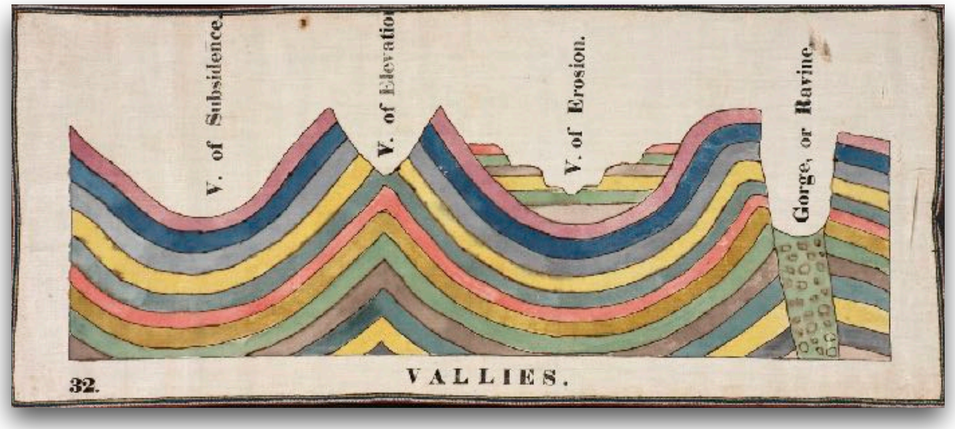
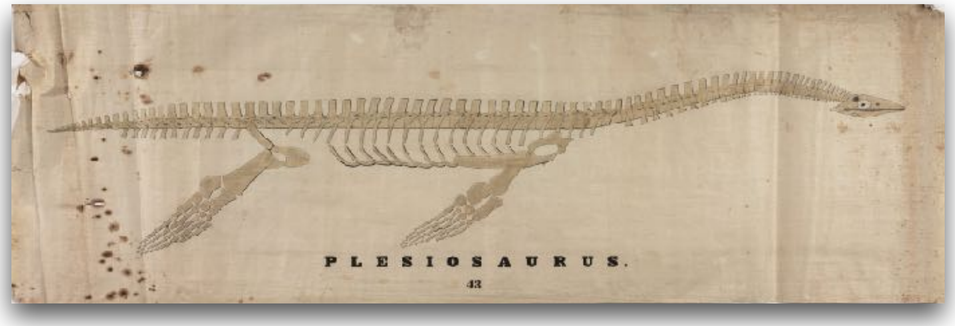




Sequence alignement and filtering

Orthology inferences are inferred **pairwise**.

When considering **multiple species**,
we should move to the concept of … **orthogroup**.

An orthogroup is defined by a **reference speciation event**.

An orthogroup is defined as a **set of homologous genes** that descend **from a single gene in the last common ancestor** of species considered.

An **orthogroup** contains:

Orthologs – genes of different species that **evolved from a single ancestral gene** through **speciation events**.

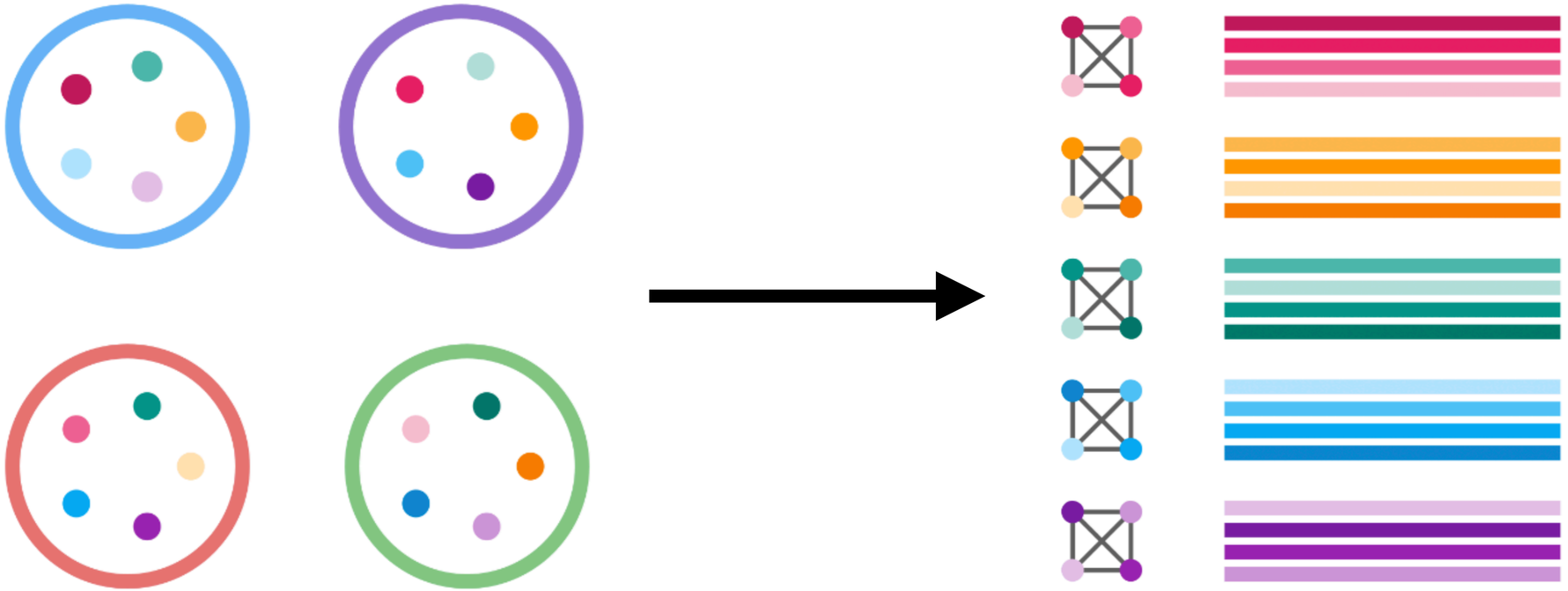
In-paralogs – gene duplicates that **occurred after the last common ancestor of the species considered**.

BEWARE:

Orthogroups **do not separate orthologs** and **in-paralogs** but group all genes descended from a single one in the LCA of the studied species.

… **out-paralogs** - gene duplications that occurred before the LCA - are **not included in an orthogroup**, as they belong to a different orthogroup.

