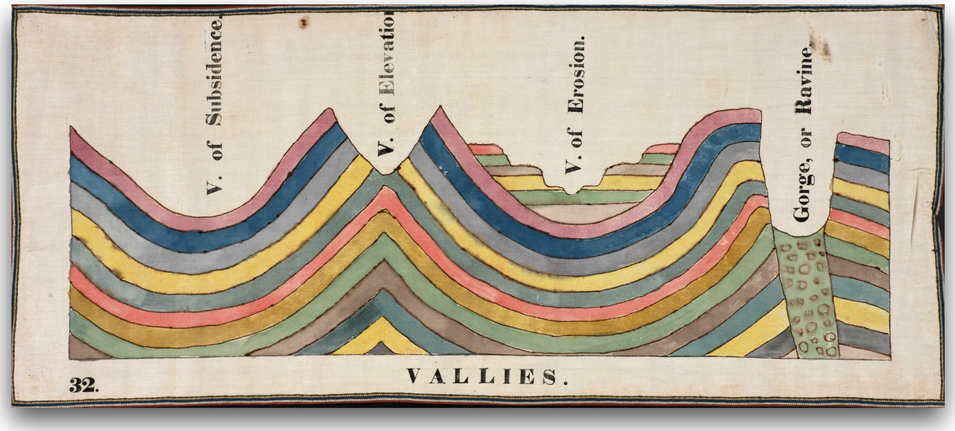
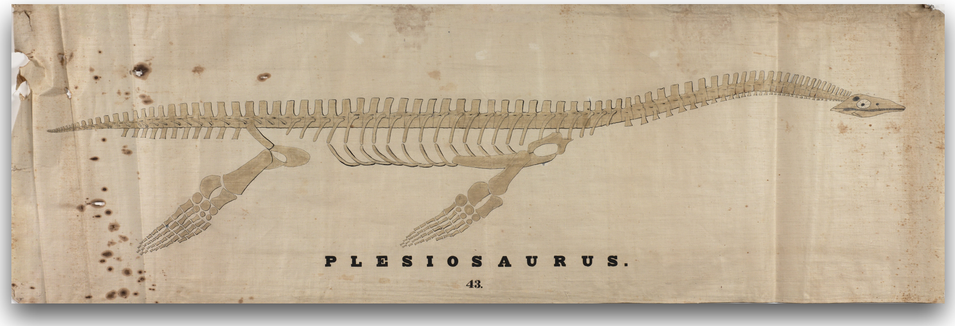
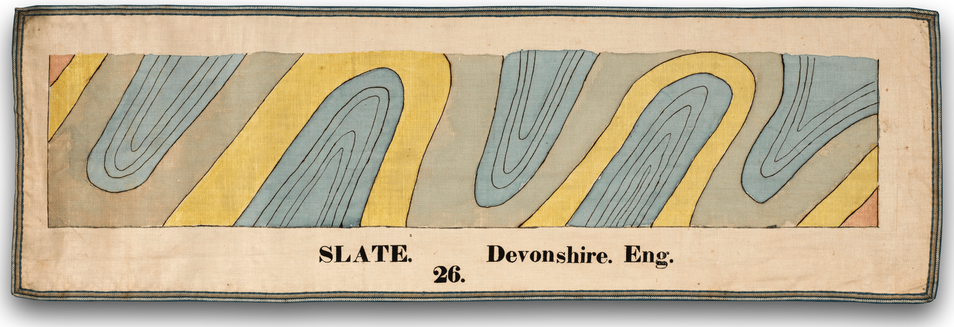
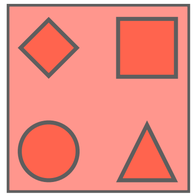






Sequence alignement and filtering

OK, I have my orthogroups … 😎



and now what?!

orthologous genes & orthologous sites!

```
species 1   AGGATCTGCAATTGCTCTTCTAATCTGTCTGATCAGGAT
species 2   AGG------AATTGCTCTTCTAATCTGTCT---CAGGAT
species 3   AGGATCTGCAATTGC---TCTAATCTGTCTGATCAGGAT
species 4   AGAATCTGCAATTGCTCTTCTGATCTGTCTGATCACGAT
species 5   AGGATCTGC---TGCTCTTCTGATCTGTCTGATCAGGAT
```

The goal of is to identify which **positions** are **orthologous**.

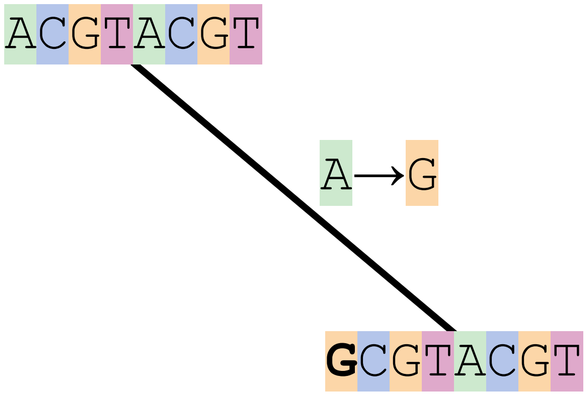
That is, their **evolutionary history is the same as that of the species**.

