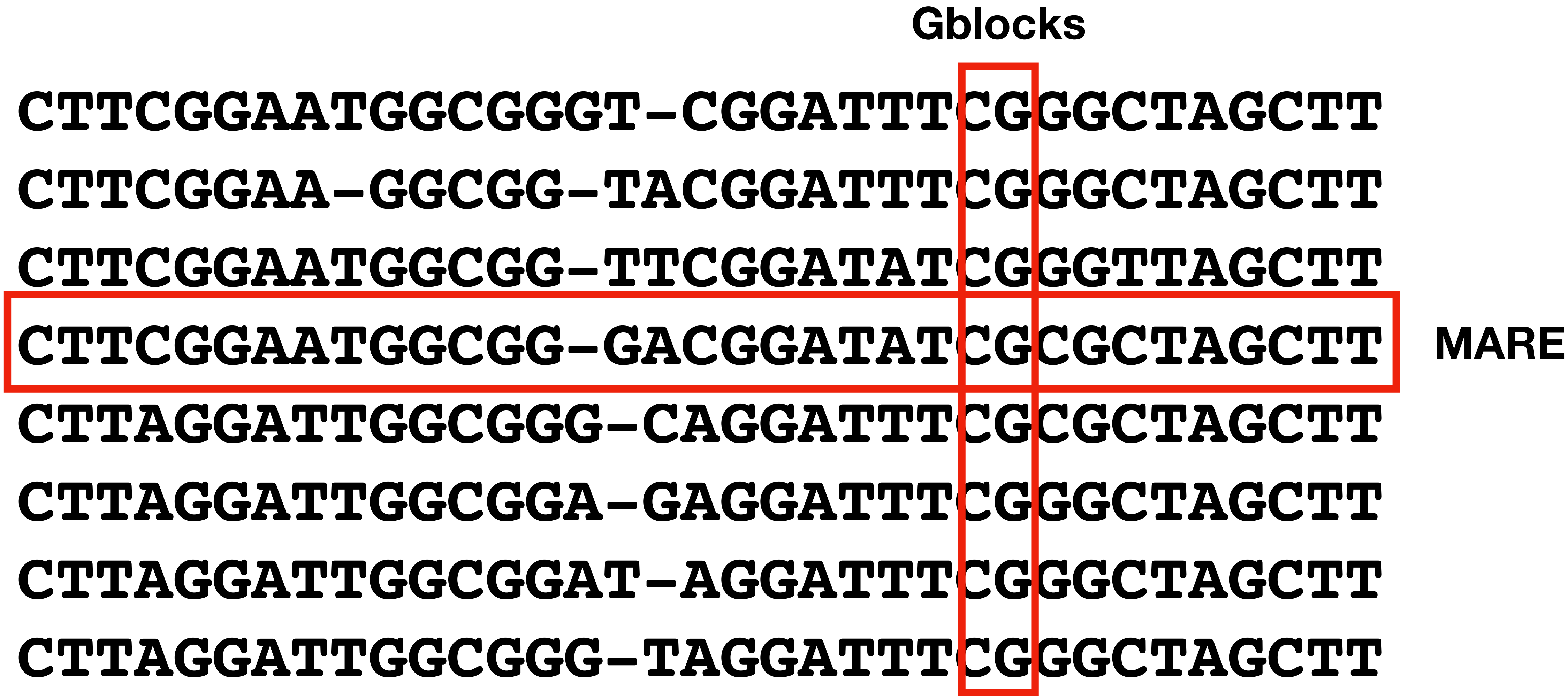
Recent data filtering method

Popular methods to filter/trim poorly aligned sequences

Remove Gap-rich and variable sites (Gblocks/TrimAl)

Matrix reduction (MARE)

What’s the difference

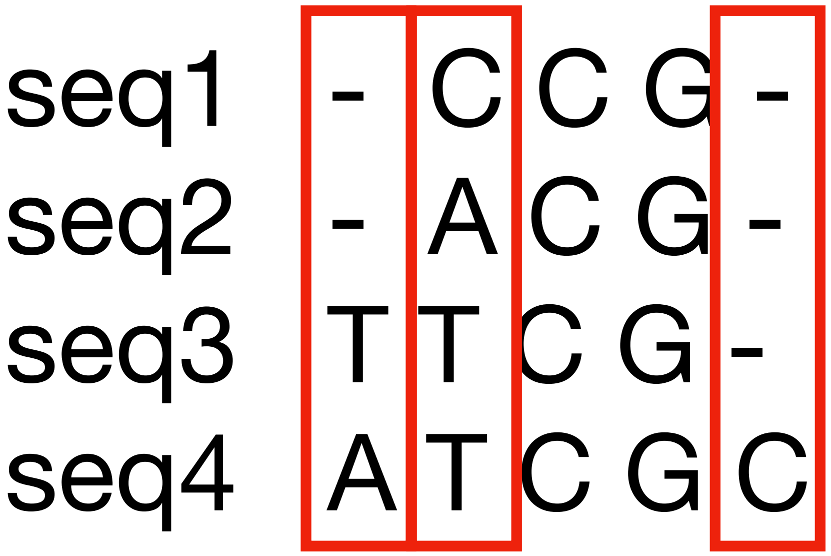


Gblocks

```
CTTCGGAATGGCGGGT-CGGATTTCGGGCTAGCTT
CTTCGGAA-GGCGG-TACGGATTTCGGGCTAGCTT
CTTCGGAATGGCGG-TTCGGATATCGGGTTAGCTT
CTTCGGAATGGCGG-GACGGATATCGCGCTAGCTT
CTTAGGATTGGCGGG-CAGGATTTCGCGCTAGCTT
CTTAGGATTGGCGGA-GAGGATTTCGGGCTAGCTT
CTTAGGATTGGCGGAT-AGGATTTCGGGCTAGCTT
CTTAGGATTGGCGGG-TAGGATTTCGGGCTAGCTT
```

MARE

Gblocks



seq1	-	C	C	G	-
seq2	-	A	C	G	-
seq3	T	T	C	G	-
seq4	A	T	C	G	C

Remove variable sites in columns

Definition

non-conserved positions: < IS identical residues or there is a gap

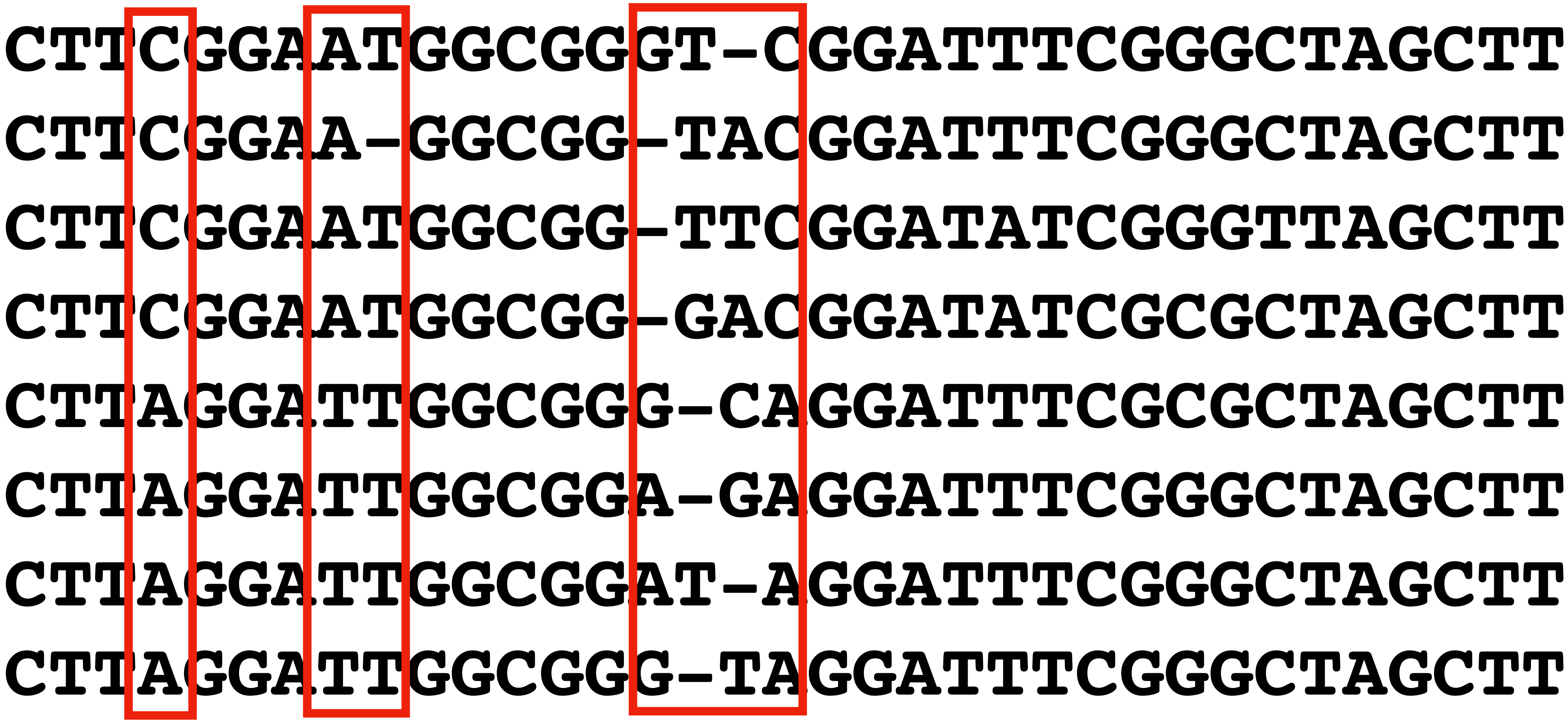
conserved positions: >= IS and < FS identical residues