OK, I have my orthogroups … 😎



and now what?!

orthologous genes…
homologous sites!

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

The goal of alignement is to identify which **positions** are **homologous**.

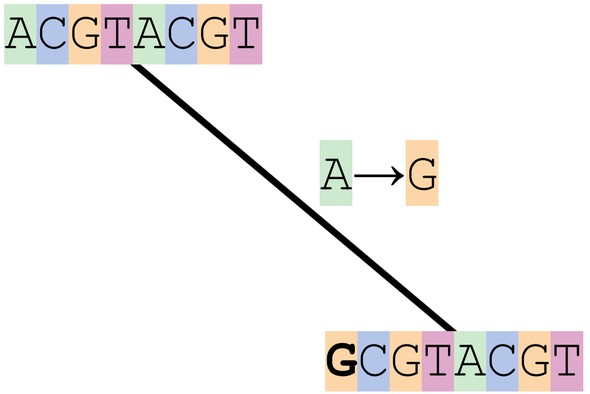
That is, their **evolutionary history** reflect the **species relationships**!

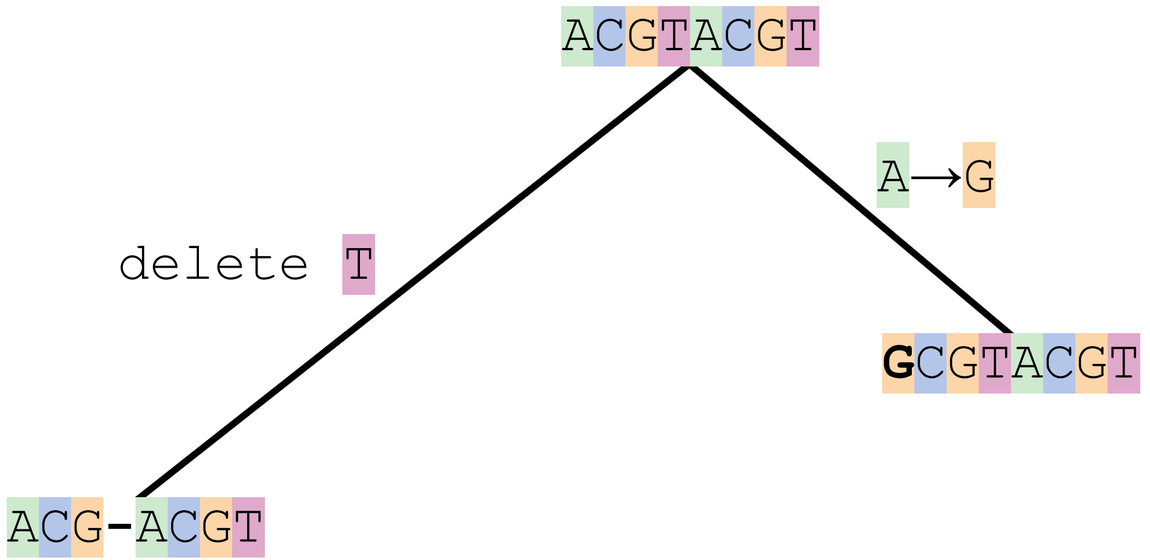
alignements (of nucleotides and proteins)

