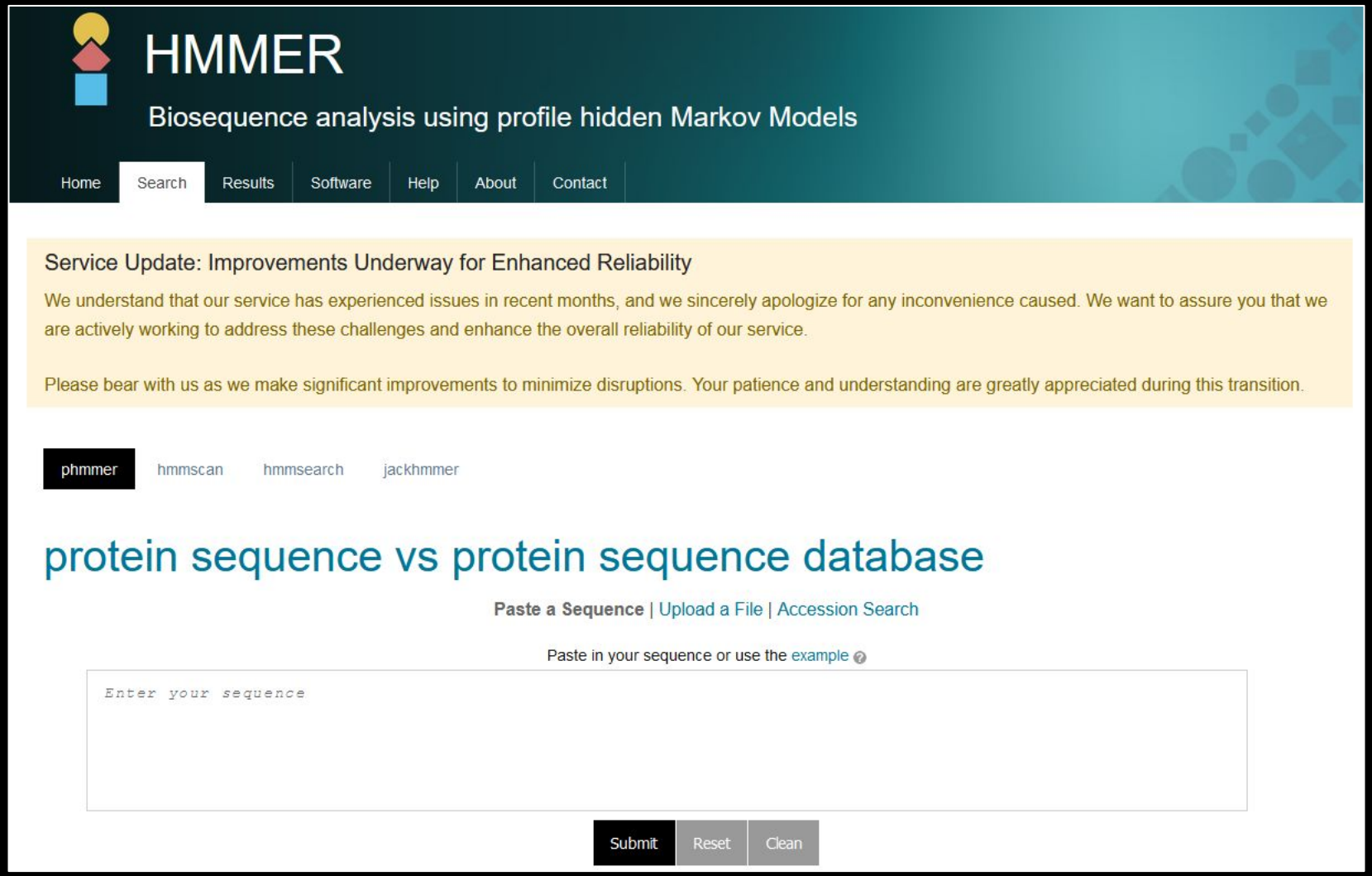
- Construct a multiple sequence alignment using some method, and build the HMM using empirical frequencies

- Supervised because we're specifying exactly WHAT sequences belong in the model



Note that we now get custom gap costs based on observed gap frequencies!