highly conserved positions: >= FS identical residues

IS = 50% of the number of sequences + 1 10*0.5+1 = 6

FS = 85% of the number of sequences 10*0.85 = 8.5

Find long stretch of non-conserved blocks

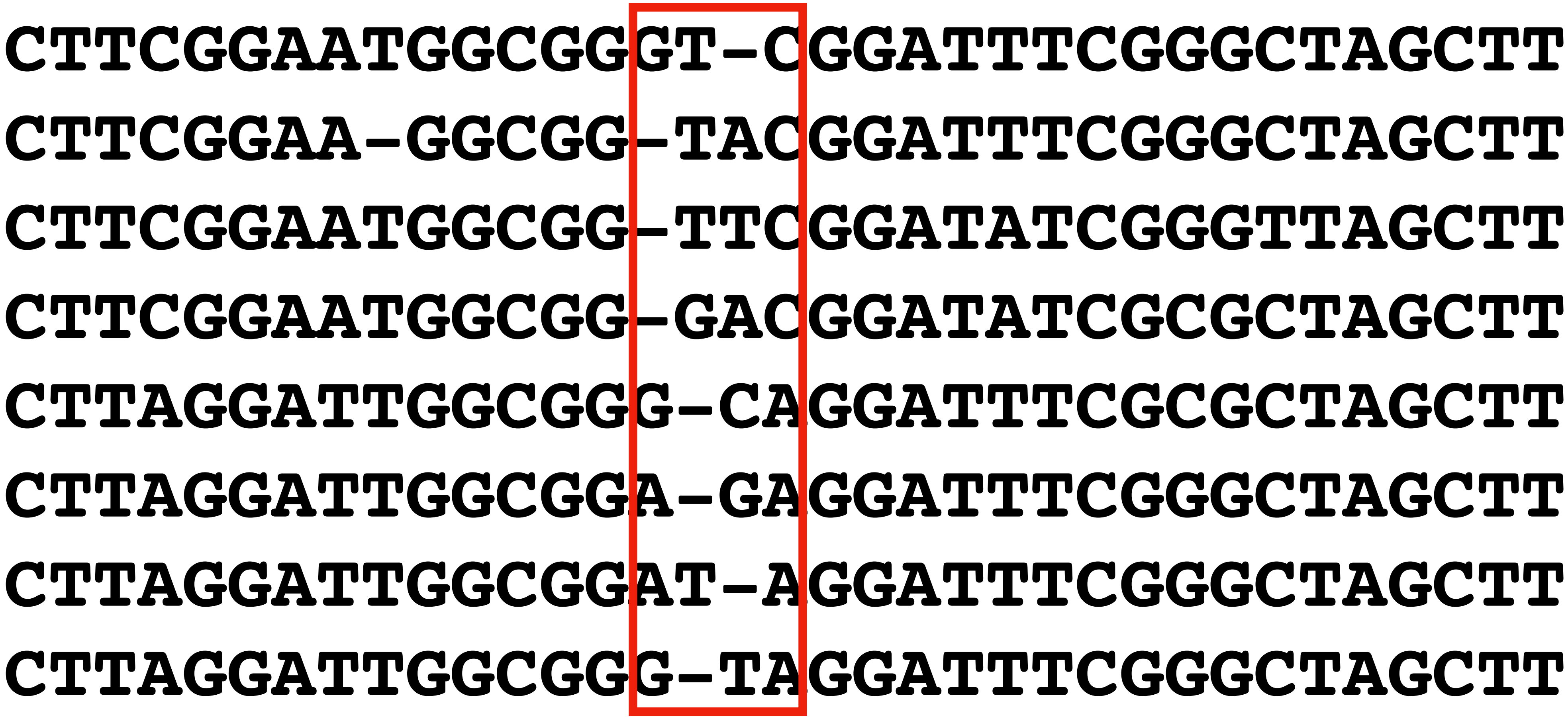


```
CTTCGGAATGGCGGGT-CGGATTTCGGGCTAGCTT
CTTCGGAA-GGCGG-TACGGATTTCGGGCTAGCTT
CTTCGGAATGGCGG-TTCGGATATCGGGTTAGCTT
CTTCGGAATGGCGG-GACGGATATCGCGCTAGCTT
CTTAGGATTGGCGGG-CAGGATTTCGCGCTAGCTT
CTTAGGATTGGCGGA-GAGGATTTCGGGCTAGCTT
CTTAGGATTGGCGGAT-AGGATTTCGGGCTAGCTT
CTTAGGATTGGCGGG-TAGGATTTCGGGCTAGCTT
```

CP = 4

maximum number of contiguous nonconserved positions

Find long stretch of non-conserved blocks



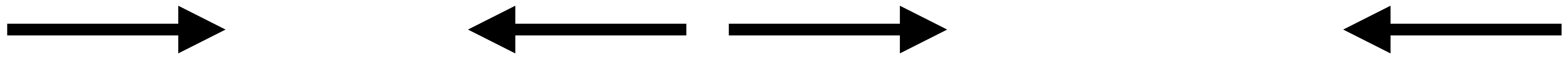
```
CTTCGGAATGGCGGGT-CGGATTTCGGGCTAGCTT
CTTCGGAA-GGCGG-TACGGATTTCGGGCTAGCTT
CTTCGGAATGGCGG-TTCGGATATCGGGTTAGCTT
CTTCGGAATGGCGG-GACGGATATCGCGCTAGCTT
CTTAGGATTGGCGGG-CAGGATTTCGCGCTAGCTT
CTTAGGATTGGCGGA-GAGGATTTCGGGCTAGCTT
CTTAGGATTGGCGGAT-AGGATTTCGGGCTAGCTT
CTTAGGATTGGCGGG-TAGGATTTCGGGCTAGCTT
```