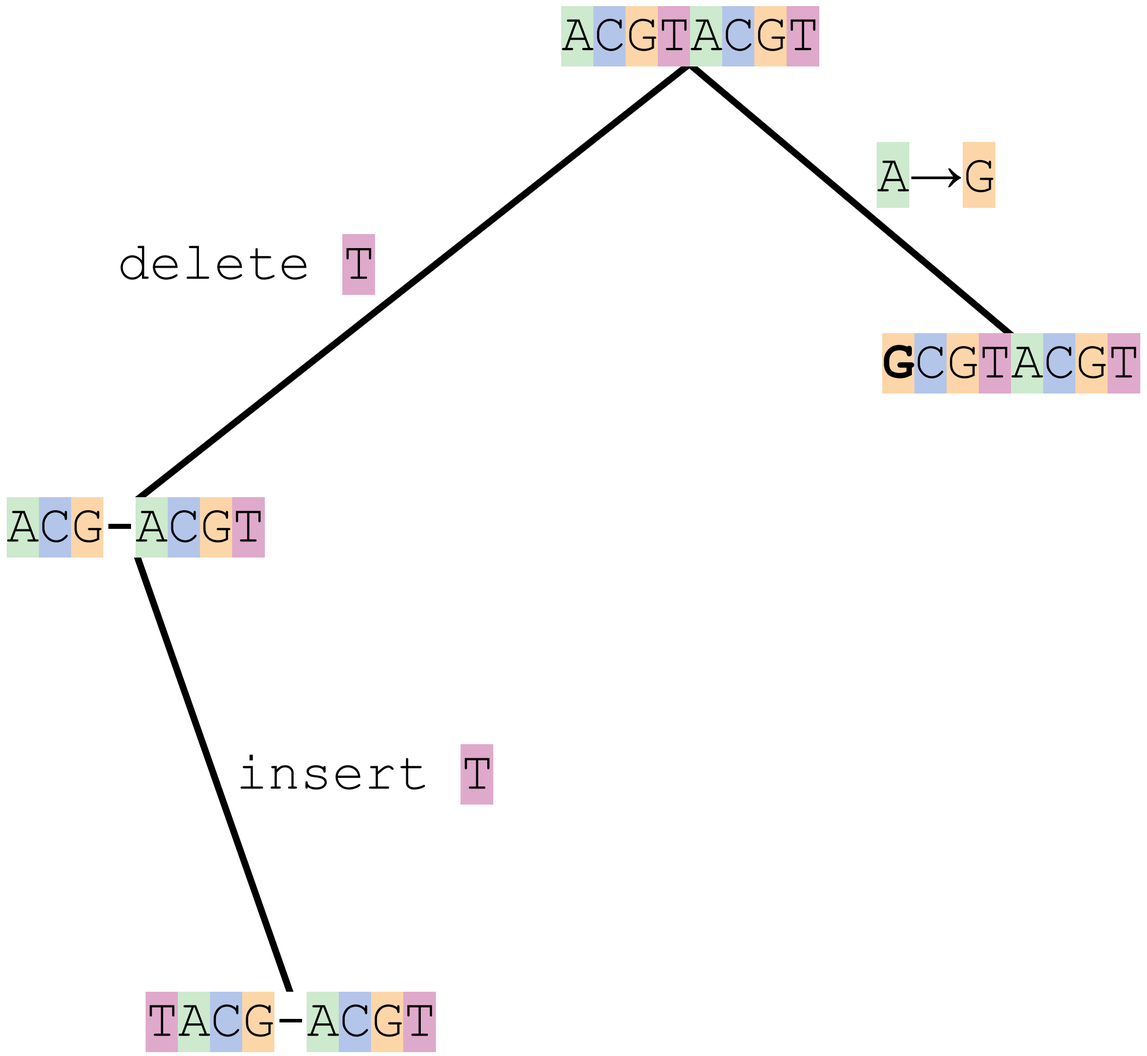
```
species 1    AGGATCTGCAATTGCTCTTCTAATCTGTCTGATCAGGAT
             ValArgSerCysSerCysValArgSerCysValValSer
species 2    AGG------AATTGCTCTTCTAATCTGTCT---CAGGAT
             Val------CysSerCysValArgSerCys---ValSer
species 3    AGGATCTGCAATTGC---TCTAATCTGTCTGATCAGGAT
