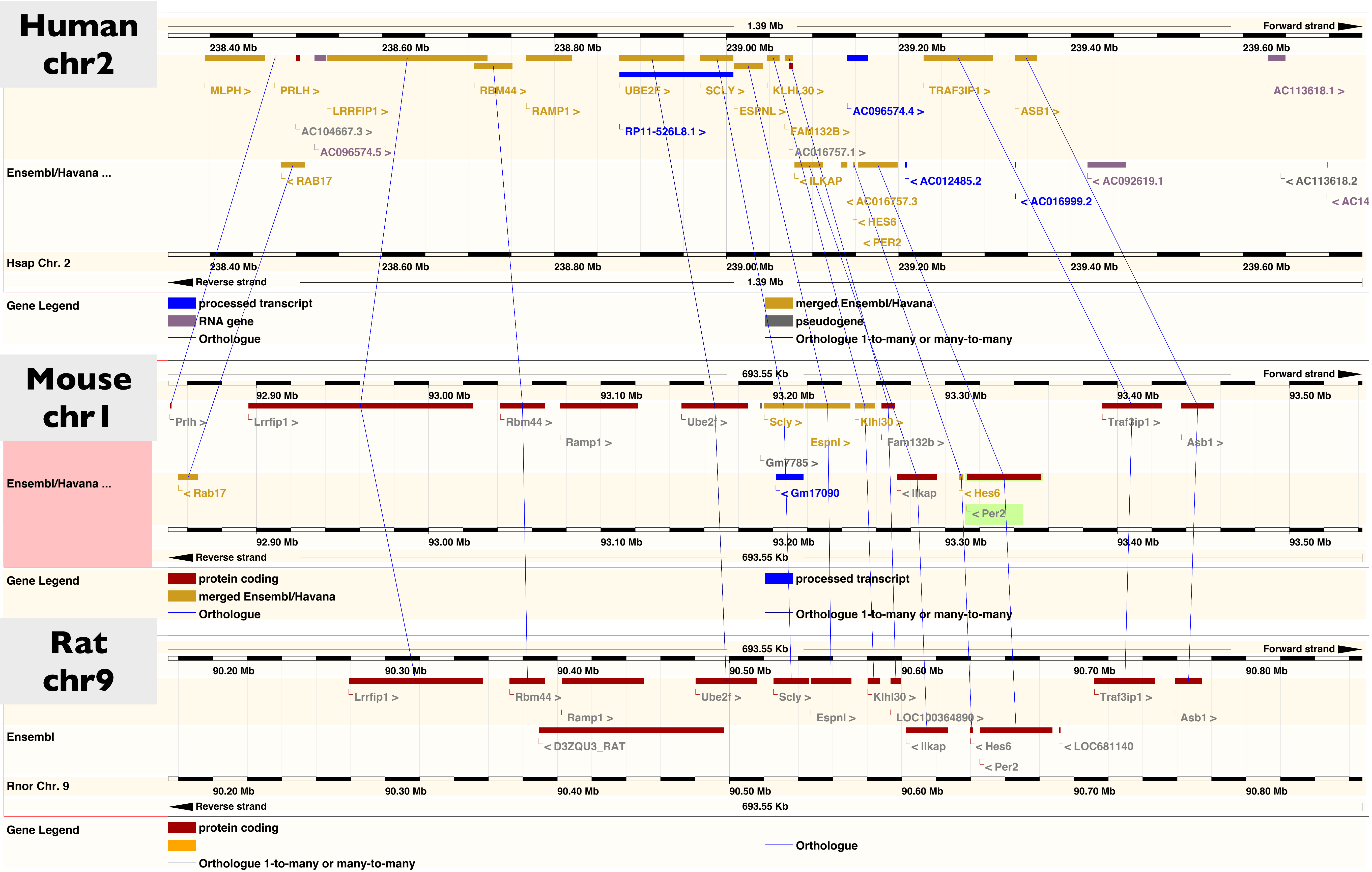
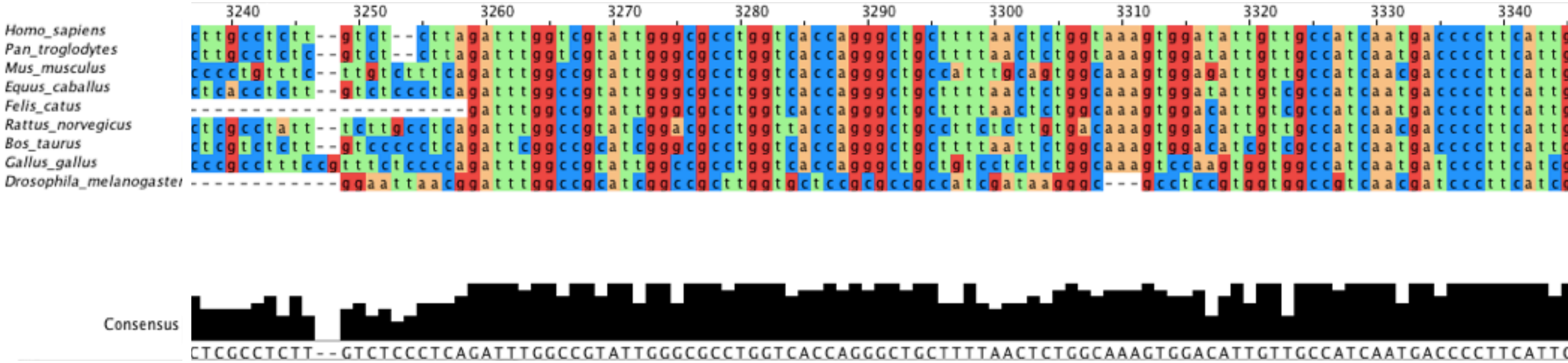
“same” gene in different species



Multiple sequence alignments

- Move from gene homology to homology at the nucleotide and residue levels:
  - Input: a set of (homologous) sequences
  - Output: an alignment of every pair of sequences that is consistent ($x_i \leftrightarrow y_j$ and $x_i \leftrightarrow z_k$ implies $z_k \leftrightarrow y_j$)