HMMER



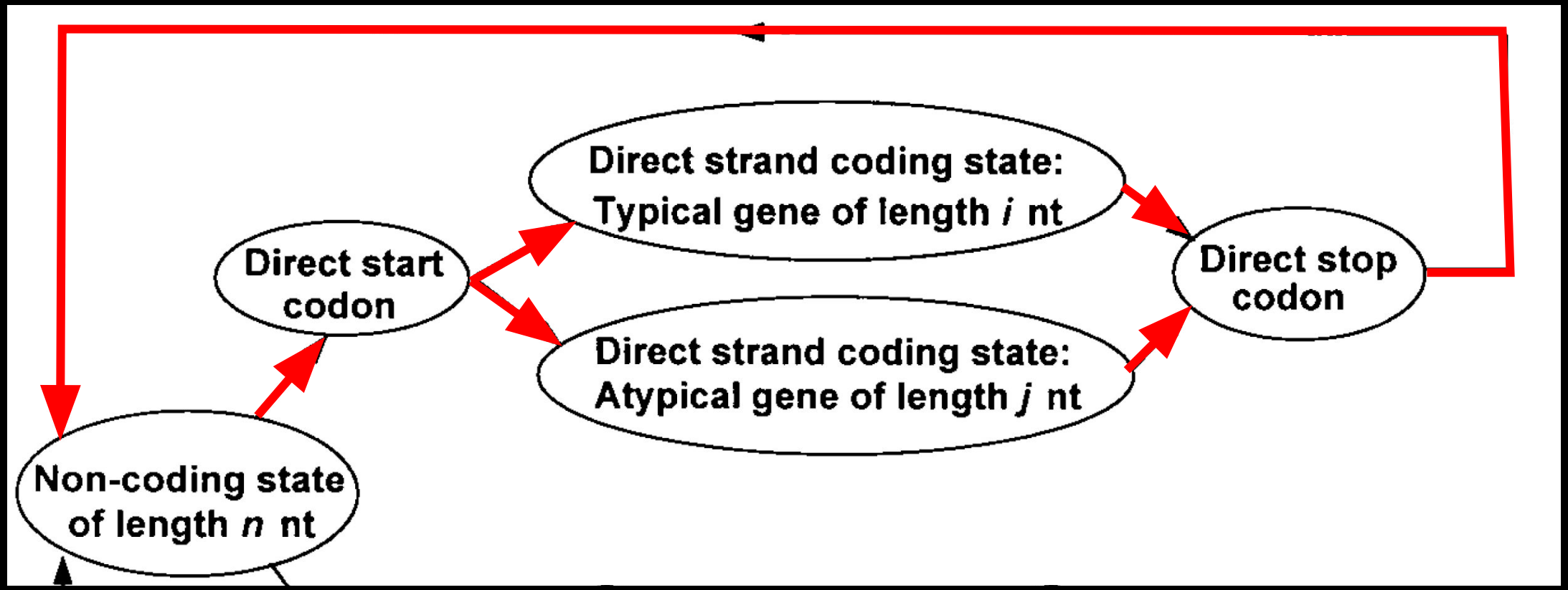
Eddy et al. (2011) *PLoS Comput Biol* (See hmmer.org)
https://www.ebi.ac.uk/Tools/hmmer/search/phmmer

HMMER Tools

hmmbuild	build profile from input multiple alignment
hmmalign	make multiple sequence alignment using a profile
hmmsearch	search profile against sequence database
hmmscan	search sequence against profile database
hmmpress	prepare profile database for `hmmscan`
phmmer	search single sequence against sequence database
jackhmmer	iteratively search single sequence against database
nhmmer	search DNA query against DNA sequence database
nhmmscan	search DNA sequence against a DNA profile database
hmmfetch	retrieve profile(s) from a profile file
hmmstat	show summary statistics for a profile file
hmmemit	generate (sample) sequences from a profile
hmmlogo	produce a conservation logo graphic from a profile
hmmconvert	convert between different profile file formats
hmmpgmd	search daemon for the `hmmer.org` website
hmmpgmd_shard	sharded search daemon for the `hmmer.org` website
makehmmerdb	prepare an `nhmmer` binary database

