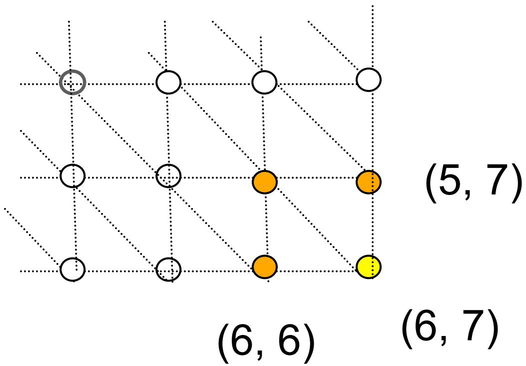
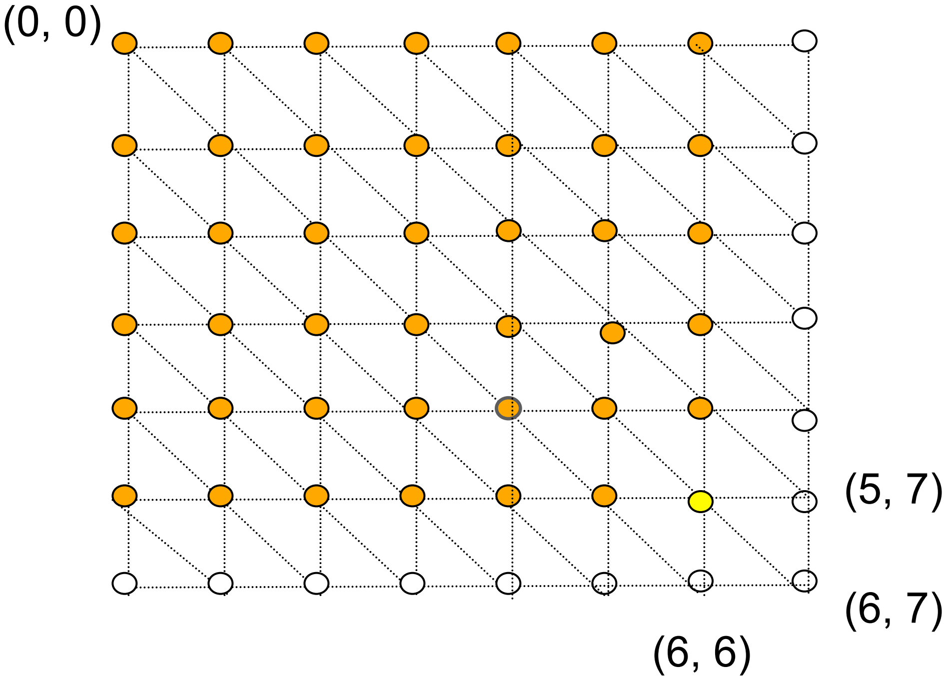
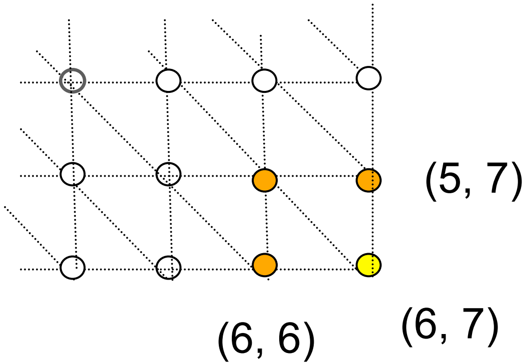
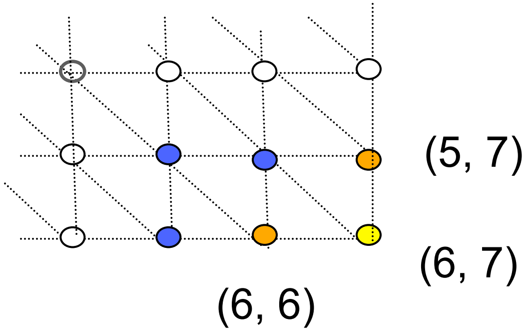
4. Sequence comparison