Multiple sequence alignments

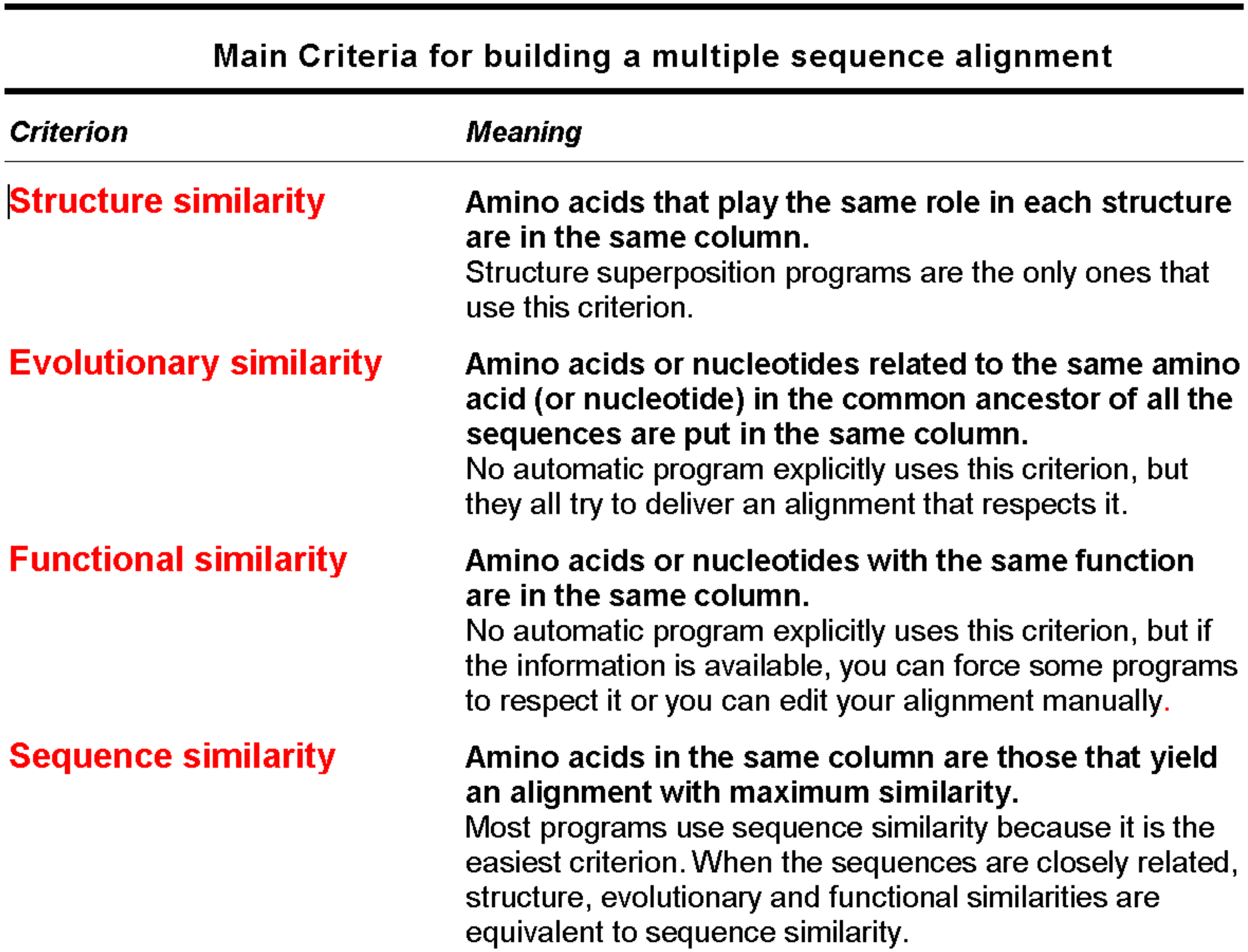


```
Homo_sapiens               cttgcctctt--gtct--cttagatttggtcgtattgggcgcctggtcaccagggctgcttttaactctggtaaagtggatattgttgccatcaatgaccccttcattg
Pan_troglodytes            cttgcctctc--gtct--cttagatttggtcgtattgggcgcctggtcaccagggctgcttttaactctggtaaagtggatattgttgccatcaatgaccccttcattg
Mus_musculus               cccctgtttc--ttgtctttcagatttggccgtattgggcgcctggtcaccagggctgccatttgcagtggcaaagtggagattgttgccatcaacgaccccttcattg
Equus_caballus             ctcacctctt--gtctccctcagatttggccgtattgggcgcctggtcaccagggctgcttttaactctggcaaagtggatattgtcgccatcaatgaccccttcattg
Felis_catus                ----------------------gatttggccgtattgggcgcctggtcaccagggctgcttttaactctggcaaagtggacattgtcgccatcaatgaccccttcattg
Rattus_norvegicus          ctcgcctatt--tcttgcctcagatttggccgtatcggacgcctggttaccagggctgccttctcttgtgacaaagtggacattgttgccatcaacgaccccttcattg
Bos_taurus                 ctcgtctctt--gtccccctcagattcggccgcatcgggcgcctggtcaccagggctgcttttaattctggcaaagtggacatcgtcgccatcaatgaccccttcattg
Gallus_gallus              cccgcctttccgtttctccccagatttggccgtattggccgcctggtcaccagggctgctgtcctctctggcaaagtccaagtggtggccatcaatgatcccttcatcg
Drosophila_melanogaster    ------------ggaattaacggatttggccgcatcggccgcttggtgctccgcgccgccatcgataagggc---gcctccgtggtggccgtcaacgatcccttcatcg

Consensus                  CTCGCCTCTT--GTCTCCCTCAGATTTGGCCGTATTGGGCGCCTGGTCACCAGGGCTGCTTTTAACTCTGGCAAAGTGGACATTGTTGCCATCAATGACCCCTTCATTG
```

Table containing all sequences (species) as rows and homologous positions as columns