             ValArgSerCysSer---ValArgSerCysValValSer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

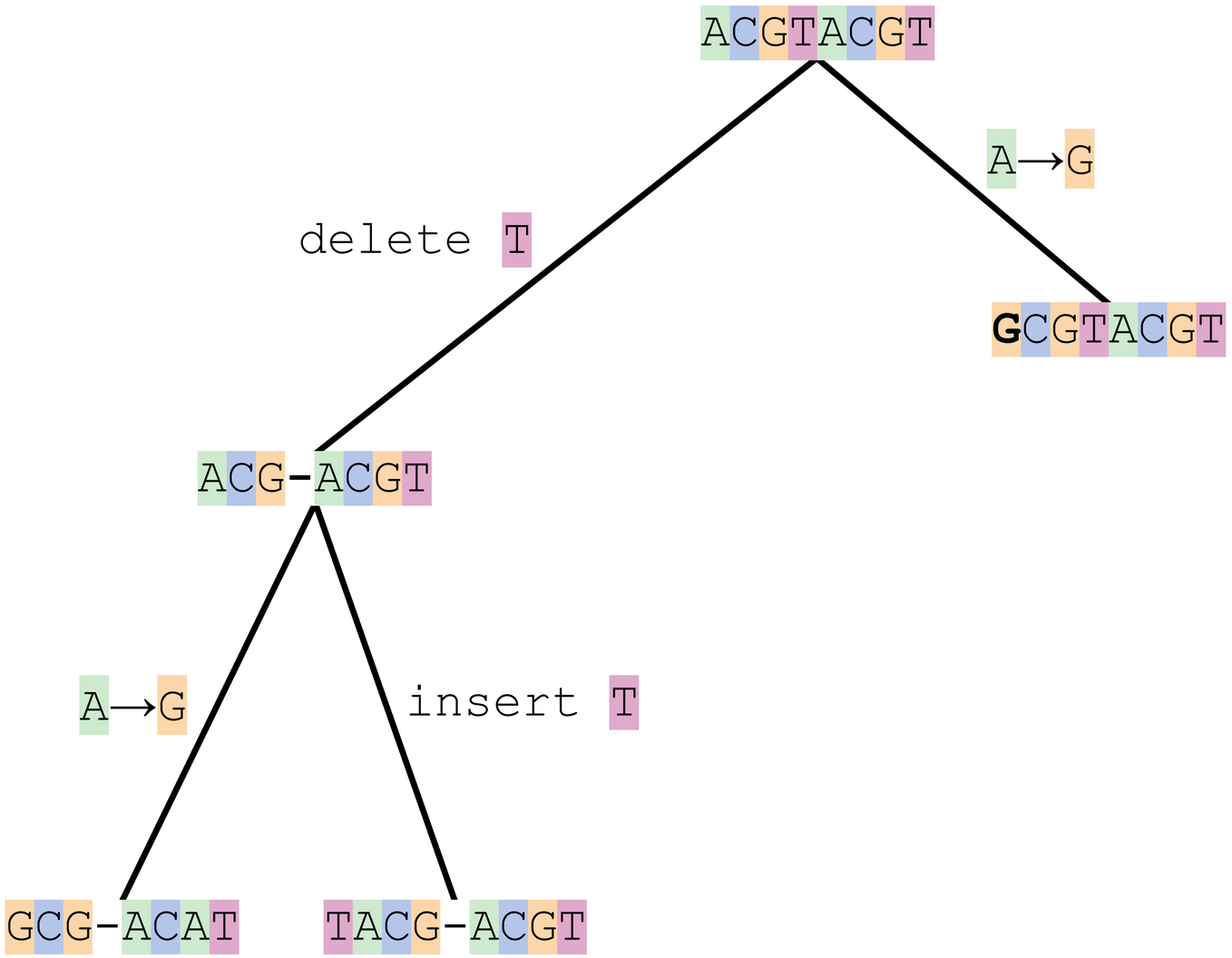
sequences evolve on a tree

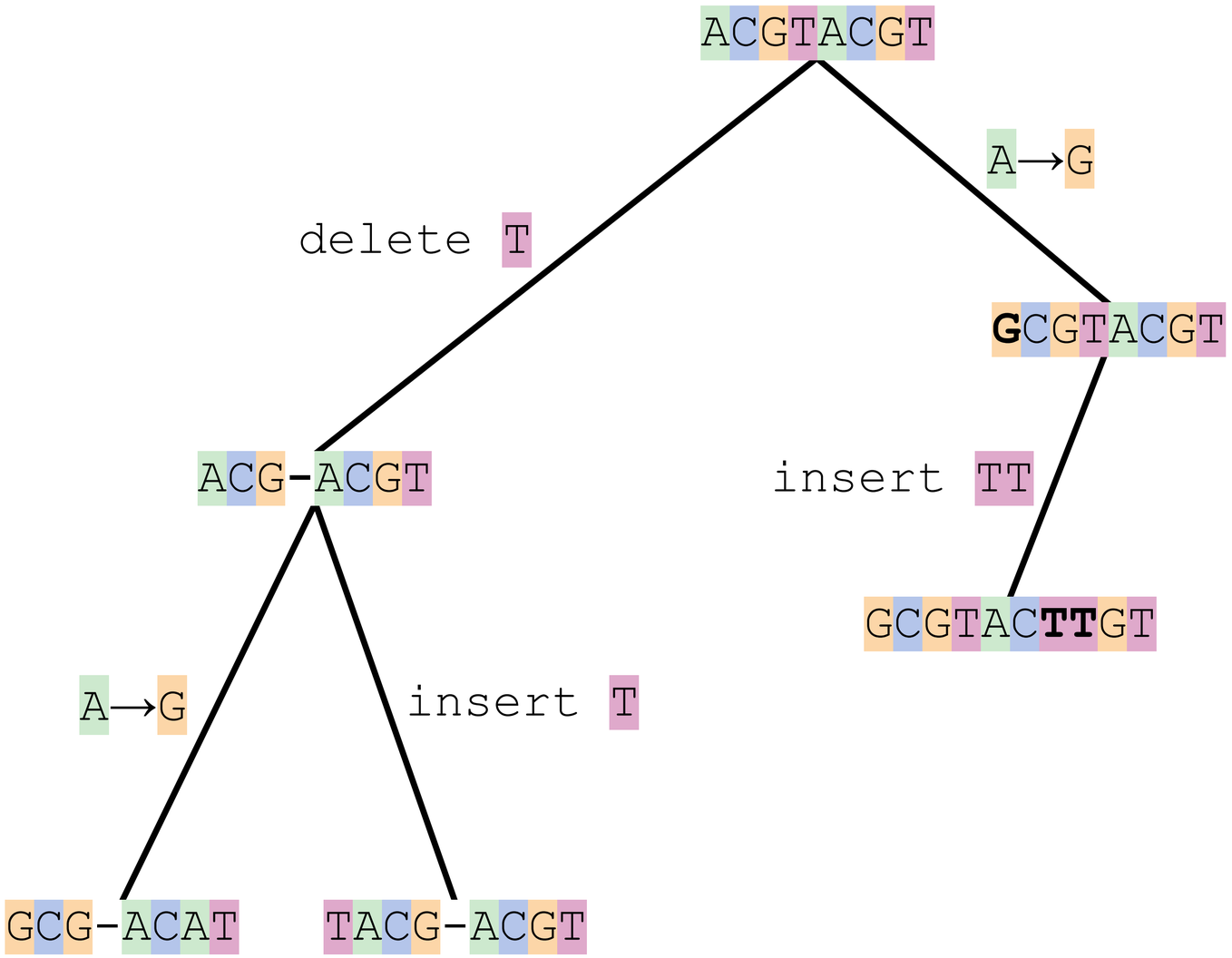

ACGTACGT

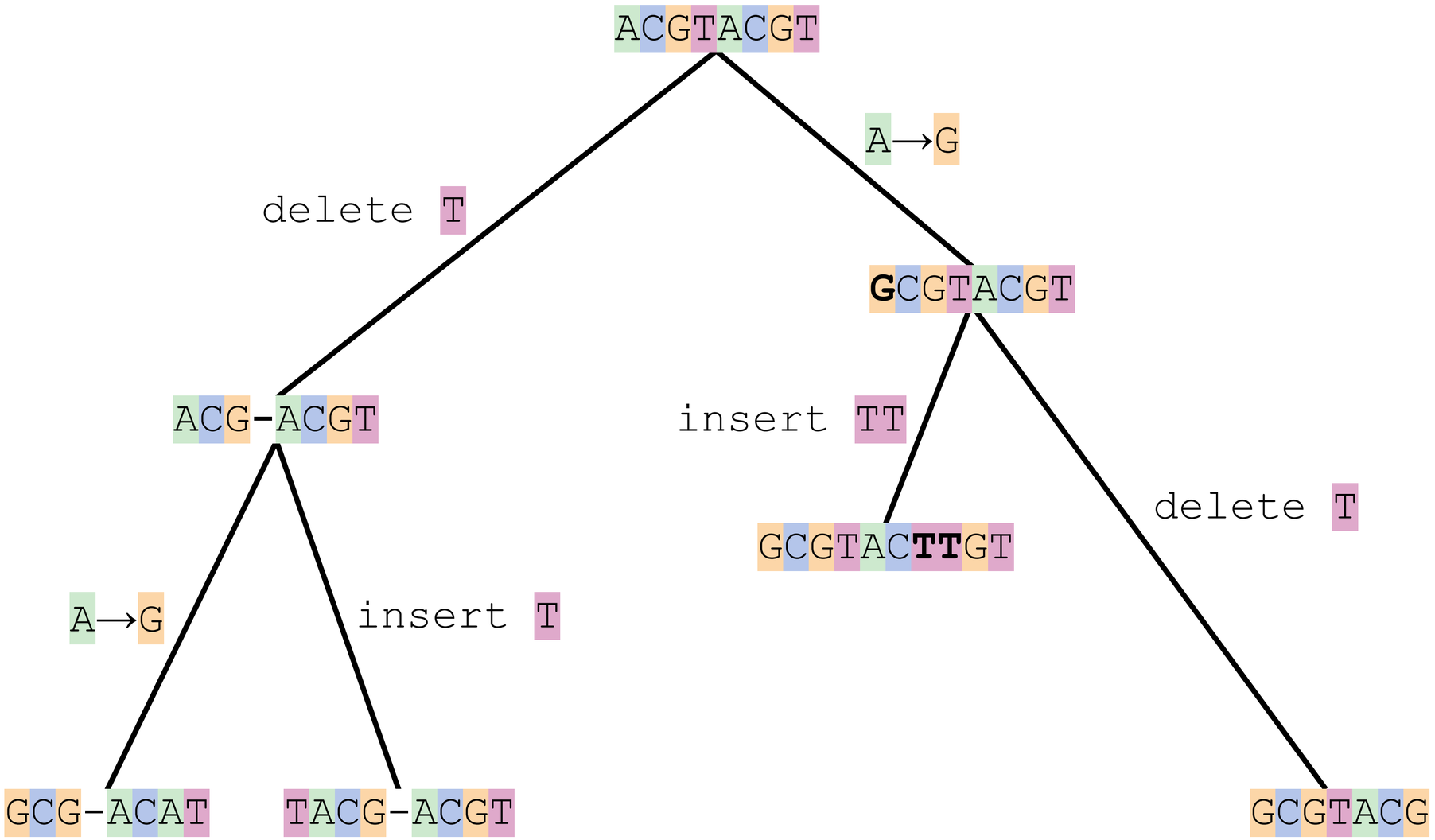