Other HMM Applications

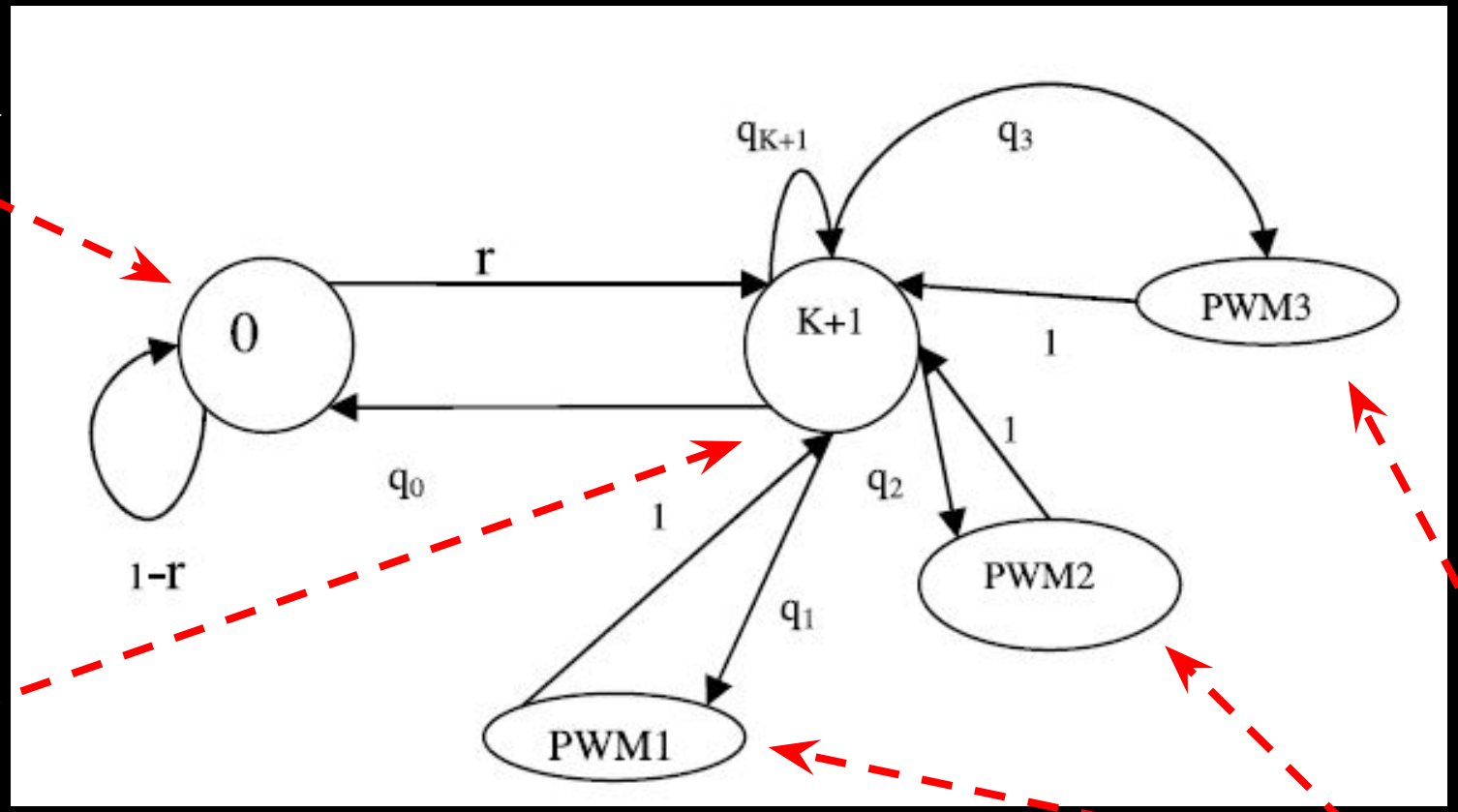
GeneMark.hmm



Lukashin and Borodovsky (1998) *Nucleic Acids Res*

Regulatory Elements
(Promoters and Adjacent patterns)