Multiple sequence alignments

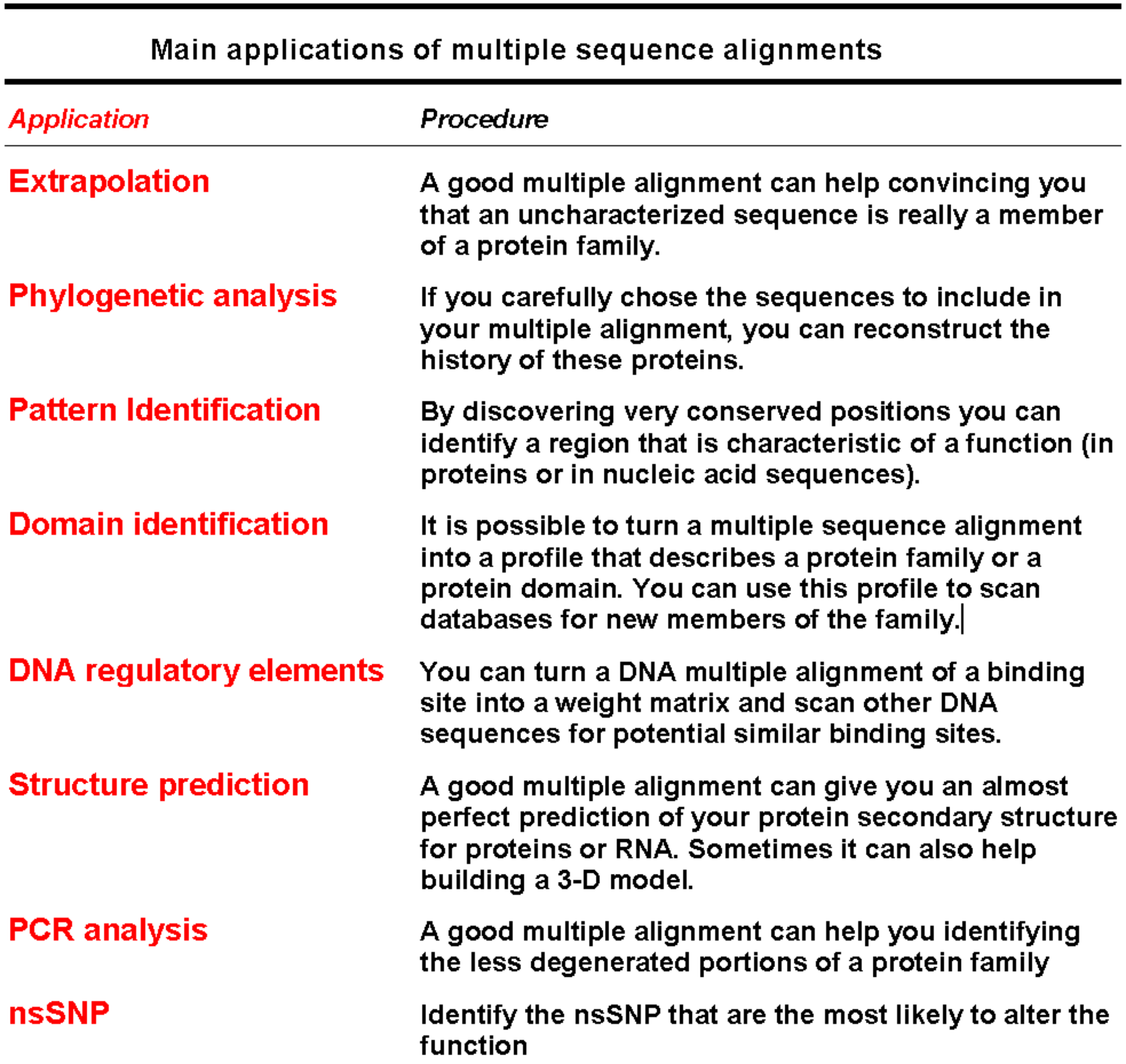


Main Criteria for building a multiple sequence alignment

Criterion	*Meaning*
Structure similarity	**Amino acids that play the same role in each structure are in the same column.** Structure superposition programs are the only ones that use this criterion.
Evolutionary similarity	**Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column.** No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	**Amino acids or nucleotides with the same function are in the same column.** No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manually.
Sequence similarity	**Amino acids in the same column are those that yield an alignment with maximum similarity.** Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structure, evolutionary and functional similarities are equivalent to sequence similarity.

From Cédric Notredame: http://www.tcoffee.org

Multiple sequence alignments



Main applications of multiple sequence alignments

Application	*Procedure*
Extrapolation	A good multiple alignment can help convincing you that an uncharacterized sequence is really a member of a protein family.
Phylogenetic analysis	If you carefully chose the sequences to include in your multiple alignment, you can reconstruct the history of these proteins.
Pattern Identification	By discovering very conserved positions you can identify a region that is characteristic of a function (in proteins or in nucleic acid sequences).
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain. You can use this profile to scan databases for new members of the family.
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potential similar binding sites.
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for proteins or RNA. Sometimes it can also help building a 3-D model.
PCR analysis	A good multiple alignment can help you identifying the less degenerated portions of a protein family
nsSNP	Identify the nsSNP that are the most likely to alter the function

From Cédric Notredame: http://www.tcoffee.org

Where's the problem?

```
TACAT
TCCAGT
TCAGT
```

pairwise alignments
(Needleman-Wunsch)

```
TACA-T   TACA-T   TCCAGT
TCCAGT   T-CAGT   T-CAGT
```

pile them up:
they are consistent!

```
TACA-T
TCCAGT
T-CAGT
```

Where's the problem?

```
TACAT
TCCAGT
TCAGT
```

pairwise alignments
(Needleman-Wunsch)

```
TACA-T    TACA-T    TCCAGT
TCCAGT    T-CAGT    T-CAGT
```

pile them up:
they are consistent!

```
TACA-T
TCCAGT
T-CAGT
```

```
TACAT
TCCAGT
TAGT
```

```
TACA-T    TACAT     TCCAGT
TCCAGT    TAG-T     T--AGT
```

Where's the problem?

```
TACAT
TCCAGT
TCAGT
```

pairwise alignments
(Needleman-Wunsch)

```
TACA-T   TACA-T   TCCAGT
TCCAGT   T-CAGT   T-CAGT
```