ACGTACGT
A⟶G
GCGTACGT

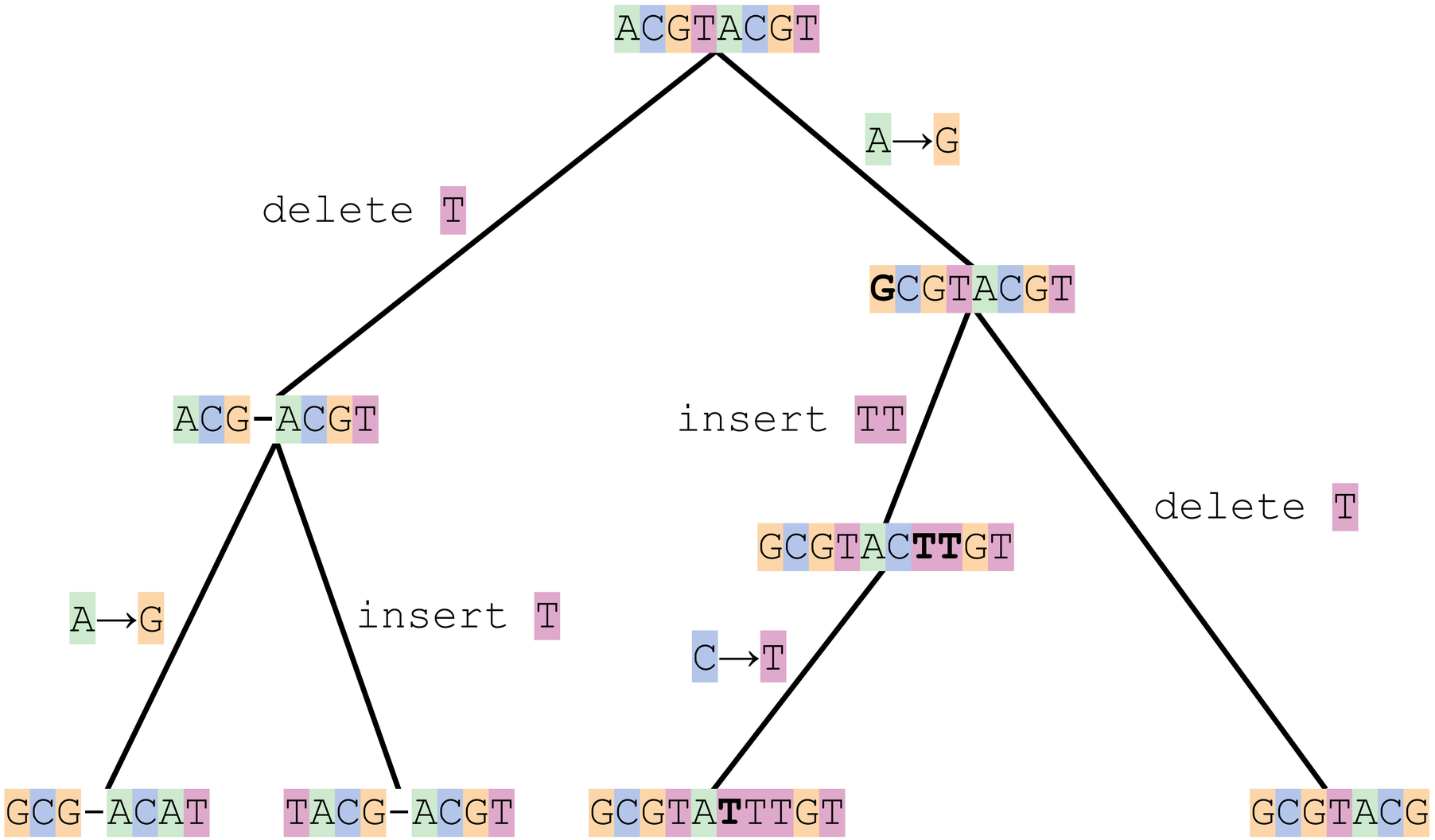

ACGTACGT
A→G
delete T
GCGTACGT
ACG-ACGT

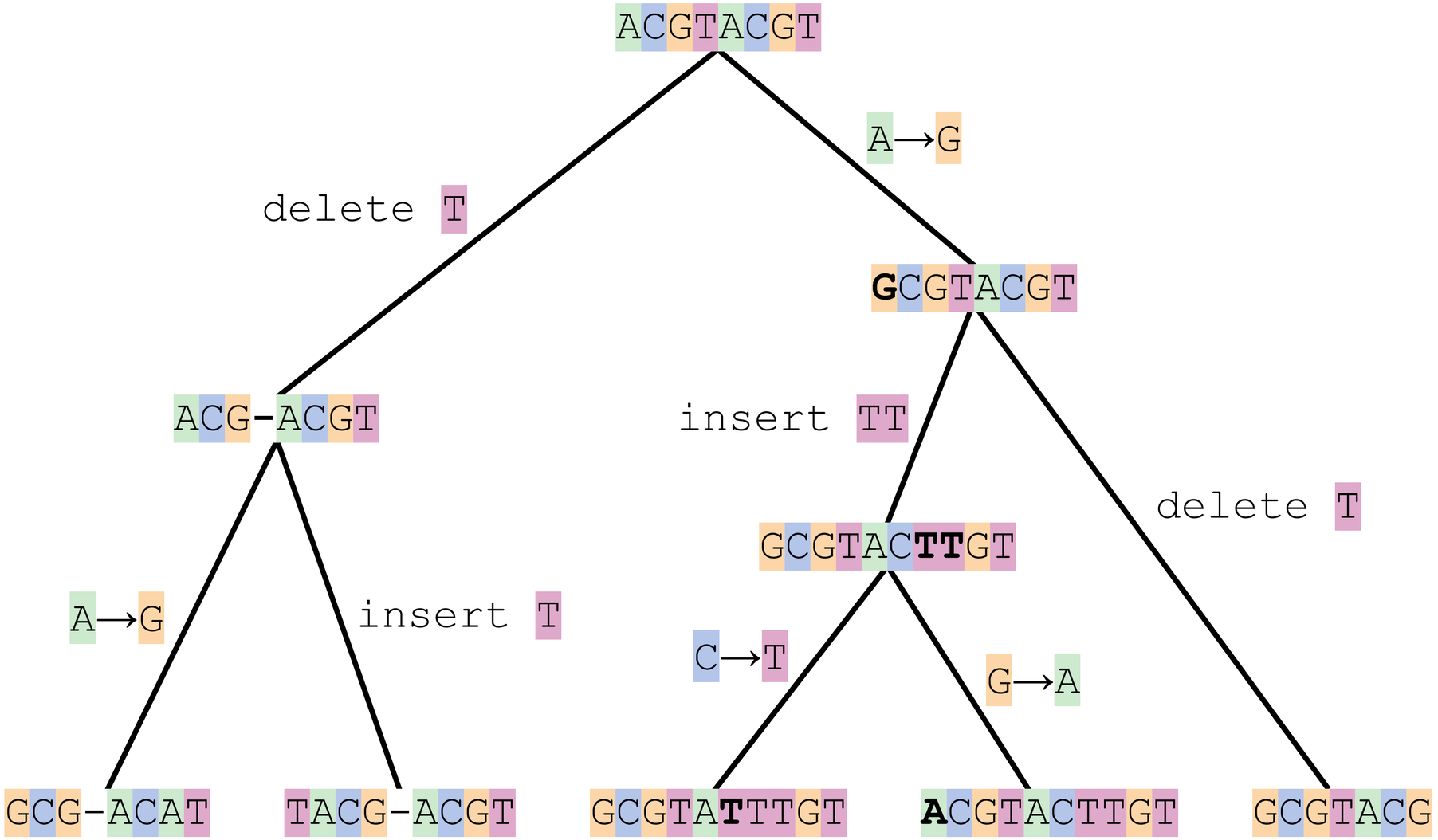

ACGTACGT
A→G
delete T
GCGTACGT
ACG-ACGT
insert T
TACG-ACGT


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

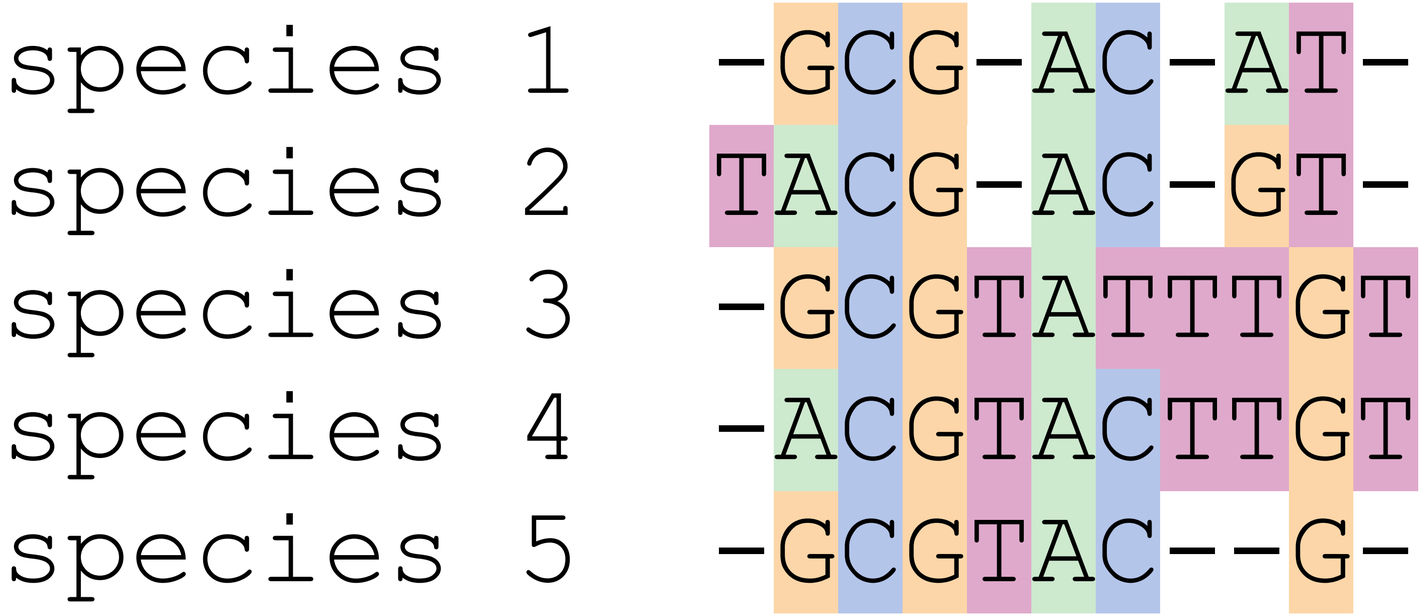