Find long stretch of non-conserved blocks

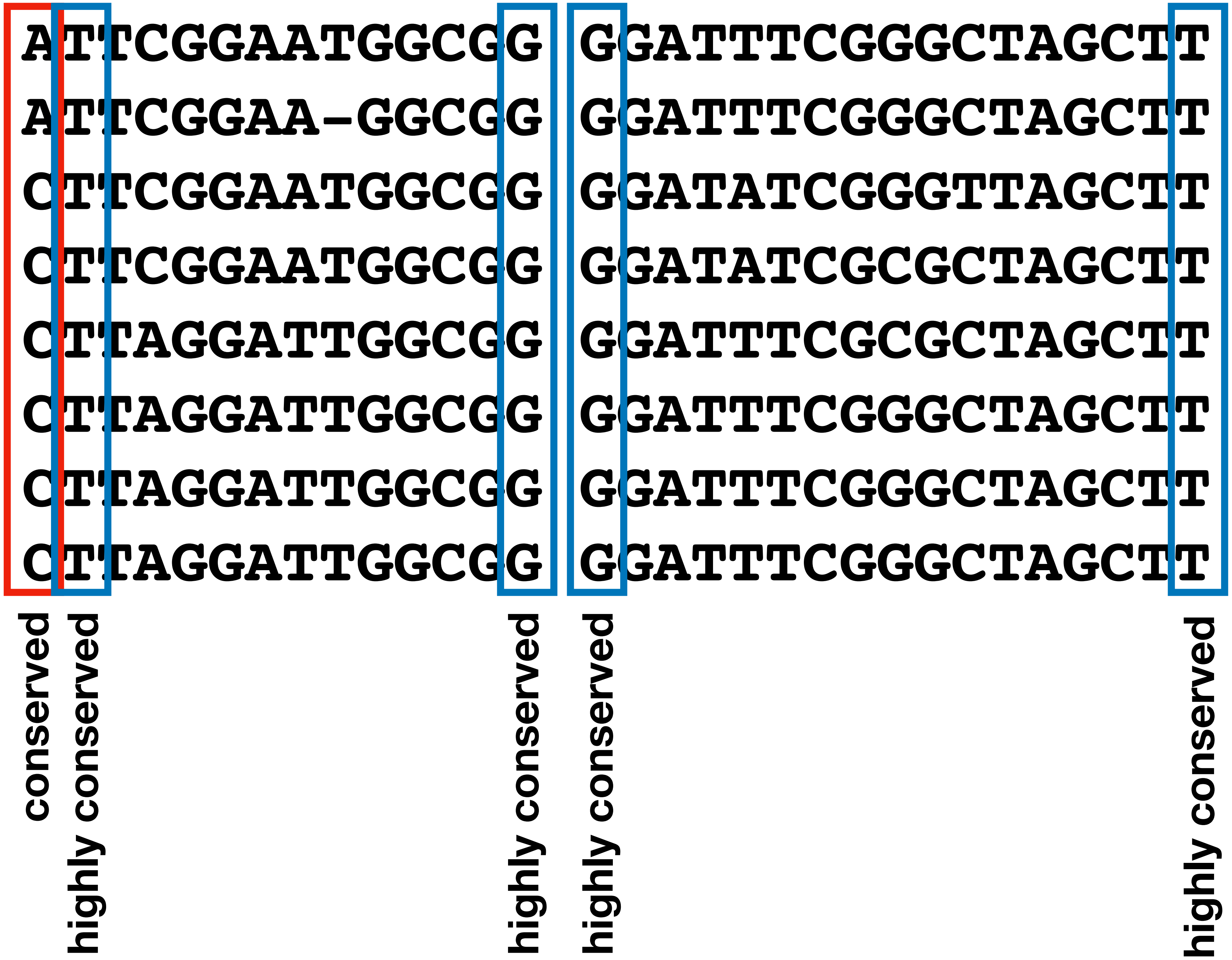
```
ATTCGGAATGGCGG GGATTTCGGGCTAGCTT
ATTCGGAA-GGCGG GGATTTCGGGCTAGCTT
CTTCGGAATGGCGG GGATATCGGGTTAGCTT
CTTCGGAATGGCGG GGATATCGCGCTAGCTT
CTTAGGATTGGCGG GGATTTCGCGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
```

Anchor blocks with highly conserved flanks

```
ATTCGGAATGGCGG GGATTTCGGGCTAGCTT
ATTCGGAA-GGCGG GGATTTCGGGCTAGCTT
CTTCGGAATGGCGG GGATATCGGGTTAGCTT
CTTCGGAATGGCGG GGATATCGCGCTAGCTT
CTTAGGATTGGCGG GGATTTCGCGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
```



Anchor blocks with highly conserved flanks



```
ATTCGGAATGGCGG GGATTTCGGGCTAGCTT
ATTCGGAA-GGCGG GGATTTCGGGCTAGCTT
CTTCGGAATGGCGG GGATATCGGGTTAGCTT
CTTCGGAATGGCGG GGATATCGCGCTAGCTT
CTTAGGATTGGCGG GGATTTCGCGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
CTTAGGATTGGCGG GGATTTCGGGCTAGCTT
```

Anchor blocks with highly conserved flanks

```
TTCGGAATGGCGG GGATTTCGGGCTAGCTT
TTCGGAA-GGCGG GGATTTCGGGCTAGCTT
TTCGGAATGGCGG GGATATCGGGTTAGCTT
TTCGGAATGGCGG GGATATCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
```

Remove short blocks

```
TTCGGAATGGCGG GGATTTCGGGCTAGCTT
TTCGGAA-GGCGG GGATTTCGGGCTAGCTT
TTCGGAATGGCGG GGATATCGGGTTAGCTT
TTCGGAATGGCGG GGATATCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
```

BL1 = 10

minimum length of an initial block

Remove columns with gaps and adjacent non-conserved position

```
TTCGGAATGGCGG GGATTTCGGGCTAGCTT
TTCGGAA-GGCGG GGATTTCGGGCTAGCTT
TTCGGAATGGCGG GGATATCGGGTTAGCTT
TTCGGAATGGCGG GGATATCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
```

Remove columns with gaps and adjacent non-conserved position

```
TTCGGAGGCGG GGATTTCGGGCTAGCTT
TTCGGAGGCGG GGATTTCGGGCTAGCTT
TTCGGAGGCGG GGATATCGGGTTAGCTT
TTCGGAGGCGG GGATATCGCGCTAGCTT
TTAGGAGGCGG GGATTTCGCGCTAGCTT
TTAGGAGGCGG GGATTTCGGGCTAGCTT
TTAGGAGGCGG GGATTTCGGGCTAGCTT
TTAGGAGGCGG GGATTTCGGGCTAGCTT
```

Remove columns with gaps and adjacent non-conserved position

```
TTCGGAGGCGG GGATTTCGGGCTAGCTT
TTCGGAGGCGG GGATTTCGGGCTAGCTT
TTCGGAGGCGG GGATATCGGGTTAGCTT
TTCGGAGGCGG GGATATCGCGCTAGCTT
TTAGGAGGCGG GGATTTCGCGCTAGCTT
TTAGGAGGCGG GGATTTCGGGCTAGCTT
TTAGGAGGCGG GGATTTCGGGCTAGCTT
TTAGGAGGCGG GGATTTCGGGCTAGCTT
```

BL2 = 7

Minimum length of a block after gap cleaning.