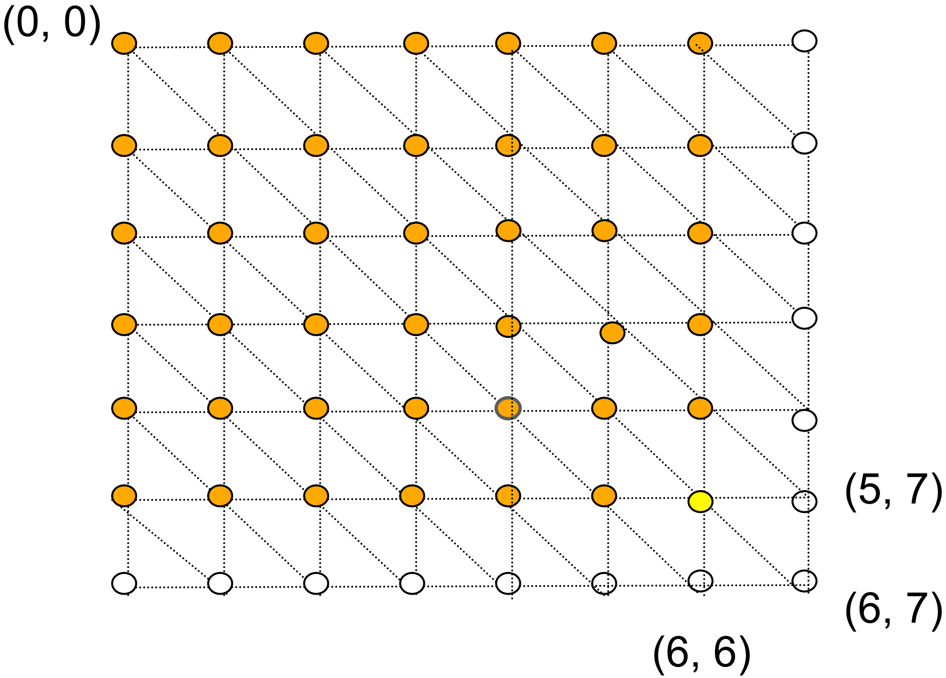
- How to predict gene/protein functions?
- Why gene/protein sequences may be similar?
- Measuring sequence similarity
- Aligning sequences is an optimization problem
- A sequence alignment as a path
- A path is optimal if all its sub-paths are optimal
- Alignment costs
- A recursive algorithm
- **Recursion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


(5, 7)
(6, 6)
(6, 7)


(0, 0)
(5, 7)
(6, 7)
(6, 6)