ACGTACGT
delete T
A→G
ACG-ACGT
GCGTACGT
insert TT
delete T
GCGTACTTGT
A→G
insert T
C→T
GCG-ACAT
TACG-ACGT
GCGTATTTGT
GCGTACG


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

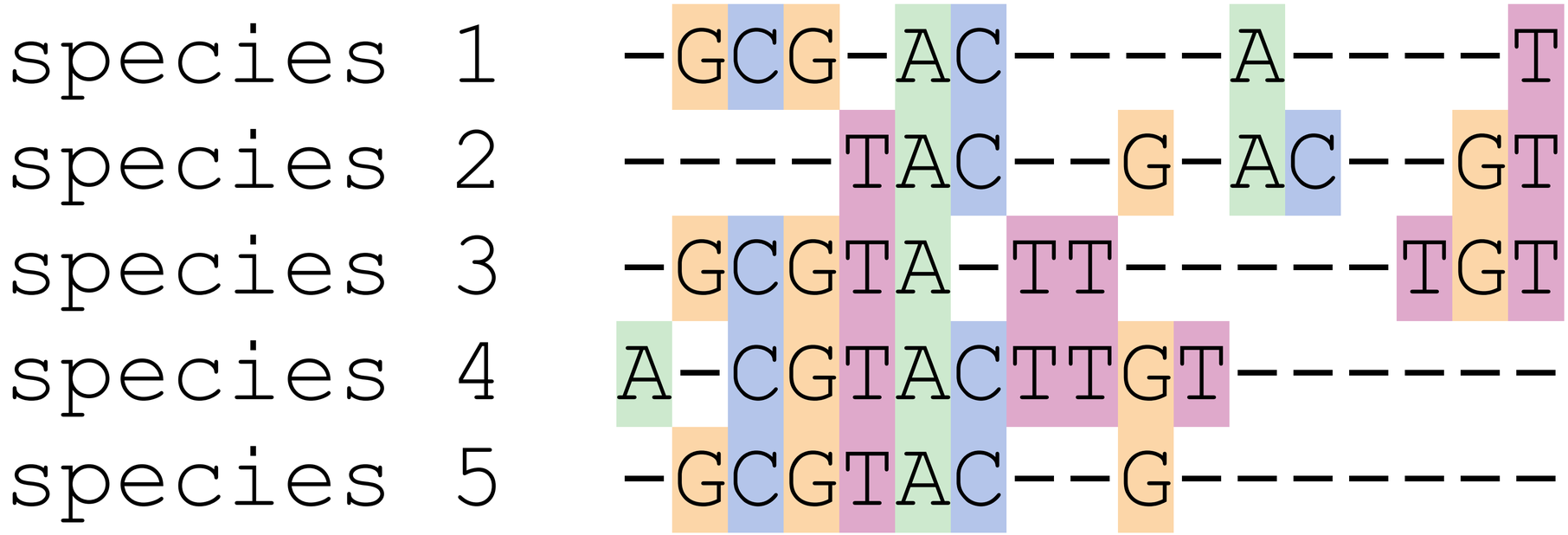
These are the sequences sampled at present!