Trimmed alignment

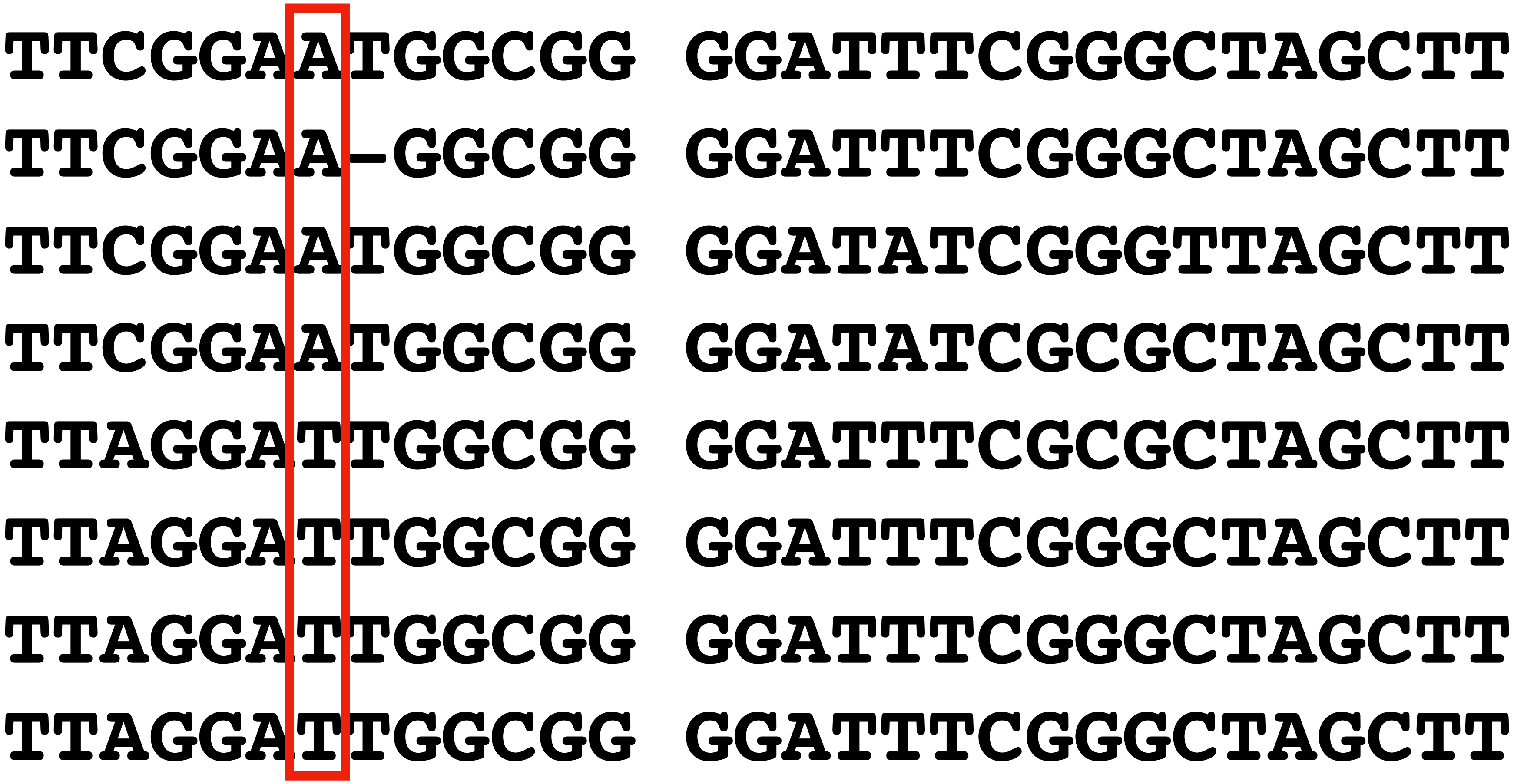
```
TTCGGAGGCGGGGATTTCGGGCTAGCTT
TTCGGAGGCGGGGATTTCGGGCTAGCTT
TTCGGAGGCGGGGATATCGGGTTAGCTT
TTCGGAGGCGGGGATATCGCGCTAGCTT
TTAGGAGGCGGGGATTTCGCGCTAGCTT
TTAGGAGGCGGGGATTTCGGGCTAGCTT
TTAGGAGGCGGGGATTTCGGGCTAGCTT
TTAGGAGGCGGGGATTTCGGGCTAGCTT
```

Pros and cons

Hard to distinguish random aligned positions
from moderate-conserved positions

Excessively trimmed columns with gaps positions

Input too few sequence may not a good idea



```
TTCGGAATGGCGG GGATTTCGGGCTAGCTT
TTCGGAA-GGCGG GGATTTCGGGCTAGCTT
TTCGGAATGGCGG GGATATCGGGTTAGCTT
TTCGGAATGGCGG GGATATCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
```

Pros and cons

Hard to distinguish random aligned positions
from moderate-conserved positions

Excessively trimmed columns with gaps positions

Input too few sequence may not a good idea

```
TTCGGAATGGCGG GGATTTCGGGCTAGCTT
TTCGGAA-GGCGG GGATTTCGGGCTAGCTT
TTCGGAATGGCGG GGATATCGGGTTAGCTT
TTCGGAATGGCGG GGATATCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGCGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
TTAGGATTGGCGG GGATTTCGGGCTAGCTT
```

Pros and cons

Hard to distinguish random aligned positions
from moderate-conserved positions

Excessively trimmed columns with gaps positions

Input too few sequence may not a good idea

4 taxa

IS = 3

FS = 3.4 ~ 4

Pros and cons

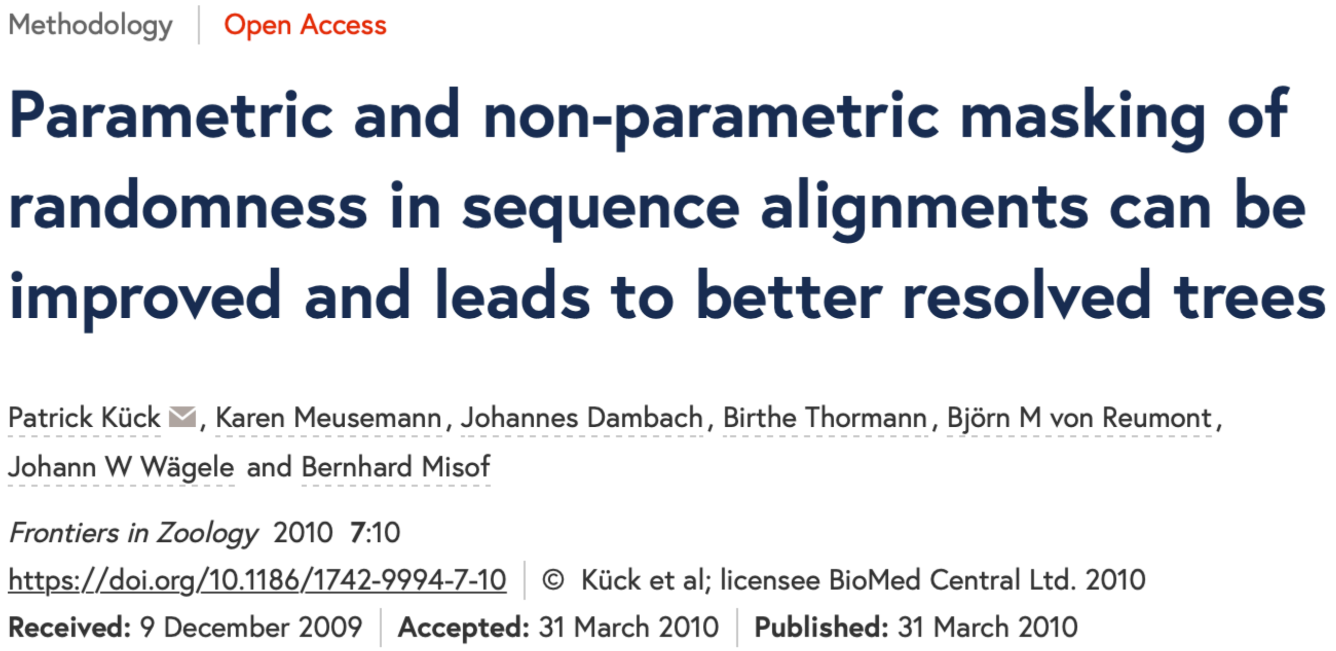
Hard to distinguish random aligned positions from moderate-conserved positions

Excessively trimmed columns with gaps positions

Input too few sequence may not a good idea

It is only suitable to trim bunch of conserved loci by Gblocks

Software accounts for uncertainty in alignments



Methodology | Open Access

Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees

Patrick Kück ✉, Karen Meusemann, Johannes Dambach, Birthe Thormann, Björn M von Reumont, Johann W Wägele and Bernhard Misof