Not a regulatory region

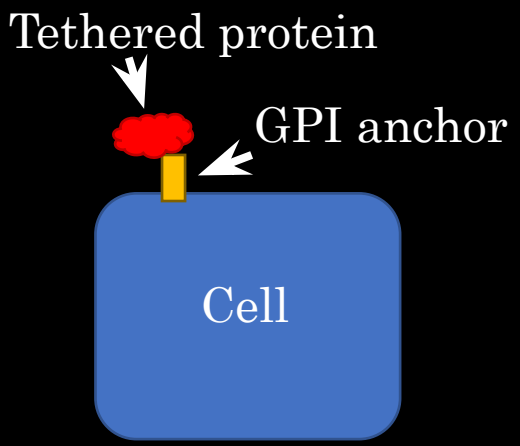


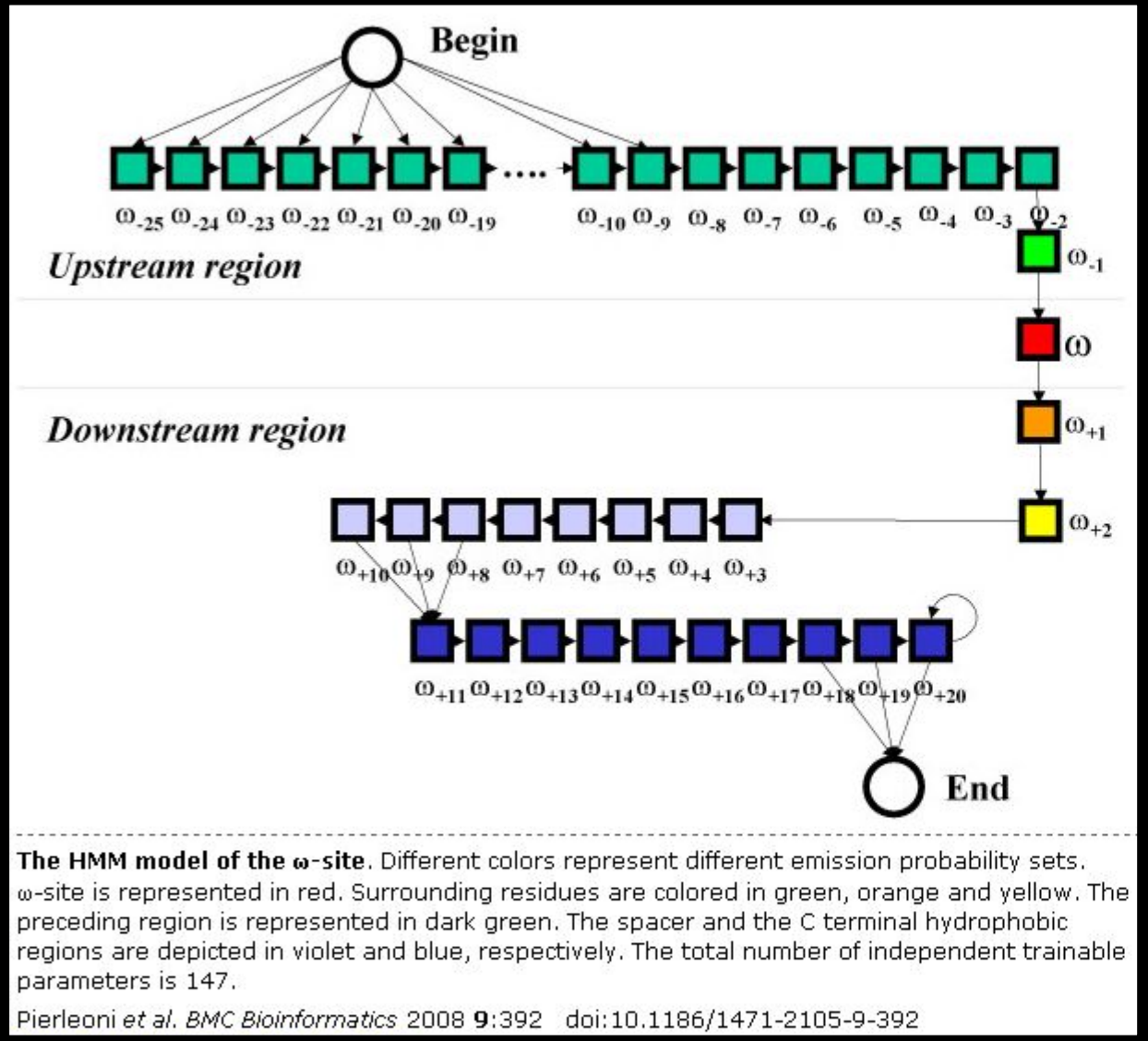
Regulatory region

3 different types of RE

Wu and Xie, *J Comput Biol* (2007)

Glycosylphasphatidylinositol anchors





The HMM model of the ω-site. Different colors represent different emission probability sets. ω-site is represented in red. Surrounding residues are colored in green, orange and yellow. The preceding region is represented in dark green. The spacer and the C terminal hydrophobic regions are depicted in violet and blue, respectively. The total number of independent trainable parameters is 147.

Pierleoni *et al. BMC Bioinformatics* 2008 **9**:392 doi:10.1186/1471-2105-9-392

Pierleoni et al. (2008) *BMC Bioinformatics*

Advantages of HMMs

- Probabilistic framework – the forward algorithm returns the probability of the data (sequence) given the model (the HMM)

- Eminently tweakable – can be designed carefully to capture the patterns in biological sequences

Disadvantages

- Must be designed carefully to adequately capture the patterns in biological sequences
  - Or, use a generic framework
- Can be computationally expensive (kind of like DP for sequence alignment)
- It's Markovian, so you cannot represent correlations of matches at different sites


12
10
2
8
20