```
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.



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



Which alignement is better?

This can be a bit of an art … 🎨🖌️

You need to balance the number of gaps with number of mismatches.

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

```
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-------
```

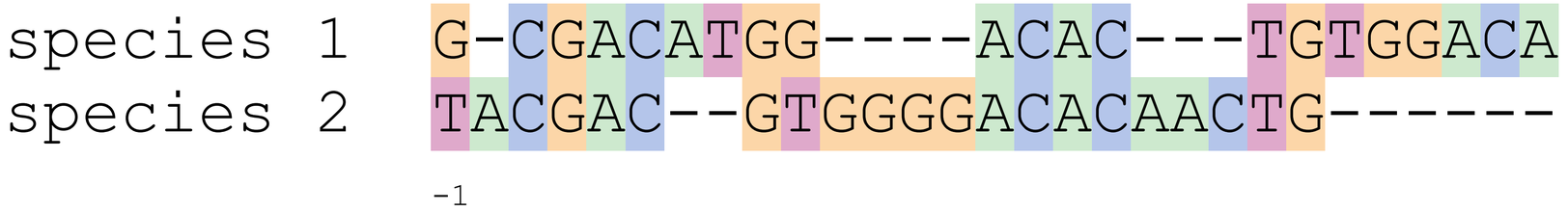
What price for a gap?

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
- e = 0.1

Nucleotide substitution scoring:
- match = +1
- mismatch = -1

Alignment scores calculation:
sum of match/mismatch scores - minus gap penalties.


species 1 G-CGACATGG----ACAC---TGTGGACA
species 2 TACGAC--GTGGGGACACAACTG------
-1


species 1 G-CGACATGG--ACAC-TGTGGACA
species 2 TACGACGTGGGGACACAACTG----
-1

```
species 1   G-CGACATGG----ACAC---TGTGGACA
species 2   TACGAC--GTGGGGACACAACTG------
```

-1-5+1+1+1+1

```
species 1   G-CGACATGG--ACAC-TGTGGACA
species 2   TACGACGTGGGGACACAACTG----
```

-1-5+1+1+1+1

```
species 1   G-CGACATGG----ACAC---TGTGGACA
species 2   TACGAC--GTGGGGACACAACTG------
```

-1-5+1+1+1+1-5

```
species 1   G-CGACATGG--ACAC-TGTGGACA
species 2   TACGACGTGGGGACACAACTG----
```

-1-5+1+1+1+1-1

```
species 1   G-CGACATGG----ACAC---TGTGGACA
species 2   TACGAC--GTGGGGACACAACTG------
```

-1-5+1+1+1+1-5-5

```
species 1   G-CGACATGG--ACAC-TGTGGACA
species 2   TACGACGTGGGGACACAACTG----
```

-1-5+1+1+1+1-1+1

```
species 1   G-CGACATGG----ACAC---TGTGGACA
species 2   TACGAC--GTGGGGACACAACTG------
            -1-5+1+1+1+1-5-5+1-1-5-5-5-5+1+1+1+1-5-5-5+1+1-5-5-5-5-5-5
```

```
species 1   G-CGACATGG--ACAC-TGTGGACA
species 2   TACGACGTGGGGACACAACTG----
            -1-5+1+1+1+1-1+1+1+1-5-5+1+1+1+1-5-1-1+1+1-5-5-5-5
```

```
species 1   G-CGACATGG----ACAC---TGTGGACA
species 2   TACGAC--GTGGGGACACAACTG------
            -1-5+1+1+1+1-5-5+1-1-5-5-5-5+1+1+1+1-5-5-5+1+1-5-5-5-5-5-5
```

final score = -71

```
species 1   G-CGACATGG--ACAC-TGTGGACA
species 2   TACGACGTGGGGACACAACTG----
            -1-5+1+1+1+1-1+1+1+1-5-5+1+1+1+1-5-1-1+1+1-5-5-5-5
```

final score = -31

There are two types of alignments:

Pairwise Alignment (just two species)

- **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.

