4. Sequence comparison

- How to predict gene/protein functions?
- Why gene/protein sequences may be similar?
- Measuring sequence similarity
- **Aligning sequences is an optimization problem**
- A sequence alignment as a path
- A path is optimal if all its sub-paths are optimal
- Alignment costs
- A recursive algorithm
- Recursion can be avoided: an iterative version
- How efficient is this algorithm?

Substitutions, but also insertions/deletions

ACCTCTAATCTATTCGTACTGCTATT

ACCTCTGAATCCATTCGTCTGCTATT

10 differences

Substitutions, but also insertions/deletions

```
ACCTCTAATCTATTCGTACTGCTATT
ACCTCTGAATCCATTCGTCTGCTATT
```

10 differences

```
ACCTCT-AATCTATTCGTACTGCTATT
ACCTCTGAATCCATTCGT-CTGCTATT
```

2 insertions/
deletions
1 substitution

Sequences may have different lengths

```
ACCTCTAATCTATTCGTACTGCTATT
    TGAATCCATTCGTCT

ACCTCT-AATCTATTCGTACTGCTATT
-----TGAATCCATTCGT-CT------
```

Sequence alignment

- To align the sequences, blank characters "**–**" can be inserted
- Several aligments of the same pair or sequences are possible
- Compute a score for each alignment
- Select the aligment for which the score is optimal

Sequence alignment for comparison
is an optimization problem