pile them up:
they are consistent!

```
TACA-T
TCCAGT
T-CAGT
```

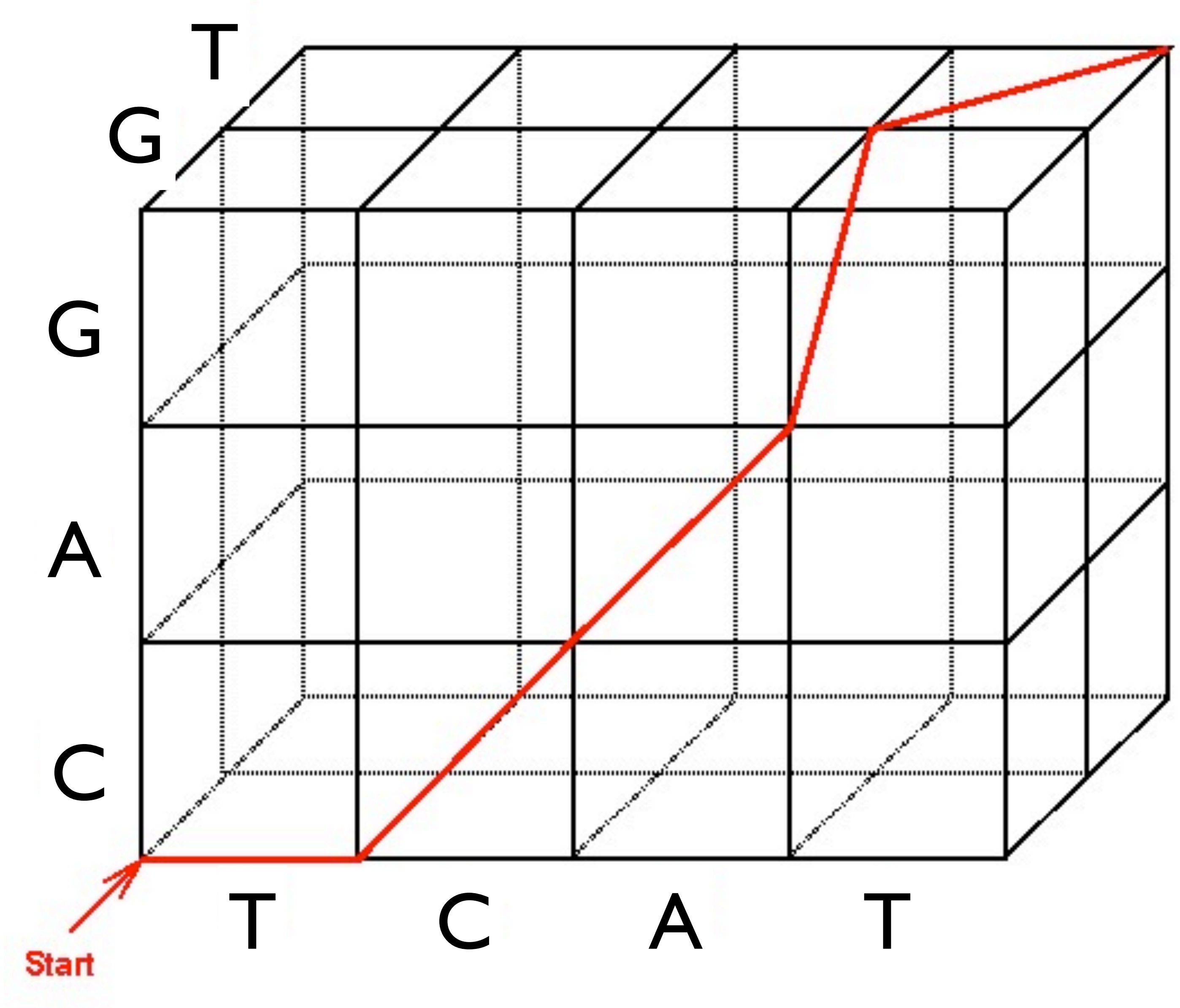
```
TACAT
TCCAGT
TAGT
```

```
TACA-T   TACAT    TCCAGT
TCCAGT   TAG-T    T--AGT
```

?　?

```
TACA-T         TACA-T
TCCAGT   or    TCCAGT   ...
TAG--T         T--AGT
```

Multi-dimensional dynamic programming



```
TCA-T
-CAG-
---GT
```

Dynamic programming:

extend Needleman-Wunsch to $n_1 \times n_2 \times n_3$ table

Scoring MSA

We want to optimize the MSA, but what is the score?

Version 1: Sum of pairs (SP) score

$$S(\mathrm{MSA}) = \sum_i S(\mathrm{col}_i) ,$$

score each column independently,
then sum them all

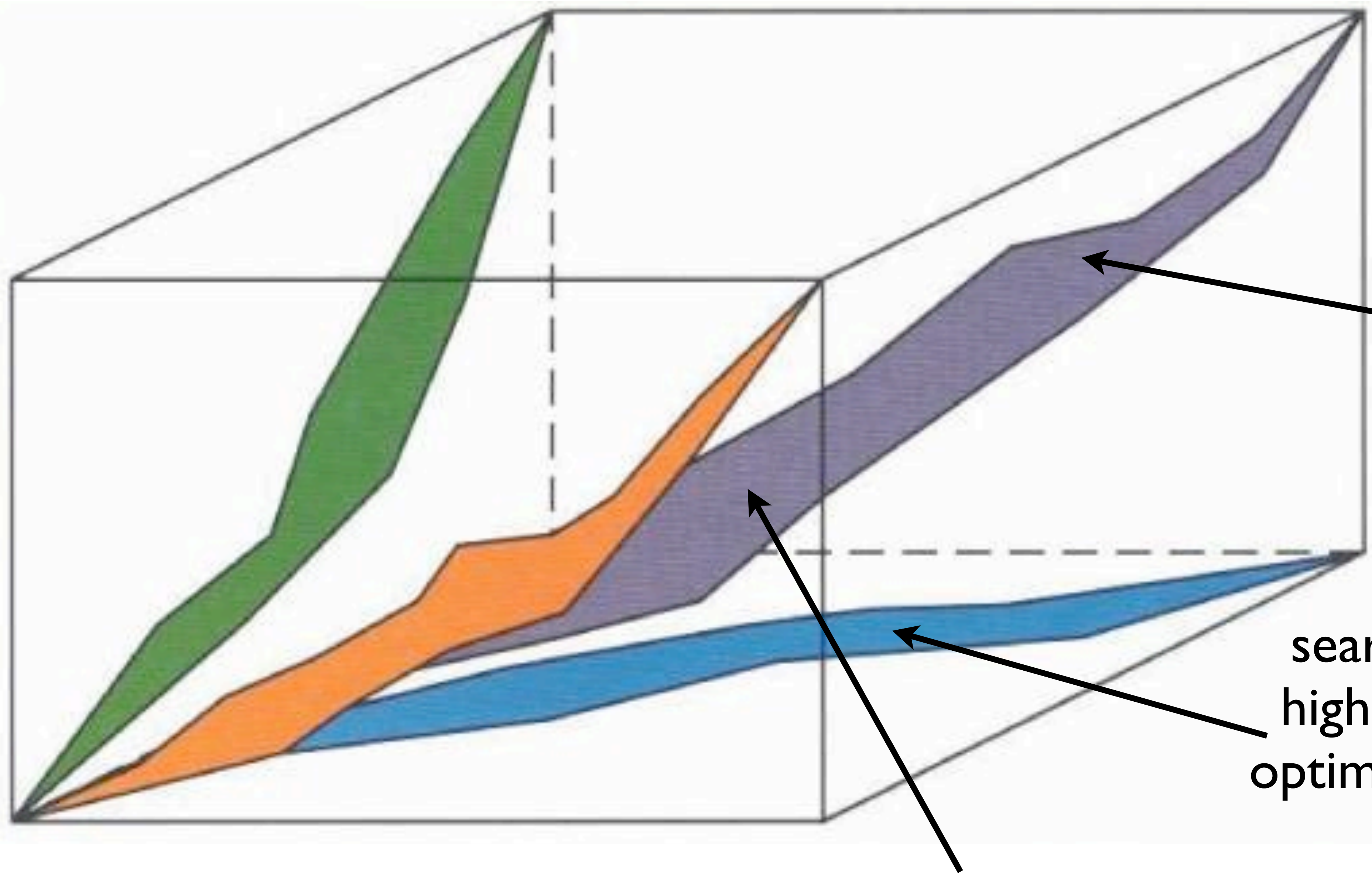
$$S(\mathrm{col}_i) = \sum_{j<k} M(x_{ji}, x_{ki}) .$$