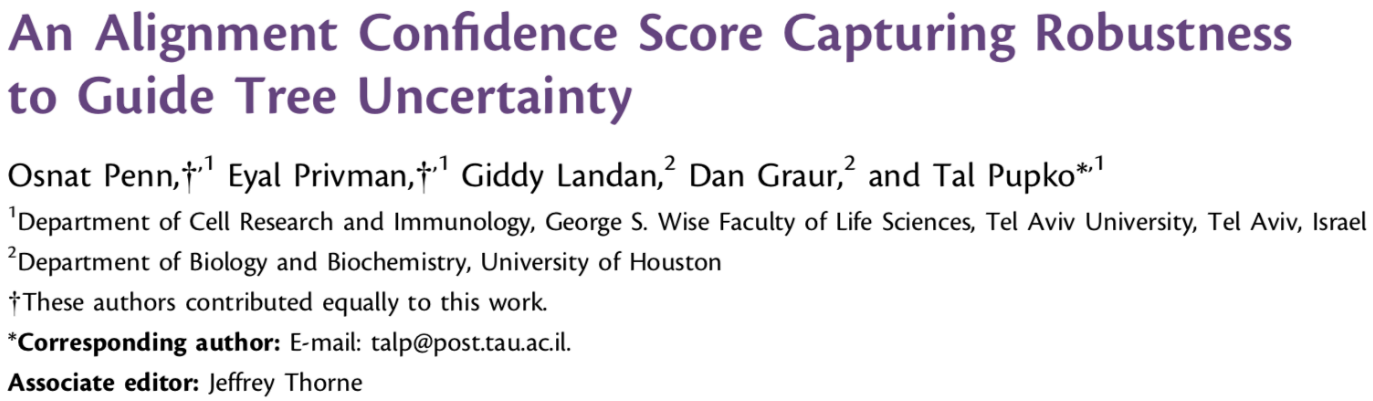
Frontiers in Zoology 2010 **7**:10

https://doi.org/10.1186/1742-9994-7-10 | © Kück et al; licensee BioMed Central Ltd. 2010

Received: 9 December 2009 | **Accepted:** 31 March 2010 | **Published:** 31 March 2010

ALISCORE

whether given alignment position is rejected by random sequences hypotheses



An Alignment Confidence Score Capturing Robustness to Guide Tree Uncertainty

Osnat Penn,†,[1] Eyal Privman,†,[1] Giddy Landan,[2] Dan Graur,[2] and Tal Pupko*,[1]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[2]Department of Biology and Biochemistry, University of Houston

†These authors contributed equally to this work.

***Corresponding author:** E-mail: talp@post.tau.ac.il.

Associate editor: Jeffrey Thorne

GUIDUANCE

whether given alignment position is sensitive to guide trees generated by bootstrapping

...

MARE

MARE (MAtrix REduction) was designed to find informative subsets of genes and taxa within a large phylogenetic dataset

