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?

Hamming distance

ACCTCTGTATCTATTCGGCATCATCAT
ACCCCTGAATCTATTCGGGATCATCAT

2 differences

ACCTCTGTATCTATTCGGGATCATCAT
ACCTCTGAATCTATCCGGGATCATGAT

3 differences

Hamming distance

- $D(S1, S1) = 0$
- $D(S1, S2) = D(S2, S1)$
- $D(S1, S2) + D(S2, S3) \geq D(S1, S3)$

- It is a mathematical distance

Computing the Hamming distance

```
function HammingDistance (Sequence1, Sequence2 : character string [1,*], Length: integer)
          return integer
    I, Distance: integer
    Distance ← 0
    for i from 1 to Length do
        if Sequence1[i] ≠ Sequence2[i] then
            Distance ← Distance + 1
    endfor
    return Distance
end HammingDistance
```