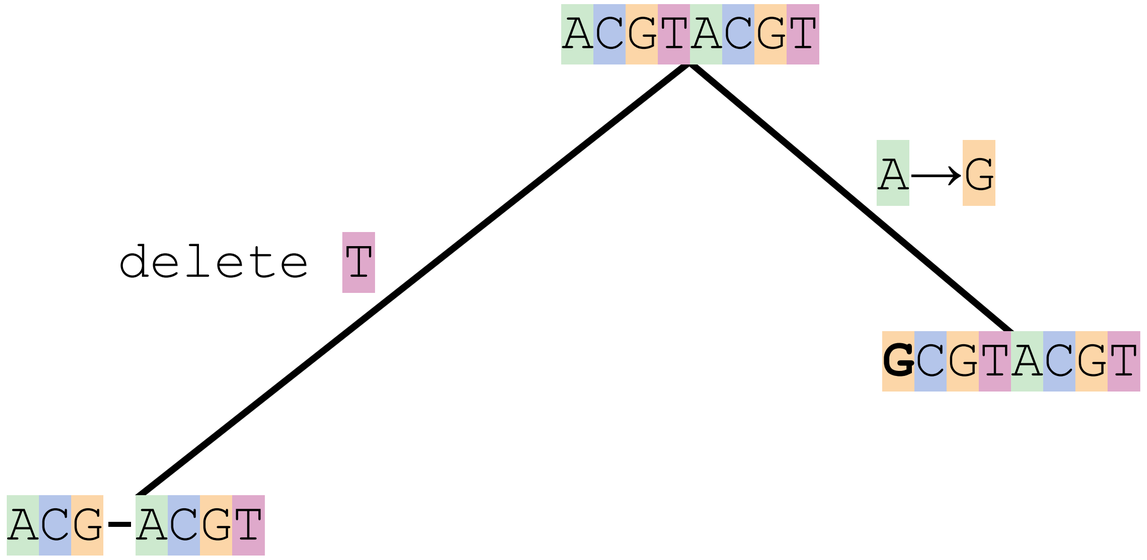
alignements (of nucleotides and proteins)

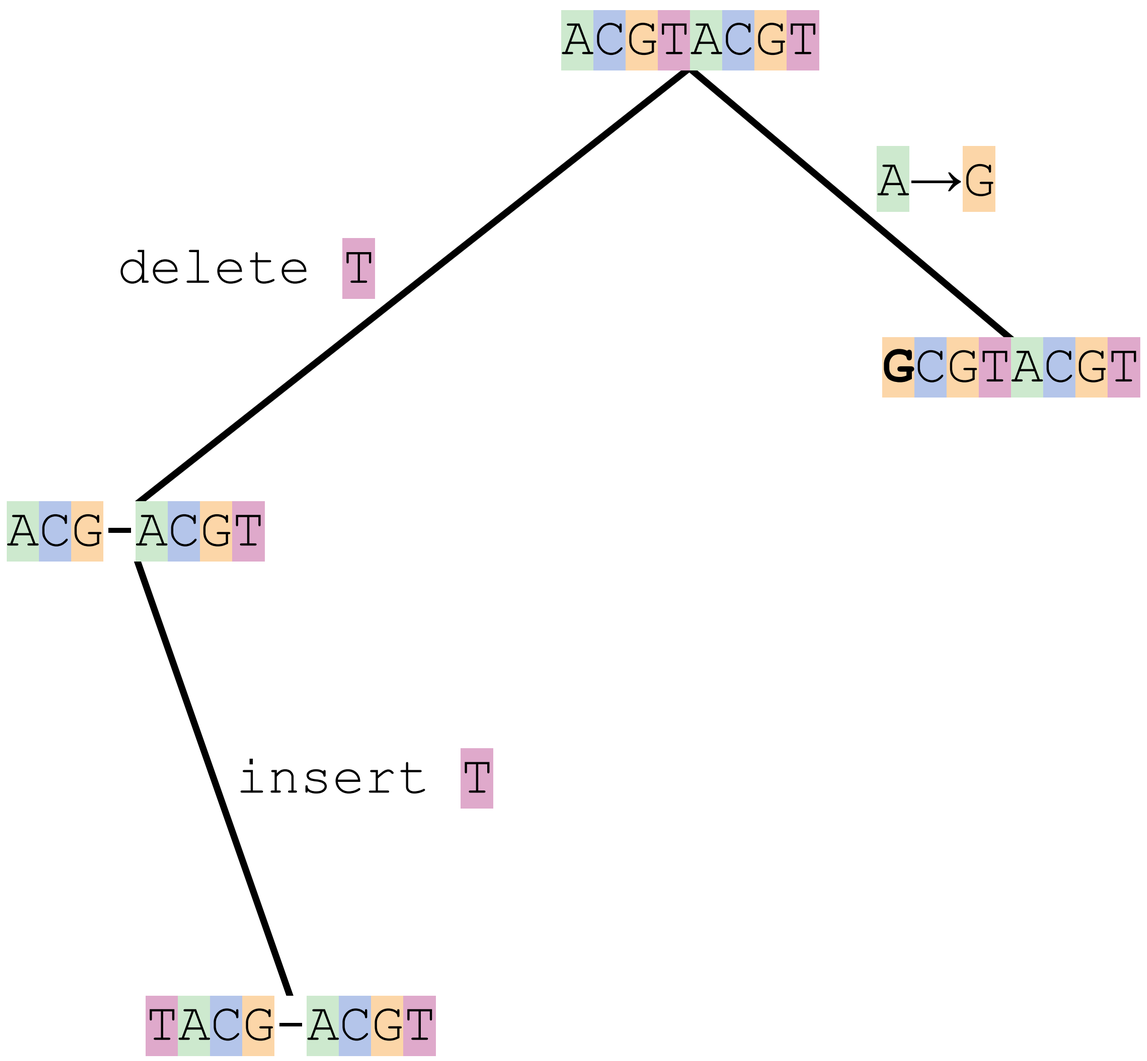
```
species 1   AGGATCTGCAATTGCTCTTCTAATCTGTCTGATCAGGAT
            ValArgSerCysSerCysValArgSerCysValValSer
species 2   AGG------AATTGCTCTTCTAATCTGTCT---CAGGAT
            Val------CysSerCysValArgSerCys---ValSer
species 3   AGGATCTGCAATTGC---TCTAATCTGTCTGATCAGGAT
            ValArgSerCysSer---ValArgSerCysValValSer
species 4   AGAATCTGCAATTGCTCTTCTGATCTGTCTGATCACGAT
            TrpArgSerCysSerCysValCysSerCysValArgSer
species 5   AGGATCTGC---TGCTCTTCTGATCTGTCTGATCAGGAT
            ValArgSer---SerCysValCysSerCysValValSer
```