Calculation of the potential information content of genes, taxa and matrix

↓

Reduction to an optimized subset of taxa and genes

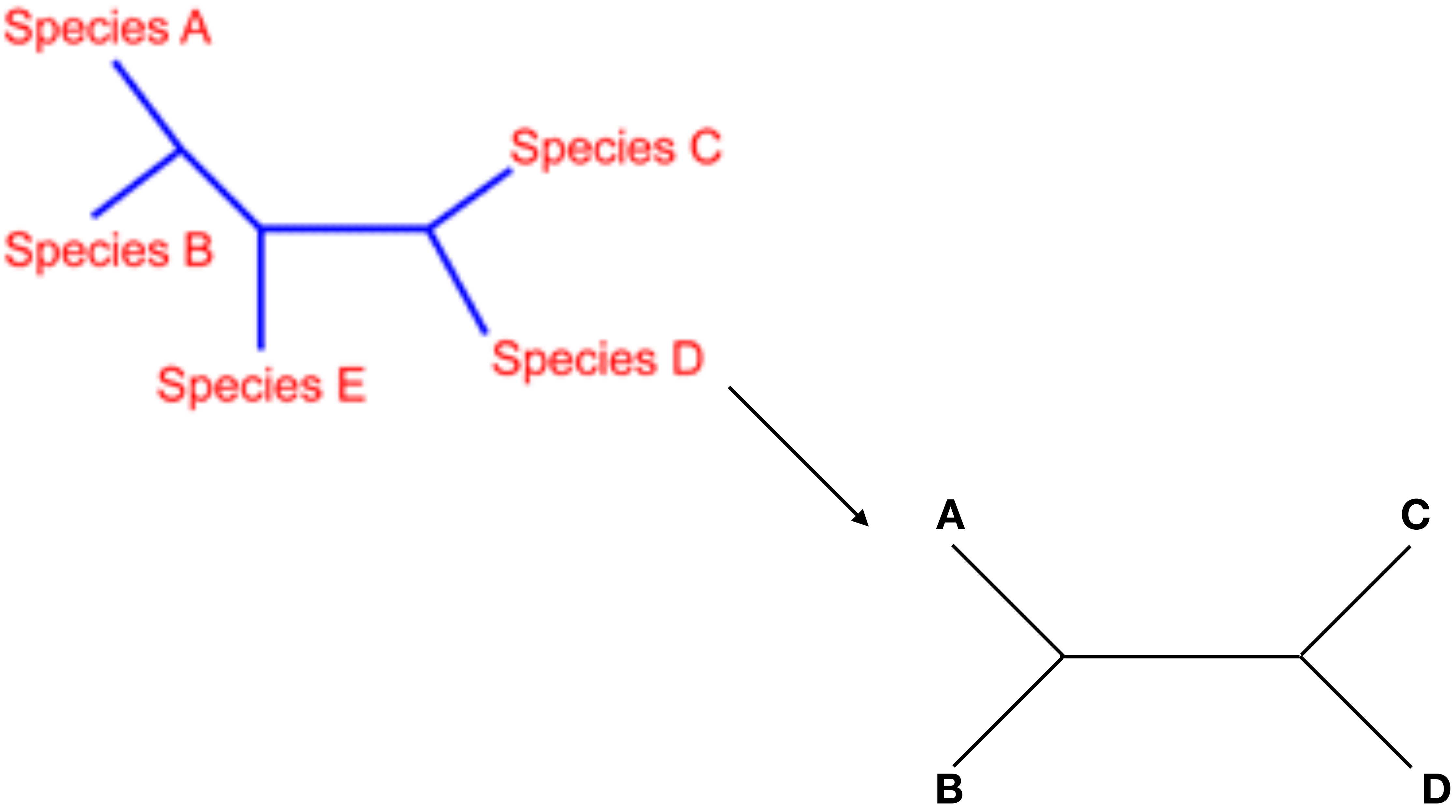
MARE

	Gene 1	Gene 2	Gene 3
Taxon 1	1	1	1
Taxon 2	1	0	1
Taxon 3	1	1	1
Taxon 4	0	1	1

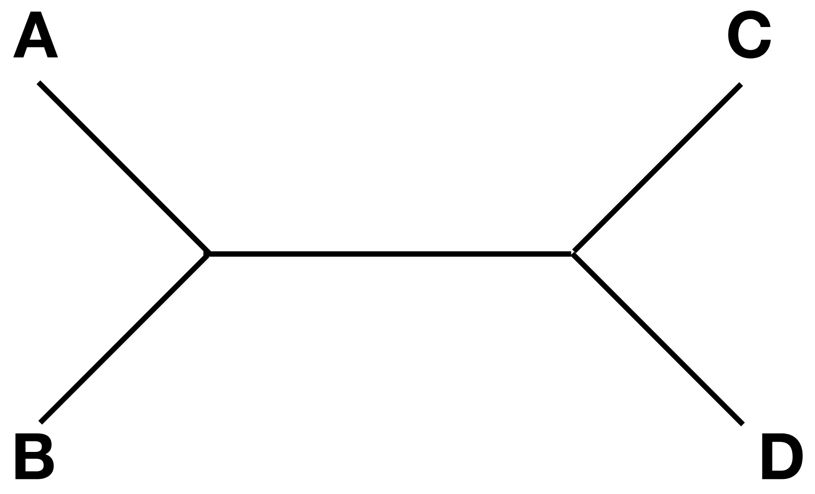
A partition

Arrange sequences into a matrix, 1 as present and 0 as absent

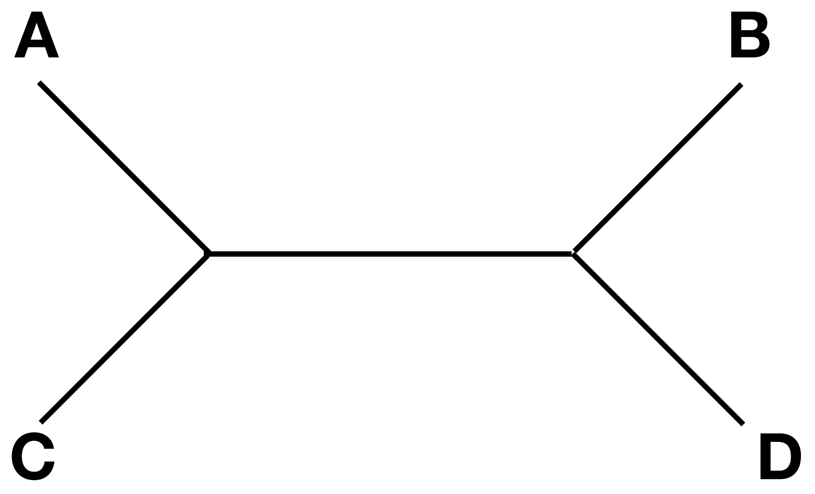
Calculation of the potential information content



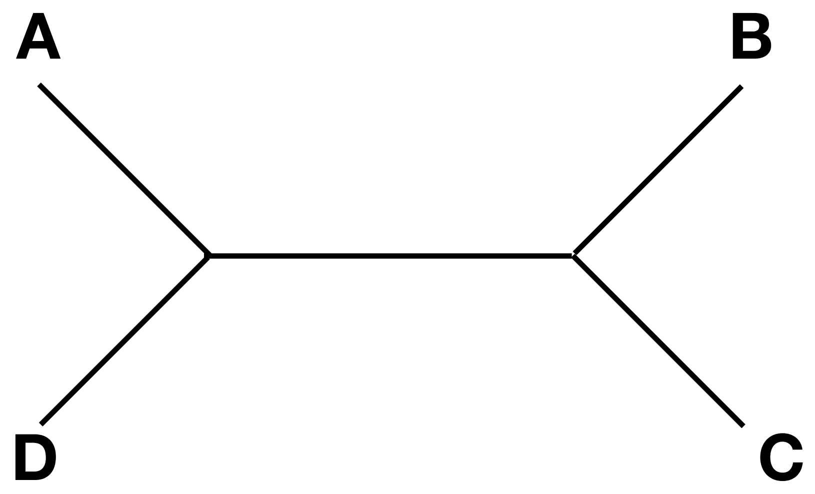
Calculation of the potential information content



P1



P2

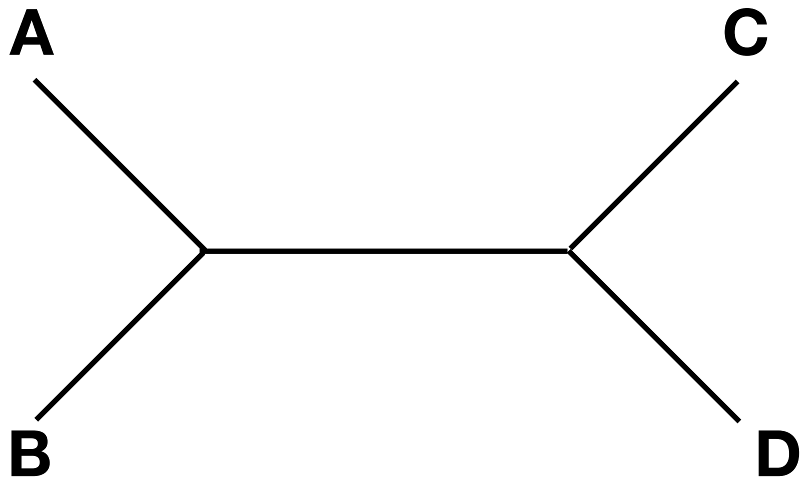


P3

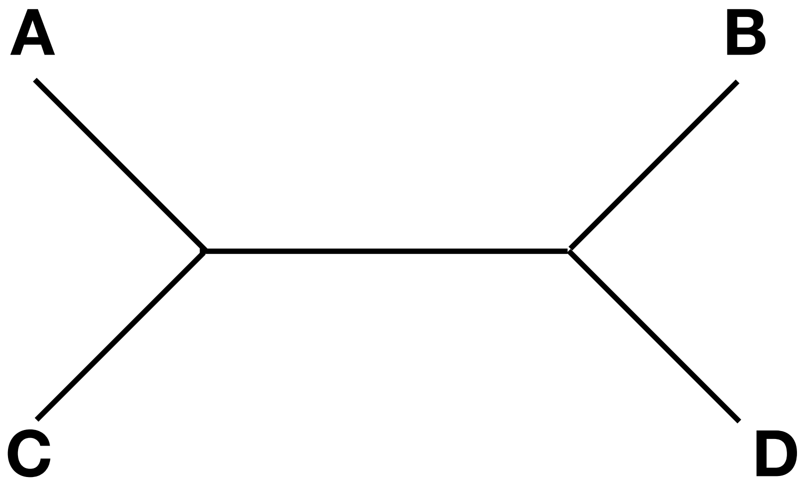
$$Si=\frac{Pi}{P1+P2+P3}$$

Posterior probability of topology i
with given alignments of a partition

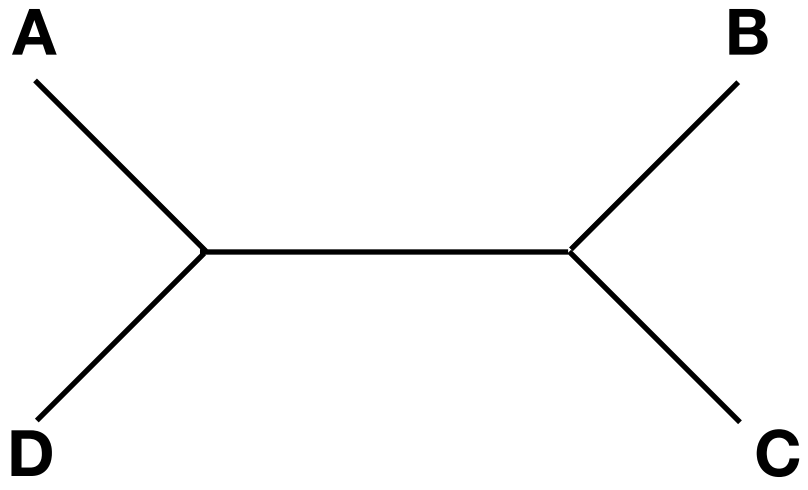
Calculation of the potential information content