*i*th column in alignment

All pairs of sequences

Substitution matrix

Carrillo-Lipman Bounds



find (heuristically)
one MSA
that is “good enough”

search pairwise scores
higher than initial MSA-optimal pairwise score for
other pairs

optimal alignment must be in the
intersection of the pairwise bounds

Restrict table to a narrow “tube” around initial find

Scoring MSA

Version 1: Sum of pairs (SP) score

Problem with the probabilistic interpretation of scores:

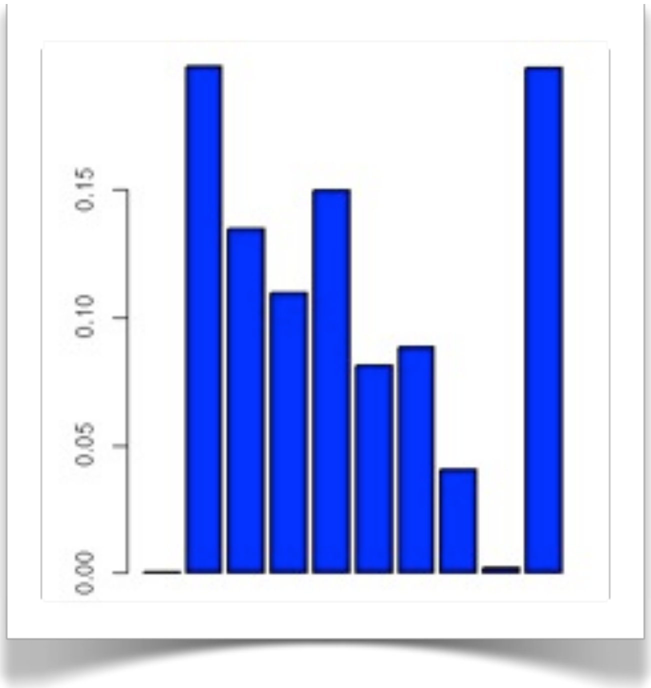
$$\begin{aligned} & M(x_{ji}, x_{ki}) + M(x_{ji}, x_{\ell i}) + M(x_{ki}, x_{\ell i}) \\ = \quad & \log p(x_{ji} x_{ki}) + \log p(x_{ji} x_{\ell i}) + \log p(x_{ki} x_{\ell i}) \\ & -2 \log q(x_{ji}) - 2 \log q(x_{ki}) - 2 \log q(x_{\ell i}) \\ \neq \quad & \log \frac{p(x_{ji} x_{ki} x_{\ell i})}{q(x_{ji}) q(x_{ki}) q(x_{\ell i})} \ , \end{aligned}$$

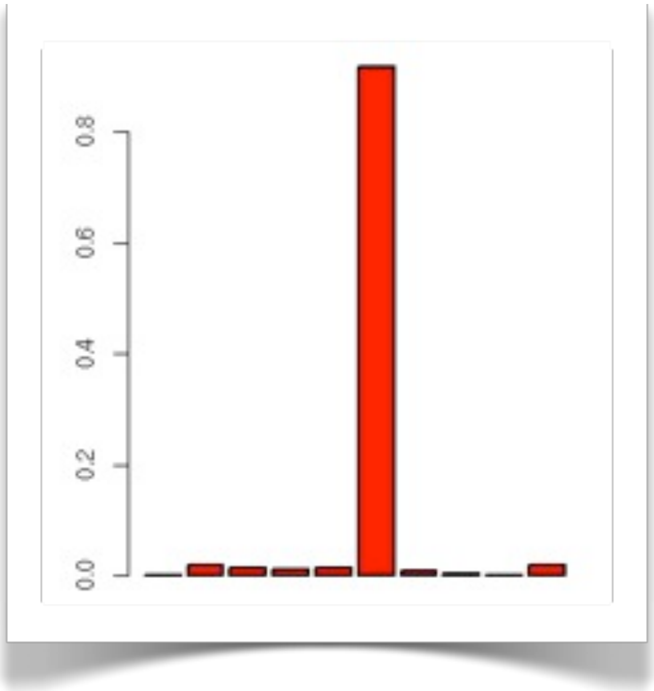
Common improvement uses weighted means of pairs scores

Scoring MSA

Version 2: Entropy

- Entropy H(p) is a measure of how flat (or peaked) is a probability distribution
- peaked = 0≤H≤log2(N) = flat





$$H(p) \;=\; -\sum_{k=1}^{N} p(k)\log_2(p(k))$$

Scoring MSA

Version 2: Entropy

- We can score a MSA by the sum of the column entropies
- Best alignment has minimum entropy

$$\begin{aligned} p_i(n) &= \frac{\#\{x_{ji} = n\}}{\#\{x_{ji} \neq -\}} \ , \\ n &= A, C, G, T \ . \end{aligned}$$

Progressive methods

- Strategy: compute all pairwise optimal alignment scores
- Build a guide tree by “Neighbour-Joining” (NJ):
  - join the two nearest items in a tree node
  - replace the pair by one item in the list, its pairwise scores are the maximum of the two scores
  - iterate these two operations
- Construct the MSA in the order of the tree
  - when aligning a sequence to a previous MSA, do what is called a profile alignment