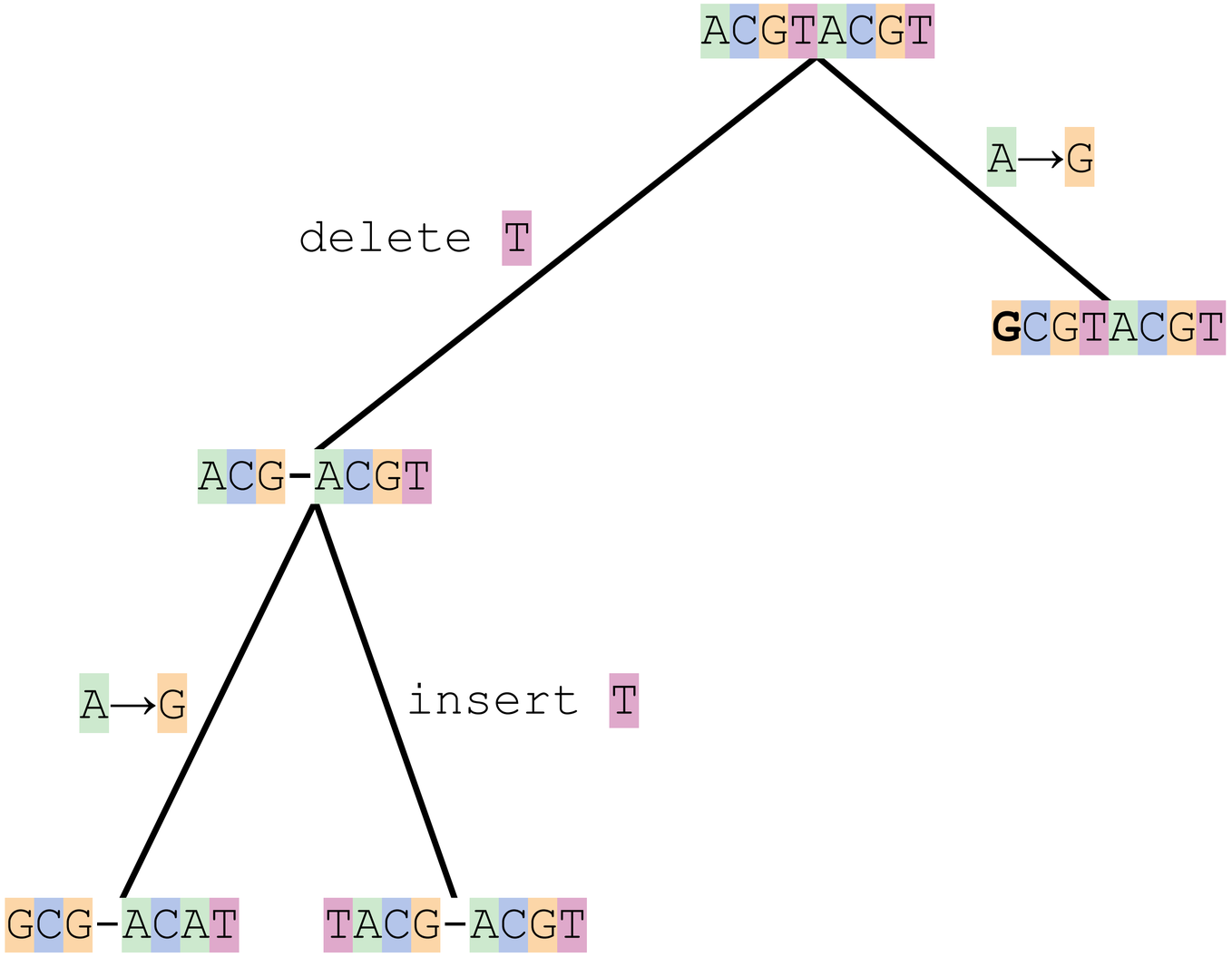
sequences evolve on a tree

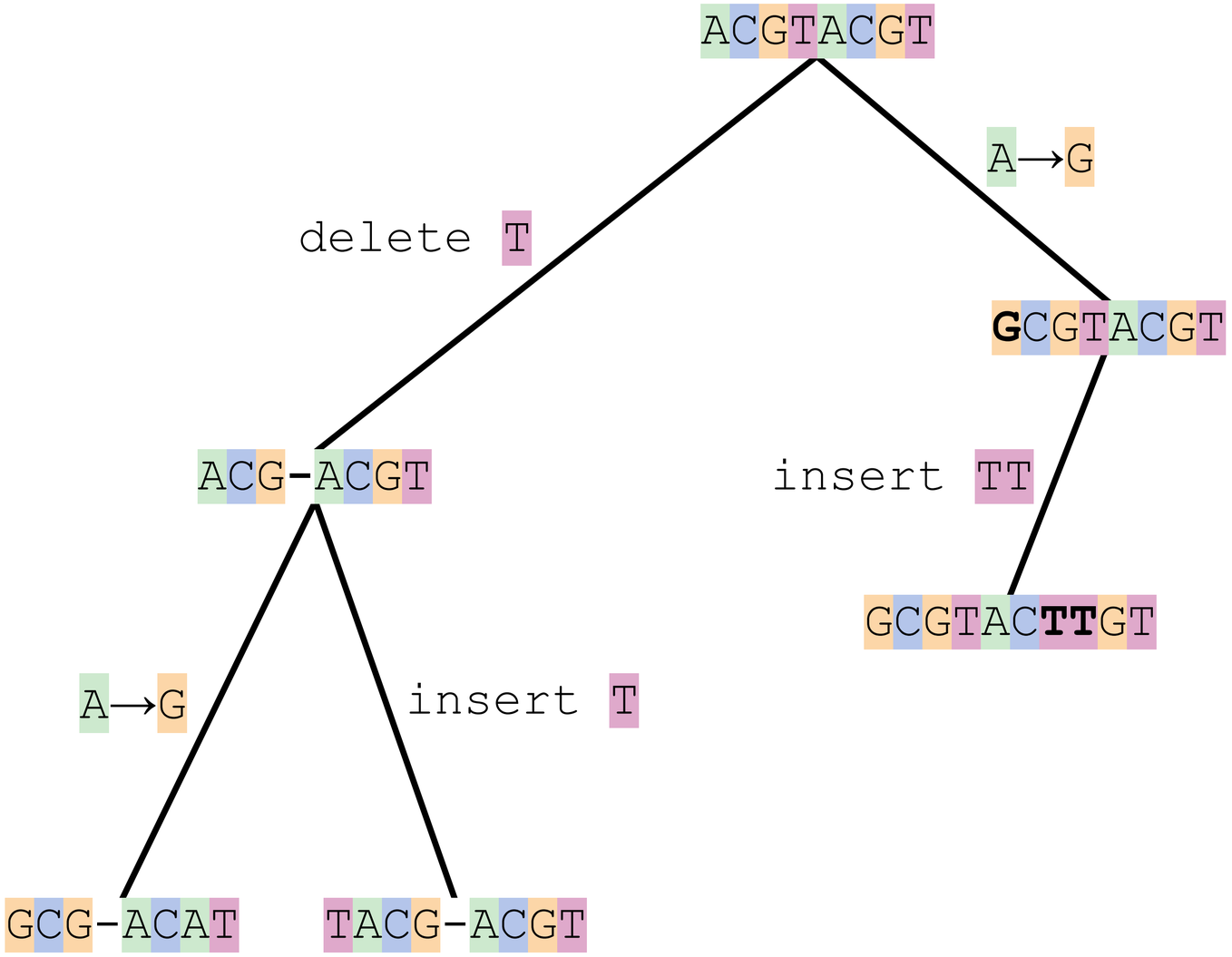
ACGTACGT

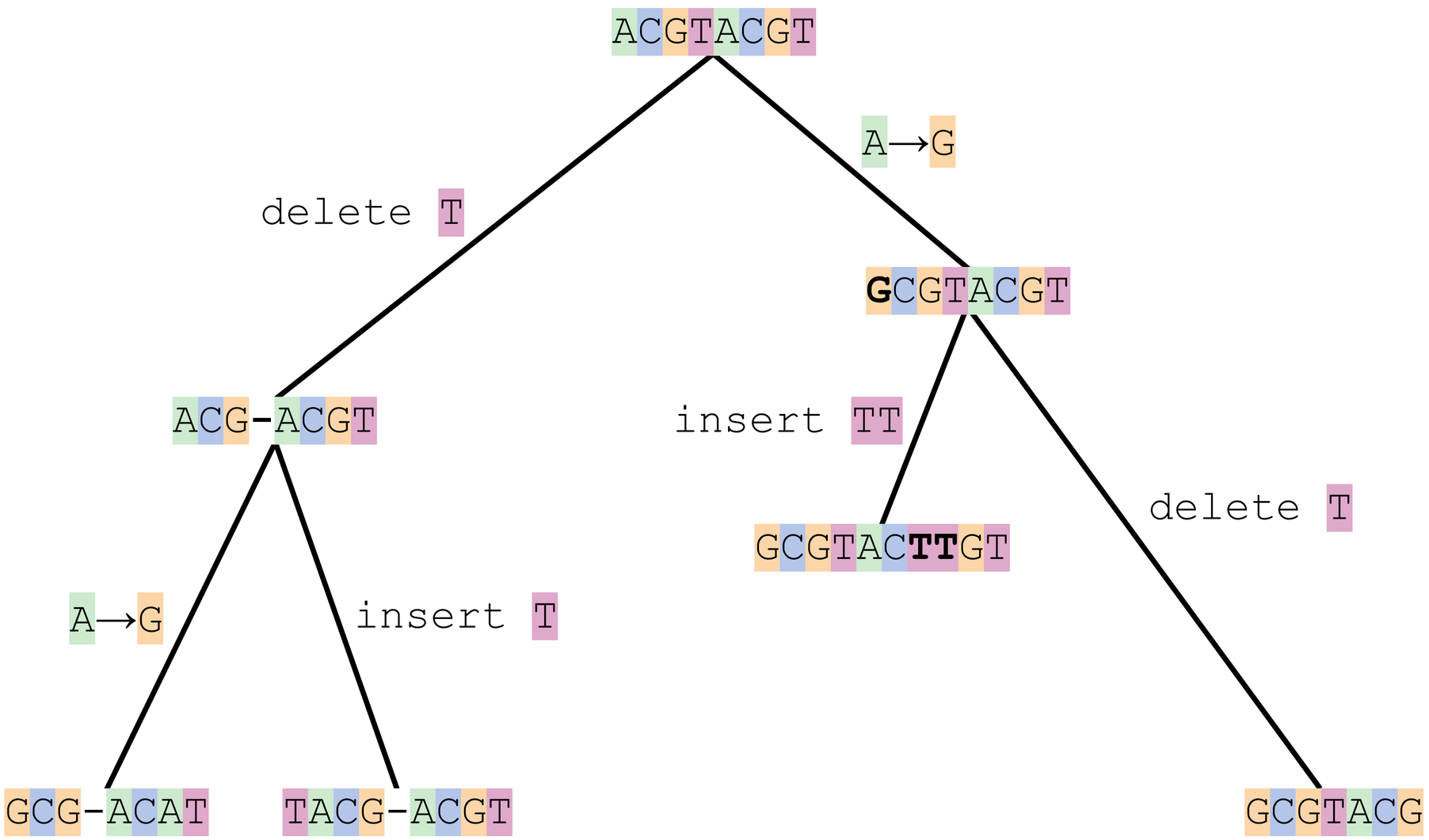

ACGTACGT
A→G
GCGTACGT

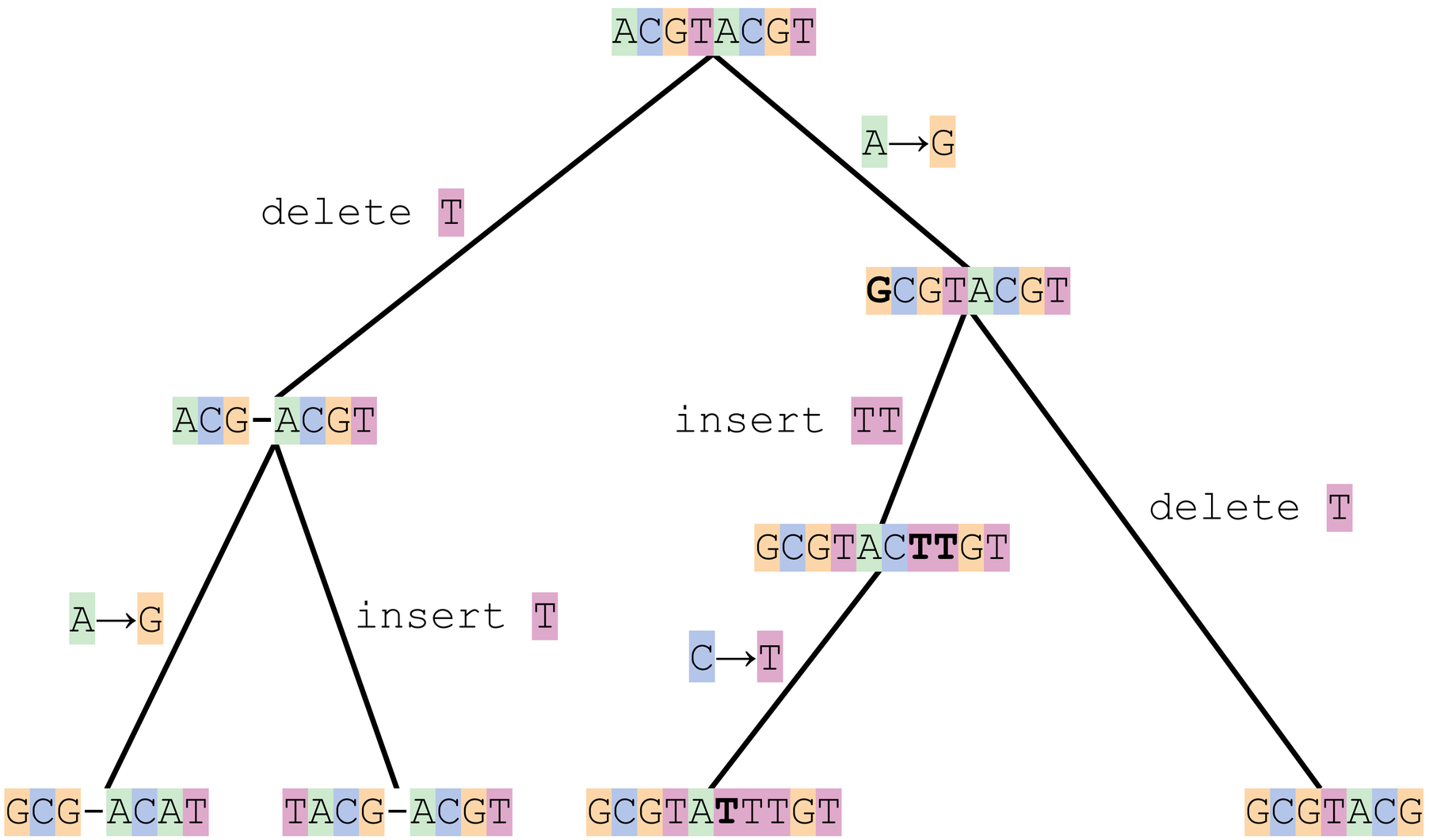

ACGTACGT
A→G
delete T
GCGTACGT
ACG-ACGT

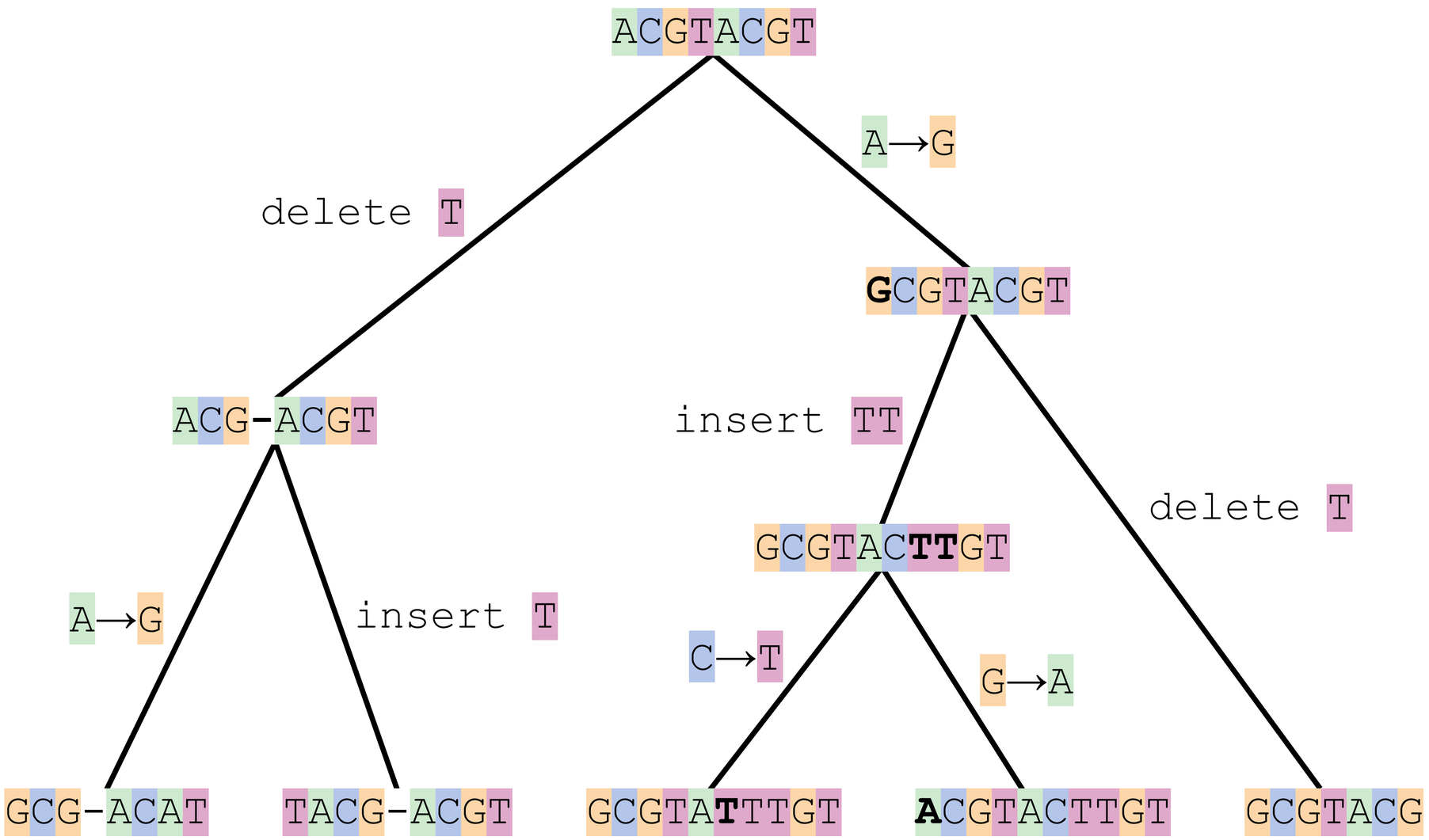

ACGTACGT
A→G
delete T
GCGTACGT
ACG-ACGT
insert T
TACG-ACGT


ACGTACGT
A→G
delete T
GCGTACGT
ACG-ACGT
A→G
insert T
GCG-ACAT
TACG-ACGT


ACGTACGT
delete T
A→G
GCGTACGT
ACG-ACGT