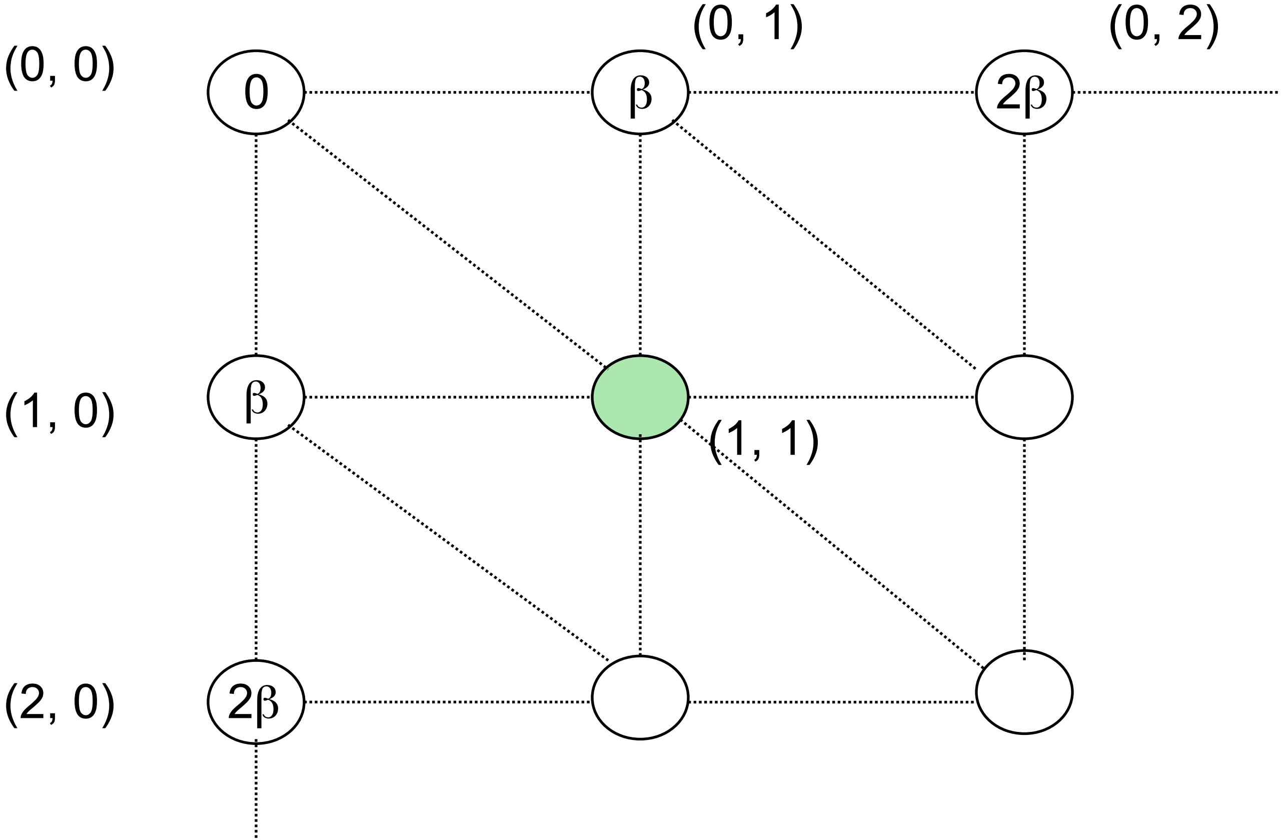

(5, 7)
(6, 6)
(6, 7)