Progressive methods

Pairwise alignment scores

	B	C	D	E
A	9	1	2	2
	B	4	2	0
		C	3	0
			D	7
				E

ACATGT

ACAAGT

TCAAGGT

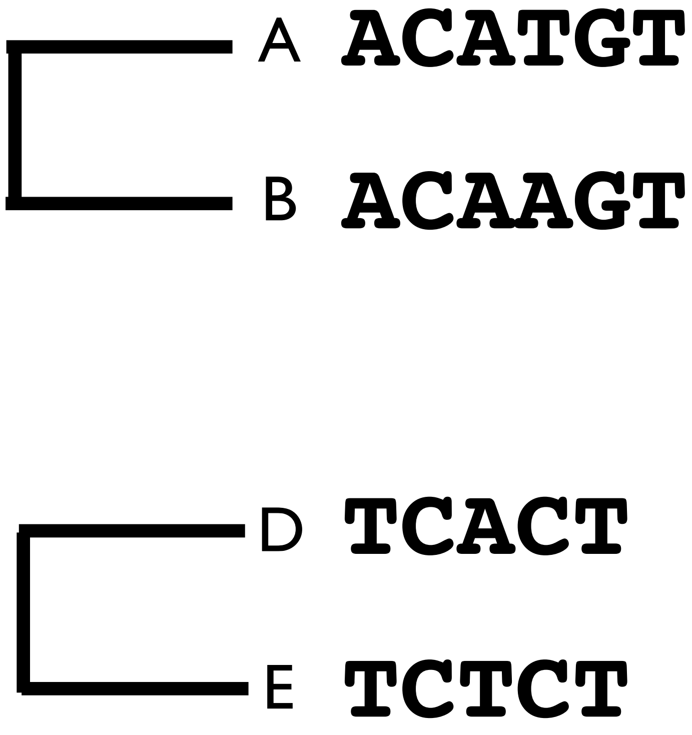
TCACT

TCTCT

Progressive methods

Pairwise alignment scores

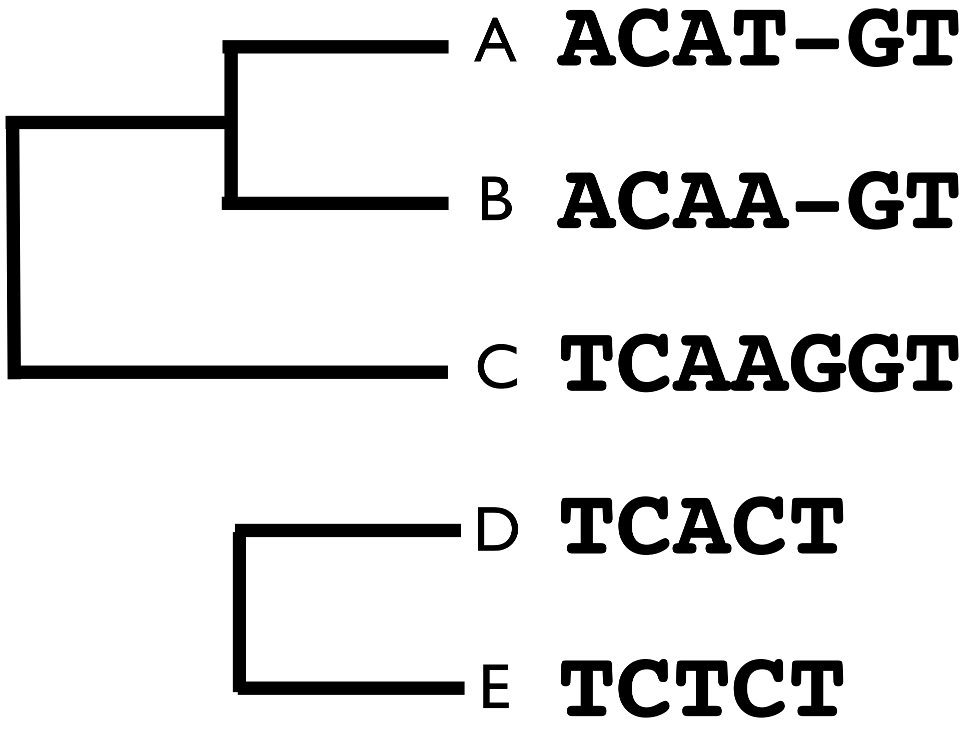
	B	C	D	E	
A	9	1	2	2	**ACATGT**
	B	4	2	0	**ACAAGT**
		C	3	0	**TCAAGGT**
			D	7	**TCACT**
				E	**TCTCT**



