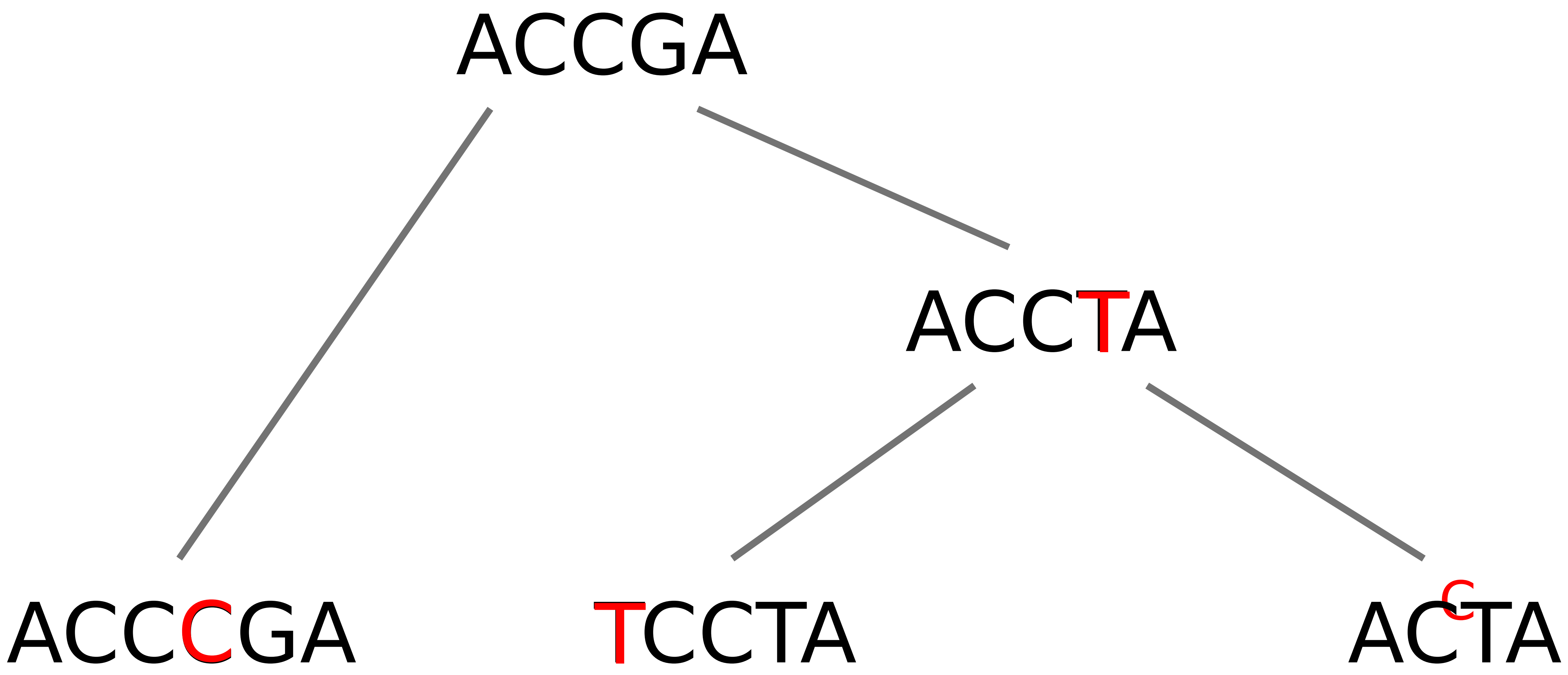
Sequence alignment, overview

Bio 208FS, Fall 2020

Paul Magwene

Over evolutionary time, sequences diverge due to mutation, drift, and selection



Example from Wills, 2011

Sequence alignment is a necessary first step for many subsequent tasks in bioinformatics

Observed sequences

```
ACCCGA
TCCTA
ACTA
```

A Possible Alignment of those Sequences

```
ACCCGA
TCC-TA
AC--TA
```

Some definitions

- Let Z be an alphabet -- the set of symbols (characters) from which a sequence can be composed (e.g. DNA nucleotides, amino acids)
- A sequence is a linear ordering of symbols from the alphabet Z
- The length of a sequence, *a*, is denoted |*a*|, and the symbol at position i in the sequence is designated a_i
- Given two sequences, *a* and *b*, the pair of sequences *a*' and *b*' is an **alignment** of *a* and *b* if:
  - The alphabet of the alignment, Z' is {Z} U {-} (the alphabet of Z plus the gap symbol '-')
  - |*a*'| = |*b*'|
  - Deleting all gap symbols of *a*' yields *a*, and deleting all gap symbols from *b*' yields *b*

Examples

a = TCCTA
b = ACTA

Some of the possible alignments between *a* and *b*

```
a'= TCCTA      a'= TCCTA      a'= TCCTA     a'= TCCTA
b'= -ACTA      b'= A-CTA      b'= AC-TA     b'= ACT-A
```

Cost of alignment and alignment distance

Let the **cost of the alignment** for the alignment (a', b') be defined as:

$$W(a', b') = \sum_{i=1}^{|a'|} w(a'_i, b'_i)$$

The **alignment distance** between a and b is:

$$D_w(a, b) = \min\{W(a', b') \,|\, (a', b') \text{ is an alignment of a and b}\}$$

Write a cost of alignment function

Where

$$w(a'_i, b'_i) = \begin{cases} 0 & \text{if } a'_i = b'_i \\ 1 & \text{otherwise} \end{cases}$$

Using your function, calculate the cost of each of the following alignments?

```
a'= TCCTA     a'= TCCTA     a'= TCCTA     a'= TCCTA    a'= TCCTA
b'= -ACTA     b'= A-CTA     b'= AC-TA     b'= ACT-A    b'= ACTA-
```

How do we find optimal alignments?

- Optimal alignments are ones that minimize the cost of alignment function
- For any pair of sequences, there are many possible alignments. How do we find the best ones?

Needleman-Wunsch algorithm

- See explanation in class