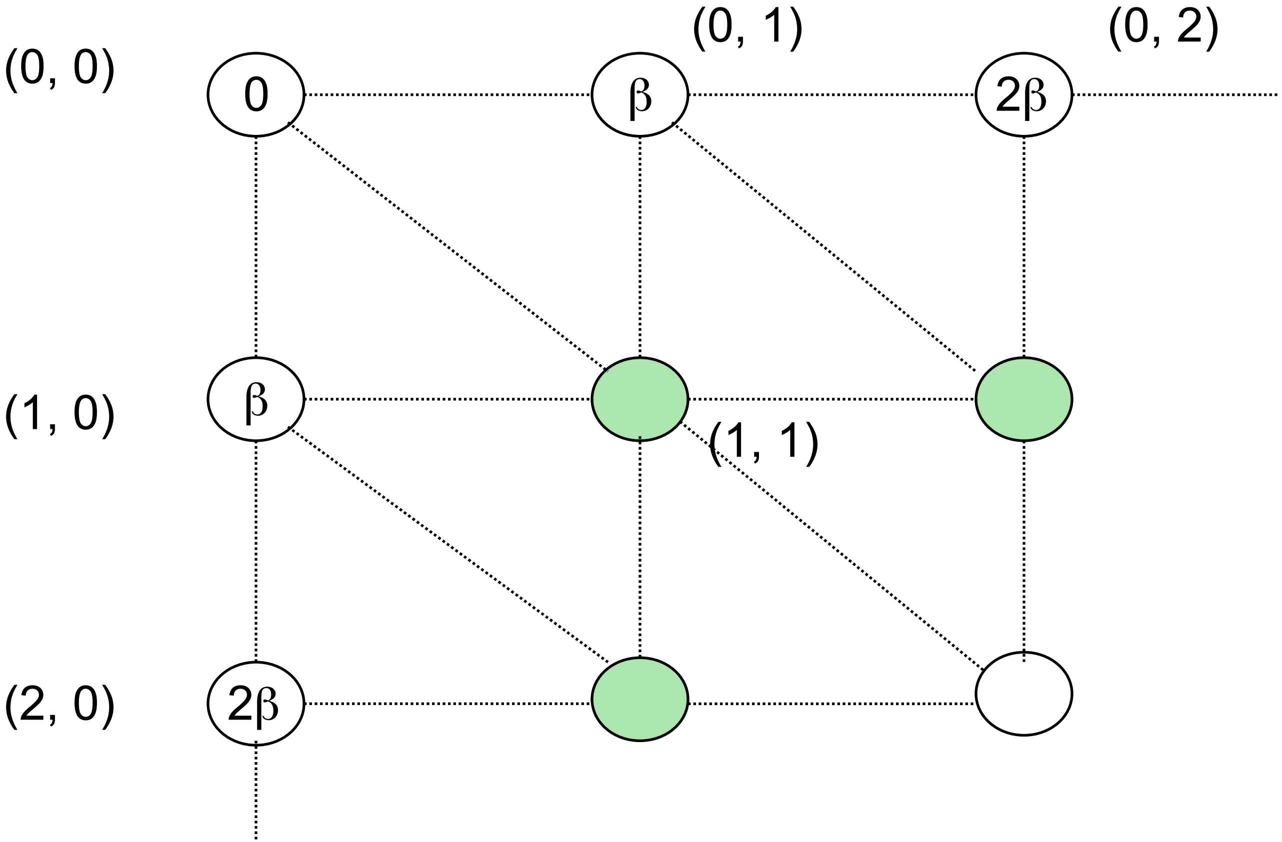

(5, 7)
(6, 7)
(6, 6)

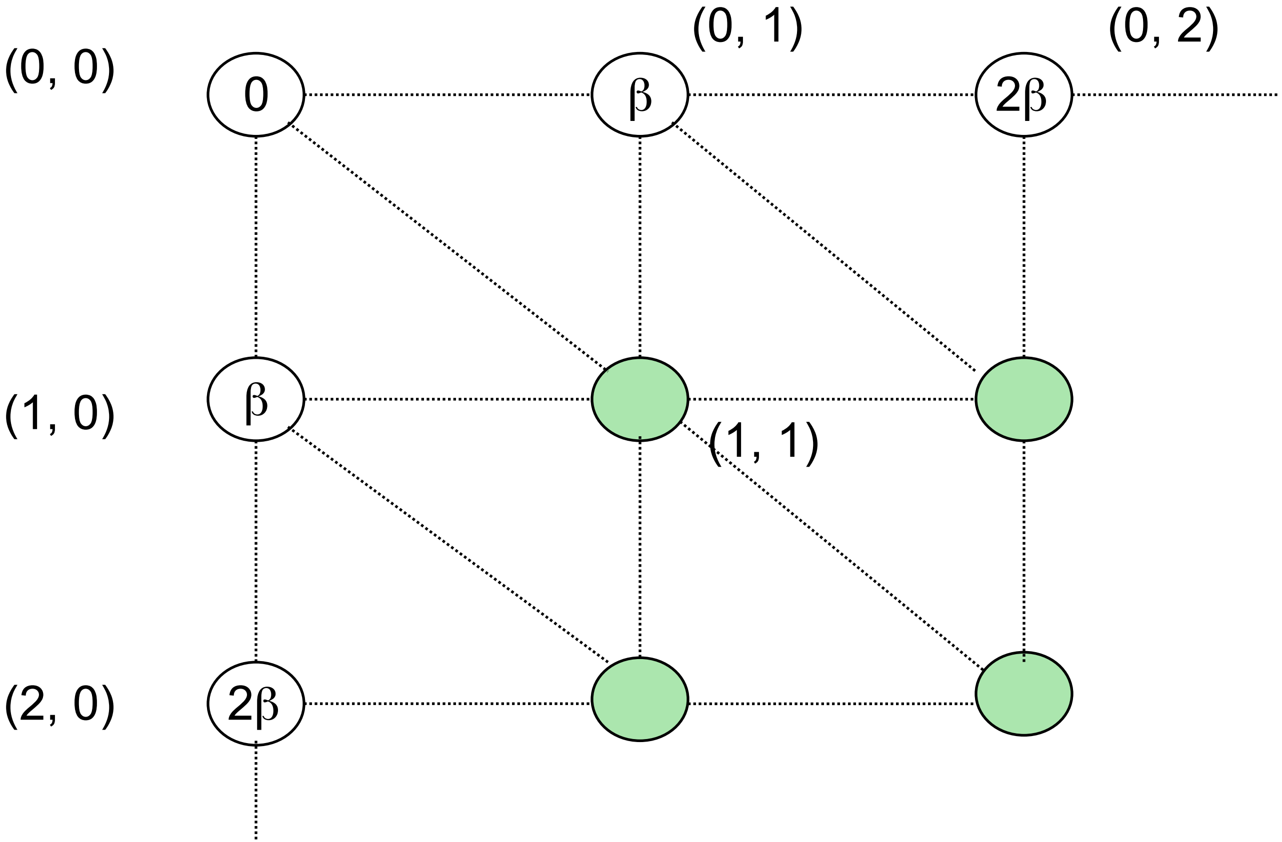

(0, 0)
(5, 7)
(6, 7)
(6, 6)