insert TT
GCGTACTTGT
A→G
insert T
GCG-ACAT
TACG-ACGT


ACGTACGT
delete T
A→G
ACG-ACGT
GCGTACGT
insert TT
delete T
GCGTACTTGT
A→G
insert T
GCG-ACAT
TACG-ACGT
GCGTACG


ACGTACGT
delete T
A→G
ACG-ACGT
GCGTACGT
insert TT
delete T
A→G
insert T
GCGTACTTGT
C→T
GCG-ACAT
TACG-ACGT
GCGTATTTGT
GCGTACG


ACGTACGT
delete T
A→G
ACG-ACGT
GCGTACGT
insert TT
delete T
A→G
insert T
GCGTACTTGT
C→T
G→A
GCG-ACAT
TACG-ACGT
GCGTATTTGT
ACGTACTTGT
GCGTACG

These are the sequences sampled at present!

```
species 1   GCGACAT
species 2   TACGACGT
species 3   GCGTATTTGT
species 4   ACGTACTTGT
species 5   GCGTACG
```

This is the “true” alignement, which is the one where the nucleotides are arranged to be orthologous.

```
species 1   -GCG-AC-AT-
species 2   TACG-AC-GT-
species 3   -GCGTATTTGT
species 4   -ACGTACTTGT
species 5   -GCGTAC--G-
```

… but also this alignement does not look too bad 🙃!

```
species 1   -GCG-AC----A----T
species 2   ----TAC--G-AC--GT
species 3   -GCGTA-TT------TGT
species 4   A-CGTACTTGT-------
species 5   -GCGTAC--G--------
```

Which alignement is better?

This can be a bit of an art.

At its core you need to balance the number of gaps with number of mismatches

What price for a gap?

Gap penalties scoring:
There are two basic methods for assigning a cost ***c*** to a gap of length ***g*** in a sequence.

- Linear cost: $c=-dg$, where ***d*** is the gap open penalty
- Affine cost: $c=-d-(g-1)e$, where ***e*** is the gap extension penalty.

Typical values are d=10 and e=0.1.

Nucleotide substitution scoring:

- match (1)
- mismatch (-1)

alignment scores calculation:
Sum of match/mismatch scores
Minus gap penalties

There are two types of alignments: **pairwise alignments** and **multiple sequence alignments**. Each type of alignment has their own methods and algorithm, but they both attempt to maximize the similarity between sequences by inserting gaps when necessary to improve overall alignment.

