4. Sequence comparison

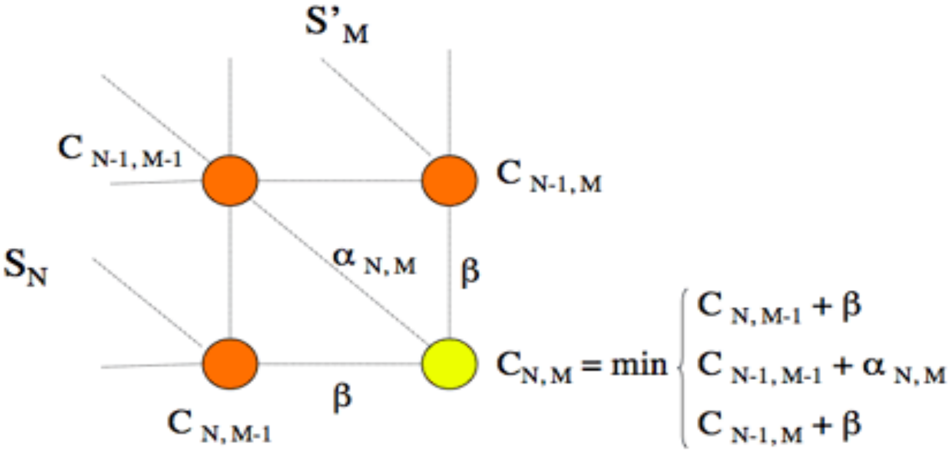
- How to predict gene/protein functions?
- Why gene/protein sequences may be similar?
- Measuring sequence similarity
- Aligning sequences is an optimization problem
- A sequence alignment as a path
- A path is optimal if all its sub-paths are optimal
- **Alignment costs**
- A recursive algorithm
- Recursion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

Computation of the cost on the last node

- β: cost of a gap "—" insertion

- $\alpha_{N,M}$: substitution cost of S1[N] by S2 [M]

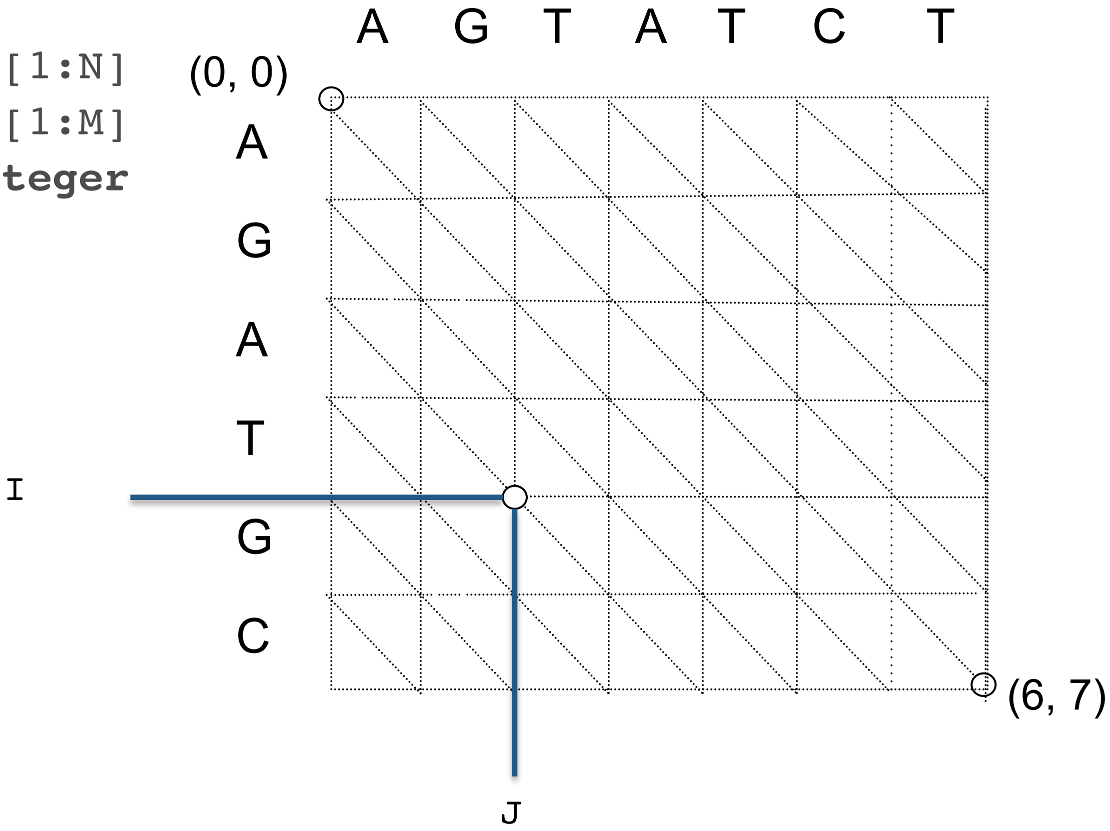
  - $\alpha_{N,M}$ is an element of the substitution matrix SubstitutionCost [1:4, 1:4] (DNA sequences) or [1:20, 1:20] (protein sequences)



$$C_{N,M} = \min \begin{cases} C_{N,M-1} + \beta \\ C_{N-1,M-1} + \alpha_{N,M} \\ C_{N-1,M} + \beta \end{cases}$$

Sequences and costs

```
Sequence1: character string [1:N]
Sequence2: character string [1:M]
Cost: array [0:N, 0:M] of integer
InsertionCost: integer
```



```
function SubstitutionCost
(Char1, Char2: character)
returns integer
```

Substitution cost function

- Accepts two characters Char1 and Char2 in the 4-letter DNA alphabet

  {A, C, G, T}

  or in the 20-letter protein alphabet

  {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, B, Z, X}

- Returns the cost of the substitution of Char1 by Char2

- Looks up a matrix of costs

- The optimal alignment

  gets the lowest cost

	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0

Substitution cost matrices

- The choice of a matrix relies on biological considerations
  - For instance, consider differentely transitions (e.g. A ↔ G)
    and transversions (e.g. G ↔ C)
- More critical for AA substitution matrices
  - Based on biophysical properties of amino acids