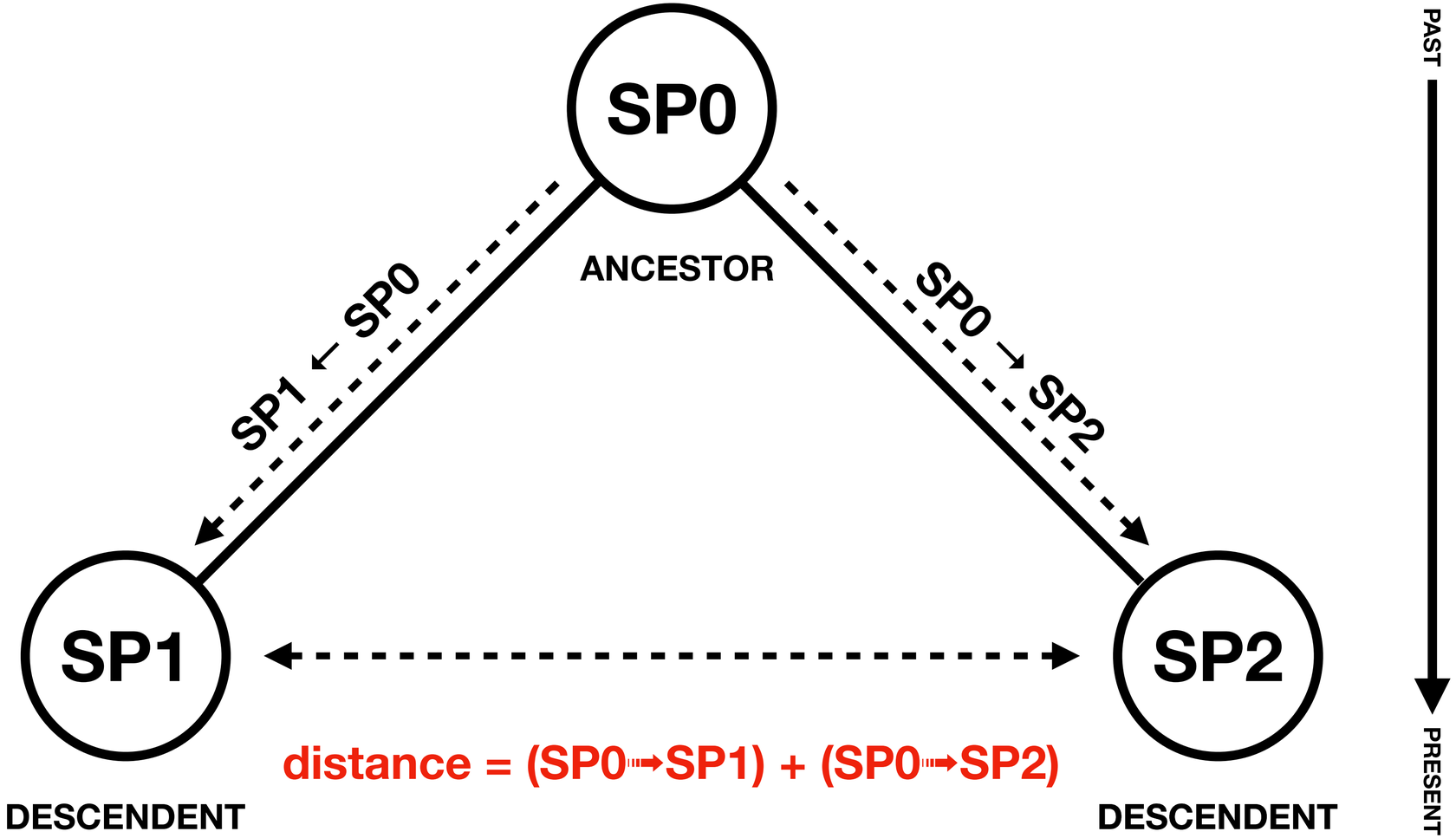
- **Local Alignment:** Smith-Waterman Algorithm

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.

Multiple Sequence Alignment (MSA)

1. Pairwise Alignments: Compute pairwise sequence distances.
2. Guide Tree Construction: Build a tree based on sequence similarity, using NJ or UPGMA.
3. Progressive Alignment: Align the closest sequences first, then aligning sequences to a profile.
4. Final Refinement: …

How to define the genetic distance between two taxa?



- **Hamming distance** measures the number of substitutions between two sequences: it only considers mismatches and ignores gaps.

- **Pairwise distance** is a more general measure of genetic distance that accounts for substitutions, insertions and deletions.

- **Distance using models** as Jukes-Cantor, Kimura-2-Parameter, Tamura-Nei, Hasegawa-Kishino-Yano, GTR … more on that next week 😛.

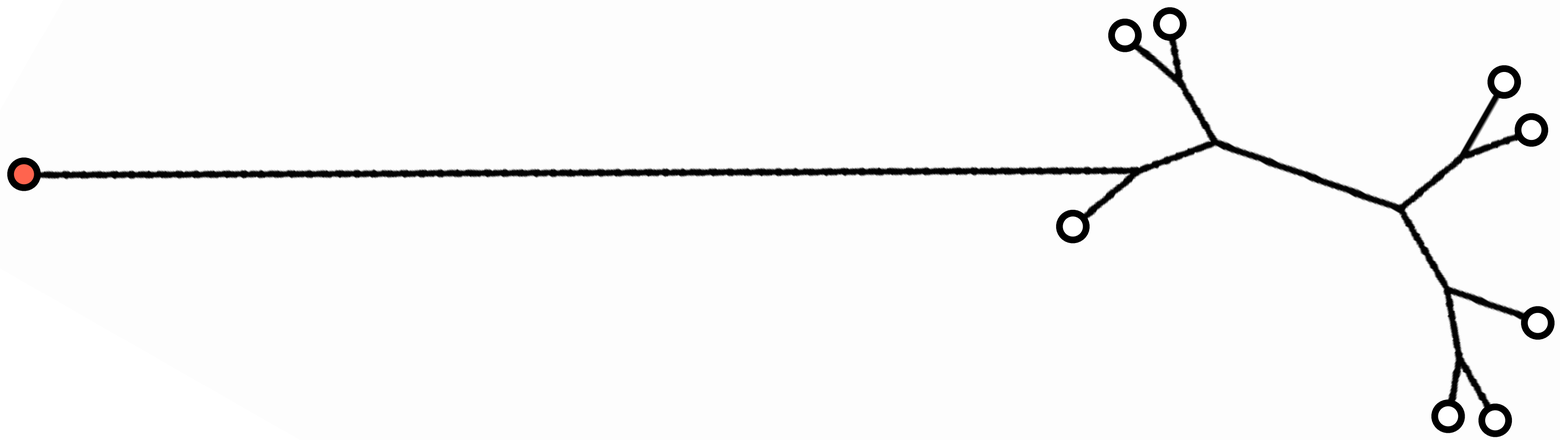
The goal of the alignment process is to identify evolutionary events associated with **homology** and ensure **aligned sequences accurately reflect evolutionary relationships of species**.

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

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

Notable tools for the process are:

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



Furthermore, **entire alignements** and **entire sequences** within alignements can be filtered out!

Several custom approaches are possible, but there is one which I like a lot: **exclude the too long terminal tips**, they might be misplaced genes or even contaminants 🐸🐝🐔🐌!

We will see more on phylogenetic subsampling (i.e. the markers choice) in the lesson on **biases**!

FINISH