P1 = 1



P2 = 0



P3 = 0

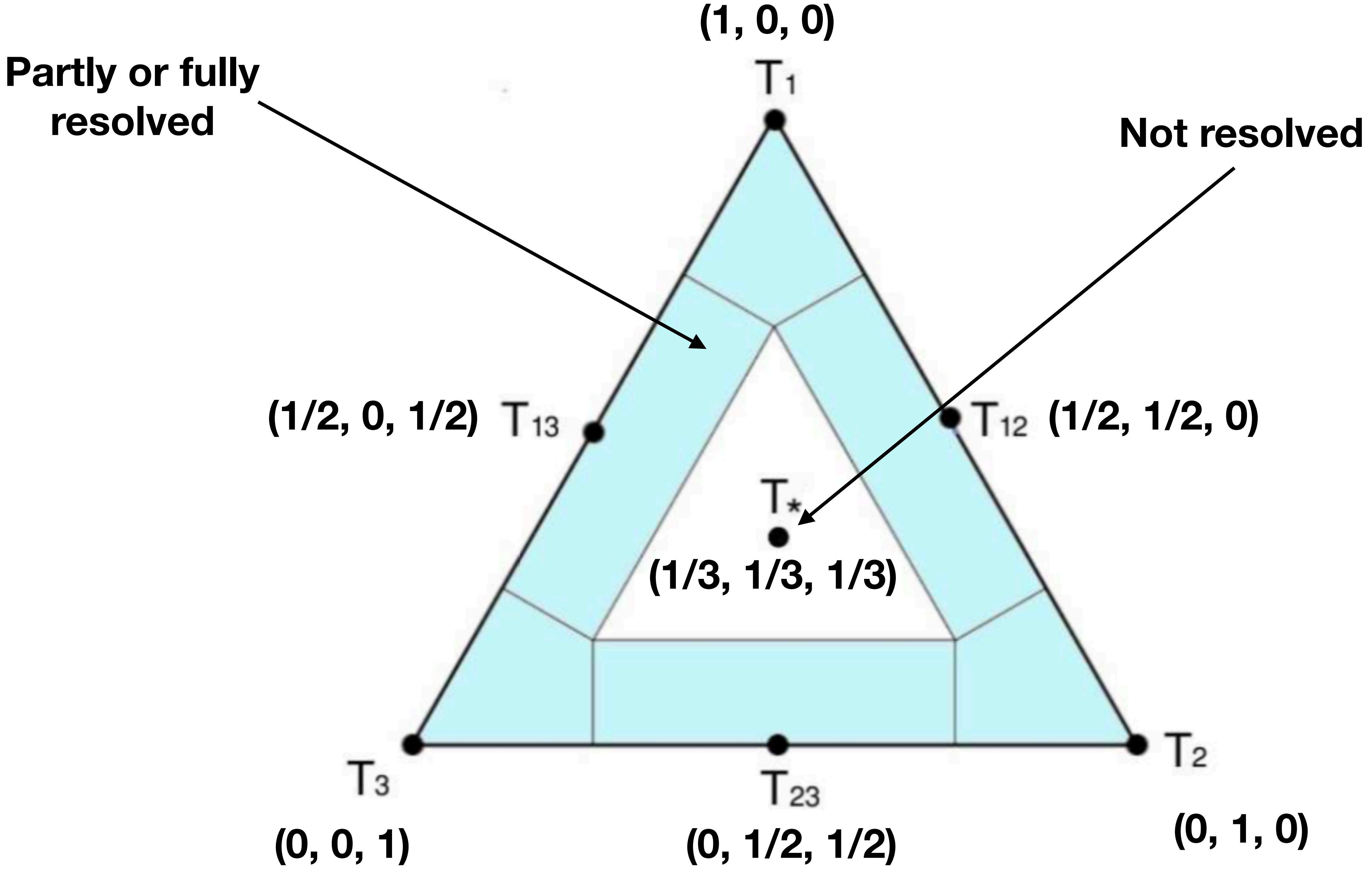
s= (1, 0, 0)

s= (0, 1, 0)

s= (0, 0, 1)

s= (1/3, 1/3, 1/3)

Tree-likeness



(1, 0, 0)

Partly or fully resolved

Not resolved

(1/2, 0, 1/2)

(1/2, 1/2, 0)

(1/3, 1/3, 1/3)

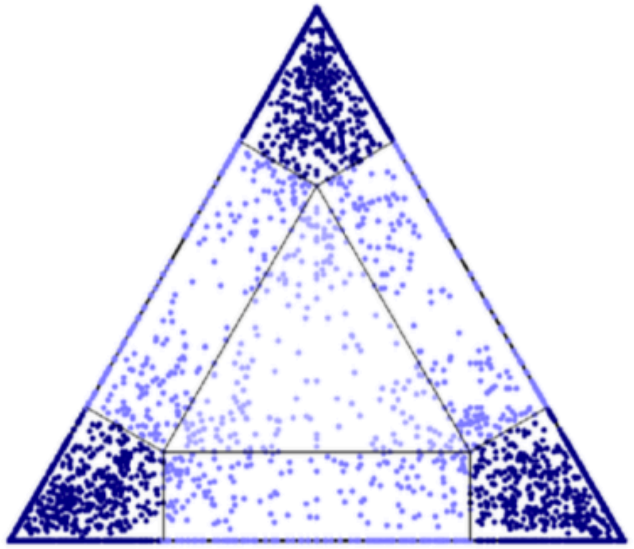
(0, 0, 1)

(0, 1/2, 1/2)

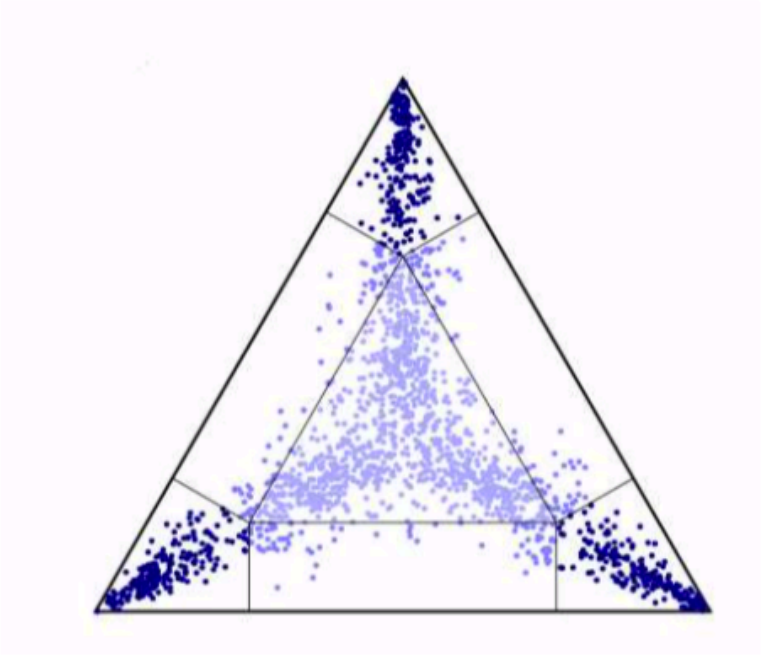
(0, 1, 0)

$$\text{t (Tree-likeness)} = \frac{\text{Number of point in blue}}{\text{All point}}$$

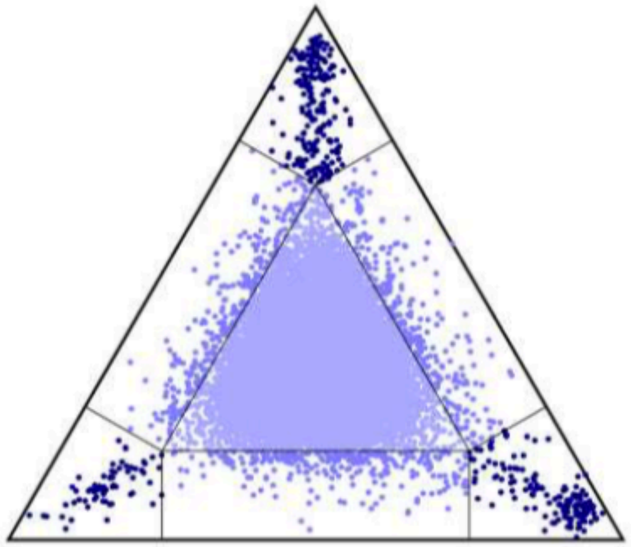
Tree-likeness



0.94



0.46



0.06

Calculation of the potential information content

	Gene 1	Gene 2	Gene 3
Taxon 1	0.94	0.05	0.46
Taxon 2	0.94	0	0.46
Taxon 3	0.94	0.05	0.46
Taxon 4	0	0.05	0.46

$$\text{Information content of taxon1} = \frac{0.94+0.05+0.46}{3}$$

$$\text{Information content of gene1} = \frac{0.94+0.94+0.94+0}{4}$$

$$\text{Information content of matrix} = \frac{\text{Information content of gene1} + \text{Information content of gene2} + \text{Information content of gene3}}{3}$$

Calculation of the potential information content

	Gene 1	Gene 2	Gene 3
Taxon 1	0.94	0.05	0.46
Taxon 2	0.94	0	0.46
Taxon 3	0.94	0.05	0.46
Taxon 4	0	0.05	0.46

$$\text{Information content of taxon1} = \frac{0.94+0.05+0.46}{3}$$

$$\text{Information content of gene1} = \frac{0.94+0.94+0.94+0}{4}$$

$$\text{Information content of matrix} = \frac{\text{Information content of gene1} + \text{Information content of gene2} + \text{Information content of gene3}}{3}$$

Calculation of the potential information content

	Gene 1	Gene 3
Taxon 1	0.94	0.46
Taxon 2	0.94	0.46
Taxon 3	0.94	0.46
Taxon 4	0	0.46

Information content
of updated taxon1

...