Progressive methods

Pairwise alignment scores

	B	C	D	E	
A	9	1	2	2	**ACATGT**
B		4	2	0	**ACAAGT**
C			3	0	**TCAAGGT**
D				7	**TCACT**
E					**TCTCT**



Progressive methods

Pairwise alignment scores

	B	C	D	E
A	9	1	2	2
	B	4	2	0
		C	3	0
			D	7
				E

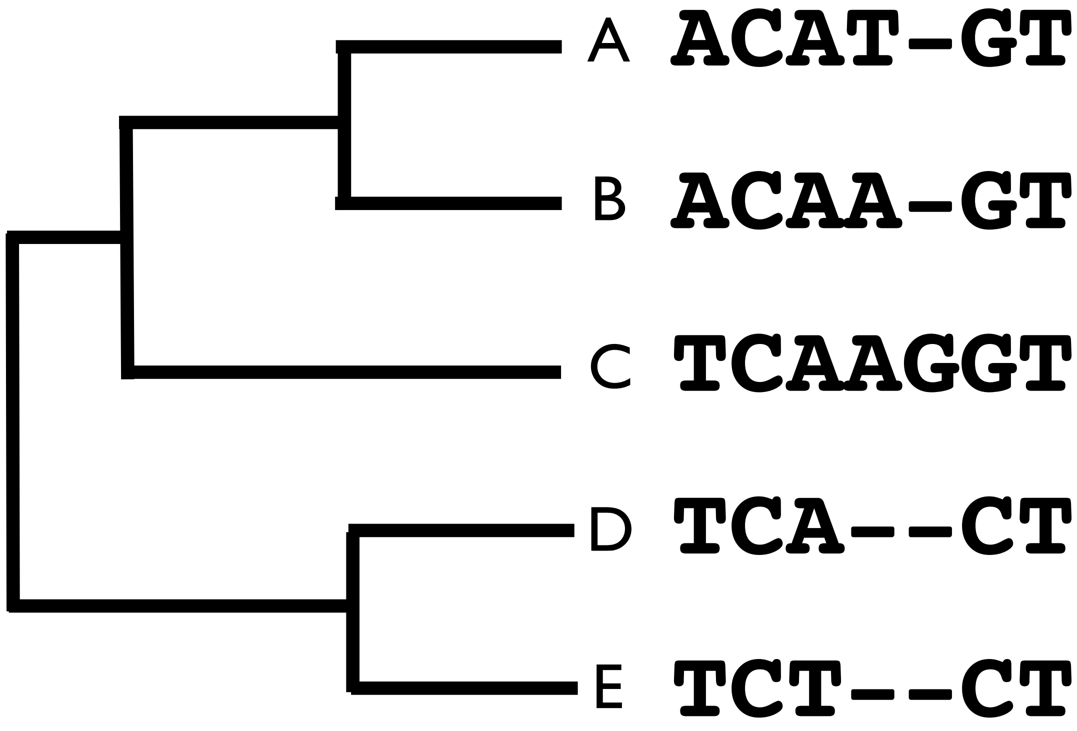
ACATGT

ACAAGT

TCAAGGT

TCACT

TCTCT



A **ACAT-GT**

B **ACAA-GT**

C **TCAAGGT**

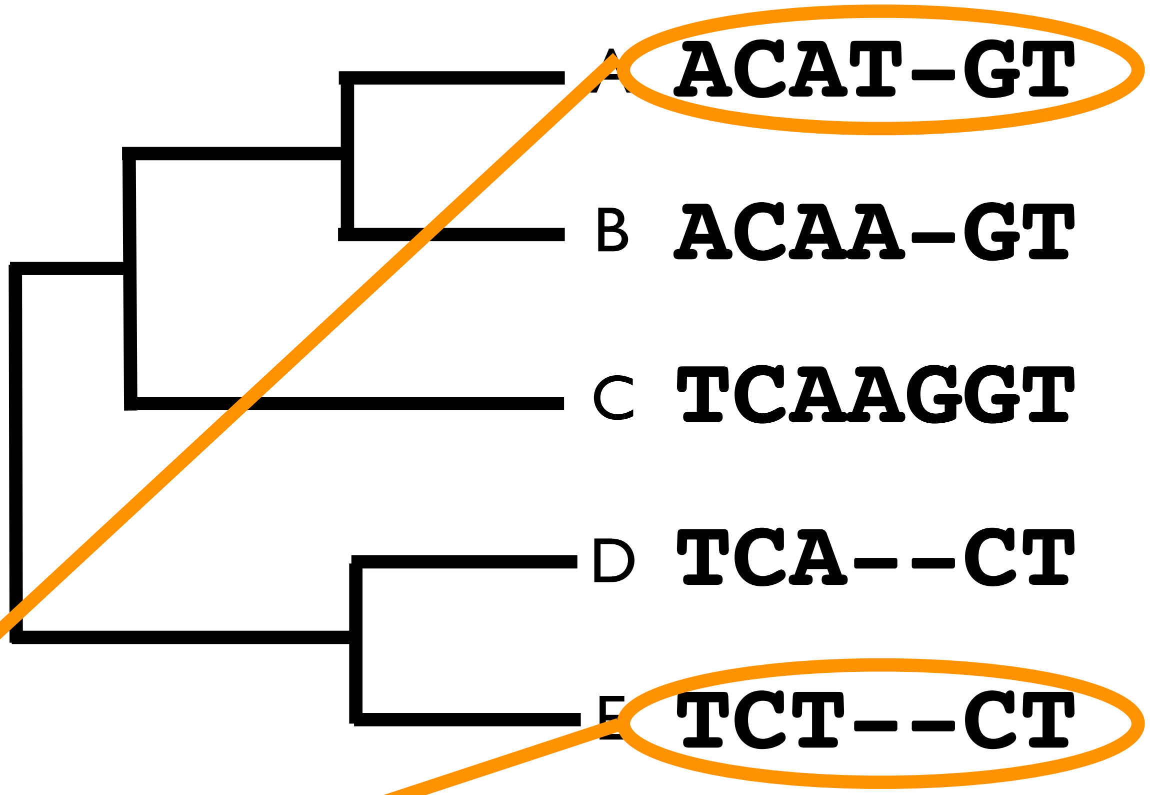
D **TCA--CT**

E **TCT--CT**

Progressive methods

Pairwise alignment scores

	B	C	D	E	
A	9	1	2	2	**ACATGT**
	B	4	2	0	**ACAAGT**
		C	3	0	**TCAAGGT**
			D	7	**TCACT**
				E	**TCTCT**

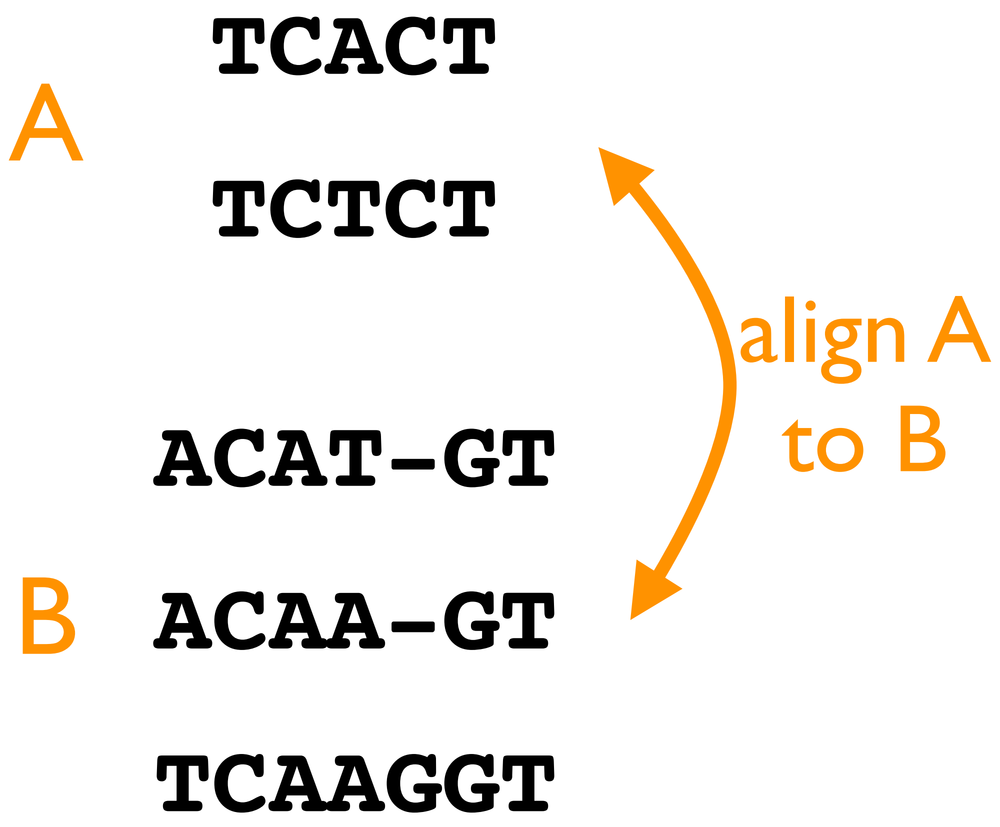


Best pairwise alignment

```
ACAT-GT
TC-T-CT
```

Alignment to a profile

A MSA can be seen as a “profile”:
a nucleotide distribution at each position



A

```
 TCACT
 TCTCT

ACAT-GT
ACAA-GT
TCAAGGT
```

B

align A to B

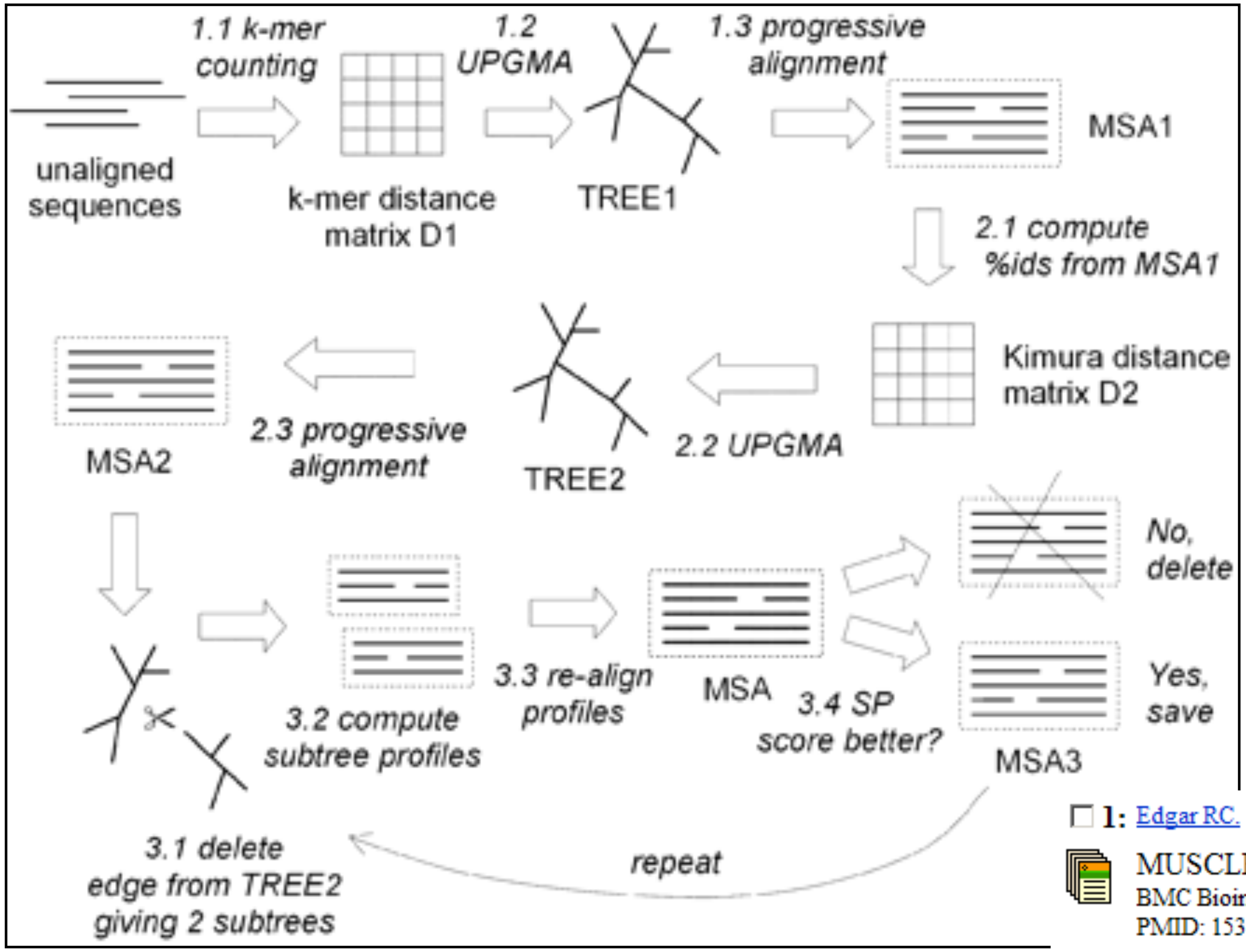
$$
\begin{aligned}
& S(\mathrm{col}_i) \\
= \; & \sum_{j<k} M(x_{ji}, x_{ki}) \\
= \; & \sum_{j,k \in A} M(x_{ji}, x_{ki}) + \sum_{j,k \in B} M(x_{ji}, x_{ki}) \\