Pairwise Alignment

- **Global alignment:** Needleman-Wunsch Algorithm

Aligns entire sequences from end to end. This algorithm is ideal when query sequences are of similar length and are expected to share homology across their full length.

- **Local Alignment:** Smith-Waterman Algorithm

Aligns specific regions of sequences rather than the full length. It is ideal when sequences are globally dissimilar but contain localized similarities, such as motifs or conserved domains.

Profile Alignment

1. Pairwise Alignments: Compute pairwise sequence distances.
2. Guide Tree Construction: Build a tree based on sequence similarity.
3. Progressive Alignment: Align the closest sequences first, then extend to profiles.
4. Final Refinement: Optimize alignment to improve accuracy.

Definition: The Hamming distance measures the number of substitutions (point differences) between two sequences of equal length. Key Characteristics: Only considers exact mismatches (substitutions). Ignores gaps (indels). Requires sequences to be of the same length.

Pairwise Distance 🌍 Definition: A more general measure of genetic distance that accounts for substitutions, insertions, and deletions (indels).

The goal of the alignment process is to identify evolutionary events associated with homologies and ensure aligned sequences accurately reflect biological relationships.

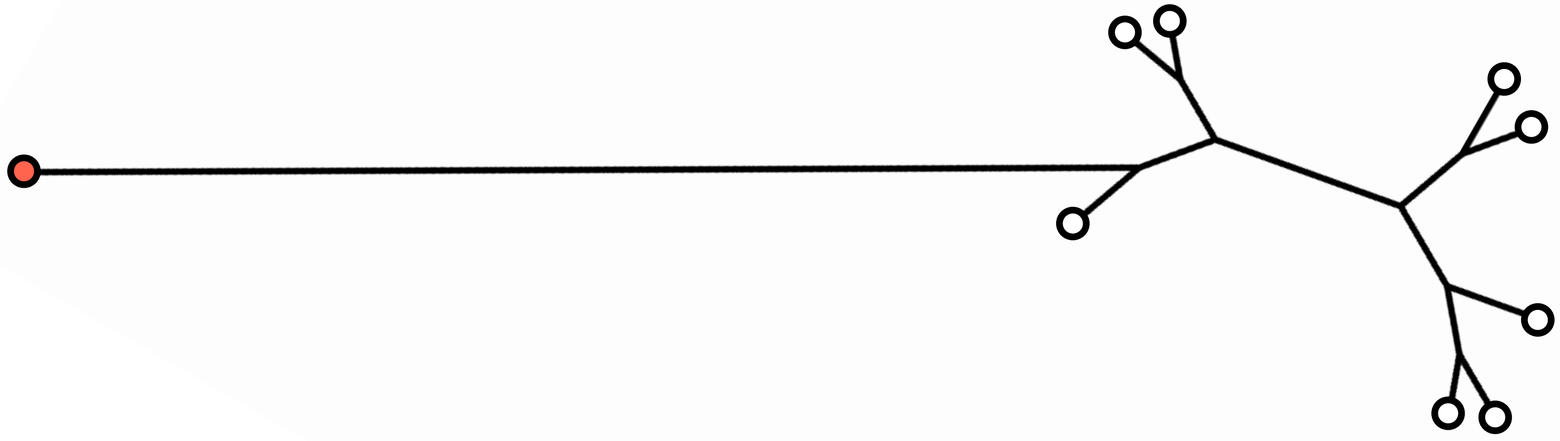
However:
- not all part of the sequences might be homologous
- for some part of the sequences homolgy might be not reconstructed correctly

If the sequences are poorly aligned, you may want to consider trimming the poorly aligned areas.

```
species 1  GCGTATTTGTCGCGAAAATCCCACGAA-GCG-AC-AT-C-AG--ATT-TATGAATCGACATG
species 2  GCGTATTTGTCGCGAAAATCCCACGAATA-G-AG-GT-G--TC-TG---ATGAATCCACATG
species 3  GCGTATTTGTCGCGAAAATCCCACGAA-GC---T-TGCC-AC--AG--CATGAATCGACATA
species 4  GCGTATTTGTCGCGAAAATCTCACGAA-ATGT-CTT-TC-AT--CT--CATGAATCGACATG
species 5  GCGTATTTGTCGCGAAAATCCCACGAA-GCTTAC--GC-AT--TAC--CATGAATCGACATG
```

Notable tools for the process are:

- **Gblocks** (Talavera & Castresana 2007)
- **Aliscore** (Kück et al. 2014)
- **BMGE** (Criscuolo & Gibaldo 2010)



Furthermore, **entire alignements** and **entiree sequences** witin alignements can be filtered out.

Several custom approaches are possible, but there is one which I like a lot:

exclude

We will see more on phylogenetic subsamling (i.e. the choiche of genes and markers) in the lessono on biases

FINISH