...

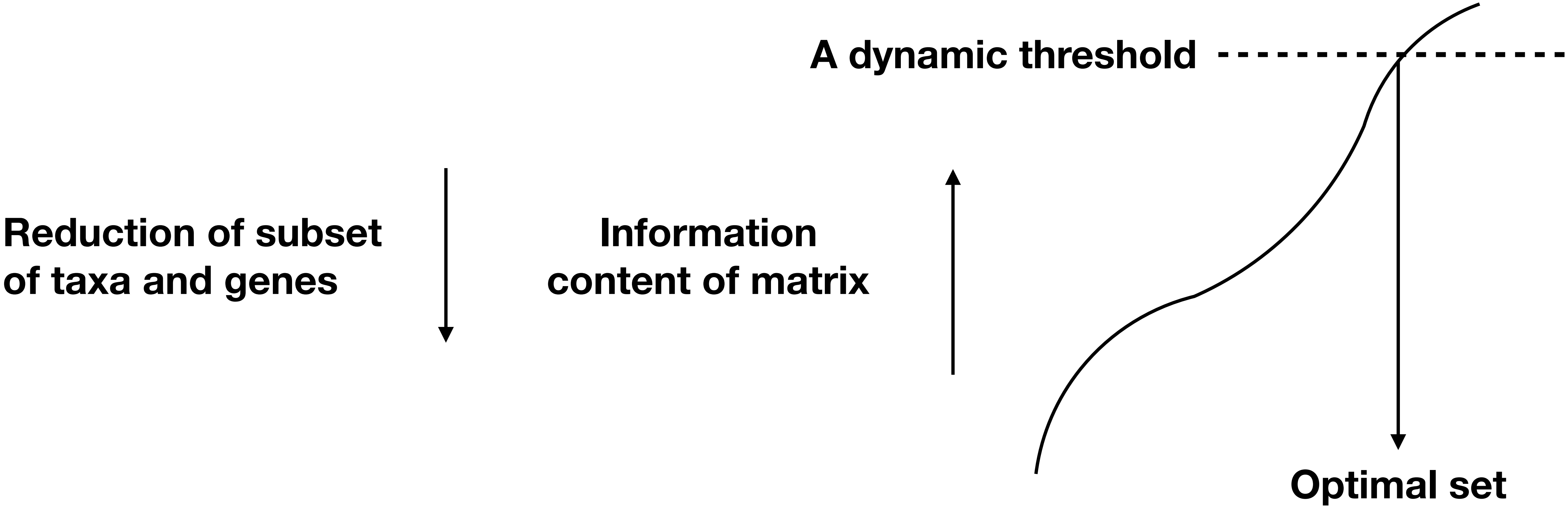
...

Information content
of updated gene1

...

Information
content of matrix

Reduction to an optimized subset of taxa and genes

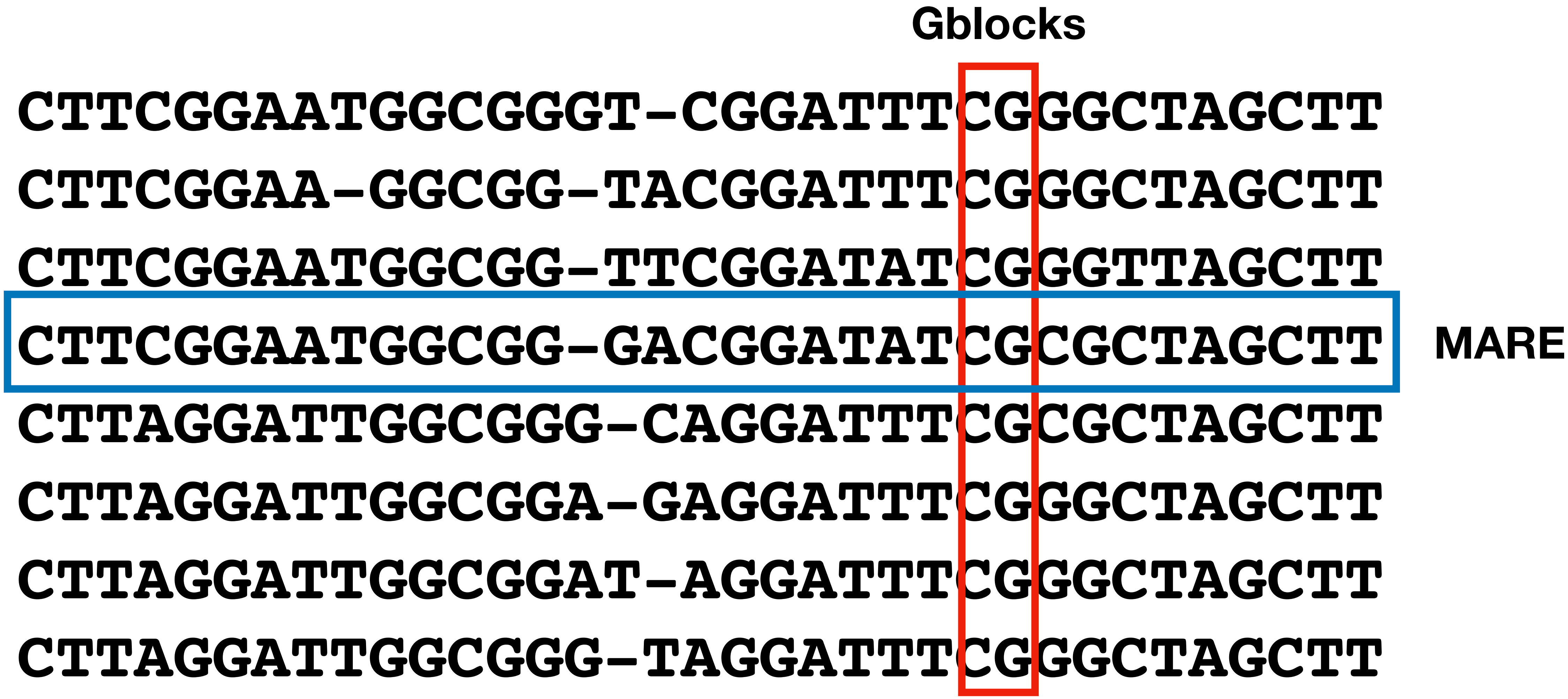


A dynamic threshold

Reduction of subset of taxa and genes

Information content of matrix

Optimal set



```
CTTCGGAATGGCGGGT-CGGATTTCGGGCTAGCTT
CTTCGGAA-GGCGG-TACGGATTTCGGGCTAGCTT
CTTCGGAATGGCGG-TTCGGATATCGGGTTAGCTT
CTTCGGAATGGCGG-GACGGATATCGCGCTAGCTT
CTTAGGATTGGCGGG-CAGGATTTCGCGCTAGCTT
CTTAGGATTGGCGGA-GAGGATTTCGGGCTAGCTT
CTTAGGATTGGCGGAT-AGGATTTCGGGCTAGCTT
CTTAGGATTGGCGGG-TAGGATTTCGGGCTAGCTT
```