& + \sum_{j \in A, k \in B} M(x_{ji}, x_{ki})
\end{aligned}
$$

Only need to optimize this sum

Requires $M(-,-) = 0$

Consistency post-processing

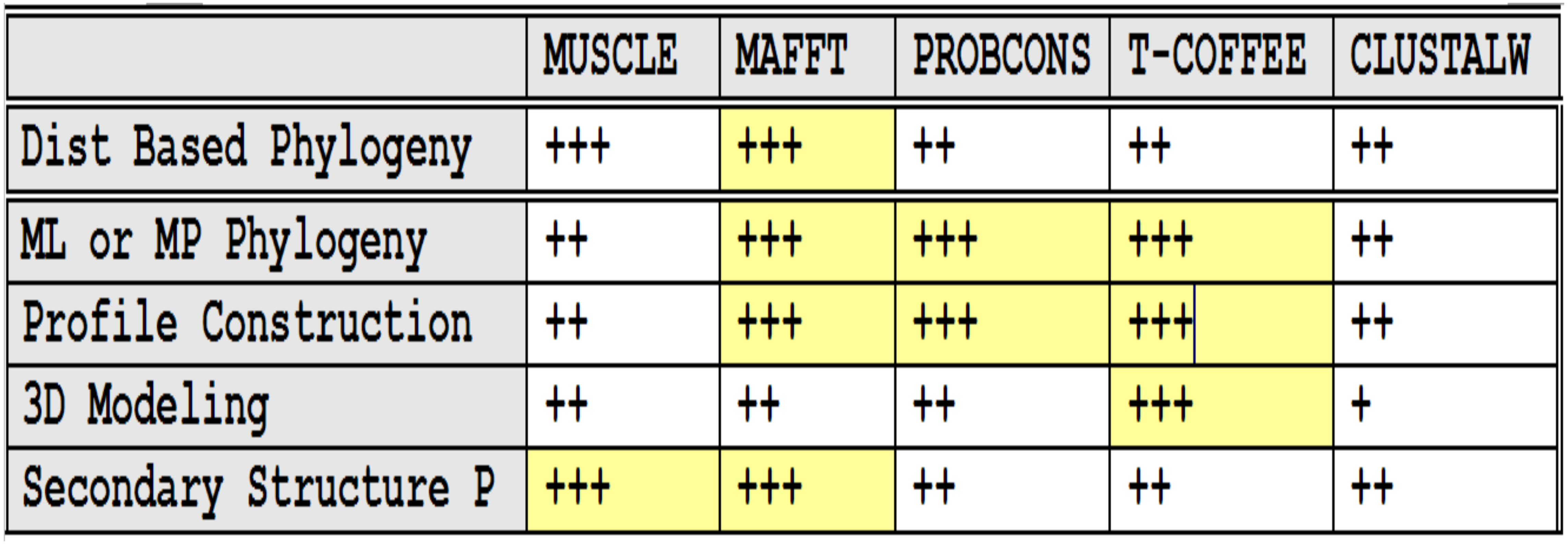
Because of the above scoring scheme, earlier alignments can never be modified by later sequence additions


unaligned sequences
1.1 k-mer counting
k-mer distance matrix D1
1.2 UPGMA
TREE1
1.3 progressive alignment
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
3.1 delete edge from TREE2 giving 2 subtrees
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
No, delete
Yes, save
MSA3
repeat

□ **1:** Edgar RC.

MUSCLE: a multiple sequence alignment method with reduced time and space complexity.
BMC Bioinformatics. 2004 Aug 19;5(1):113.
PMID: 15318951 [PubMed - indexed for MEDLINE]

Purpose ⇔ Solution



	MUSCLE	MAFFT	PROBCONS	T-COFFEE	CLUSTALW
Dist Based Phylogeny	+++	+++	++	++	++
ML or MP Phylogeny	++	+++	+++	+++	++
Profile Construction	++	+++	+++	+++	++
3D Modeling	++	++	++	+++	+
Secondary Structure P	+++	+++	++	++	++