An iterative algorithm in two phases

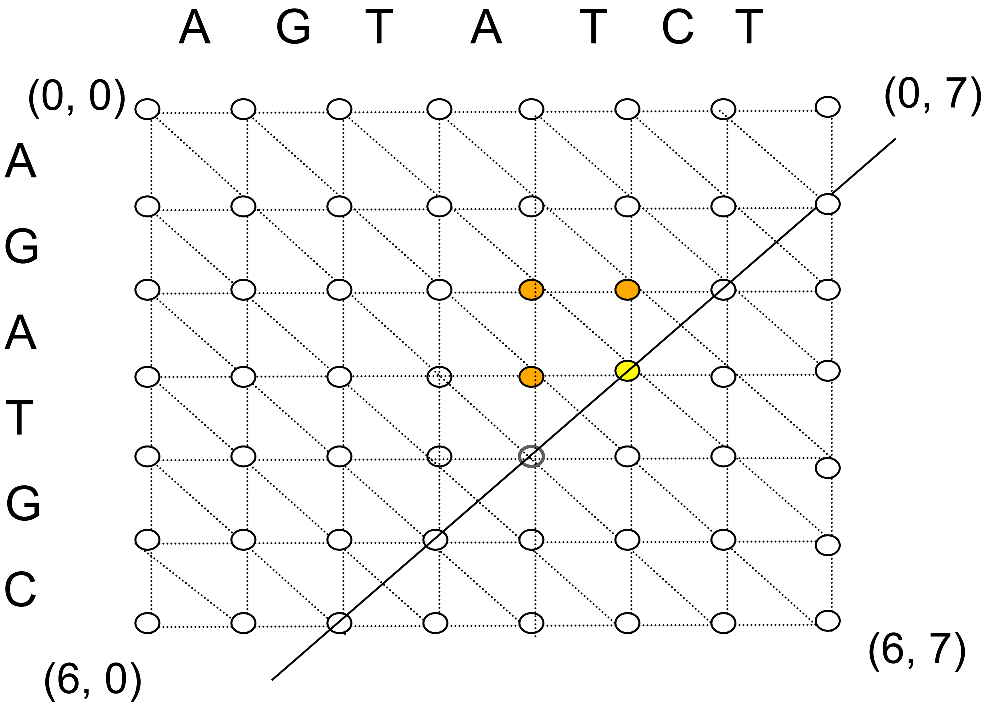
- First phase
  - Compute the cost of the optimal path ending on each node [I, J] starting from node [0, 0] until node [N, M]
  - and store them in a matrix


(0, 0)
0
(0, 1)
β
(0, 2)
2β
(1, 0)
β
(1, 1)
(2, 0)
2β


(0, 0)
0
(0, 1)
β
(0, 2)
2β
(1, 0)
β
(1, 1)
(2, 0)
2β


(0, 0)
(0, 1)
(0, 2)
(1, 0)
(1, 1)
(2, 0)
0
β
2β
β
2β

Parallelization is possible

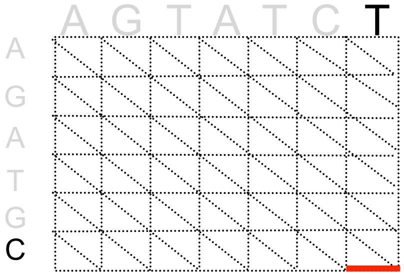


An iterative algorithm in two phases

- First phase
  - Compute the cost of the optimal path ending on each node [I, J] starting from node [0, 0] until node [N, M]
  - and store them in a matrix
- Second phase
  - Compute the optimal paths backwards, from node [N, M] to node [0, 0], using the precomputed costs in the matrix

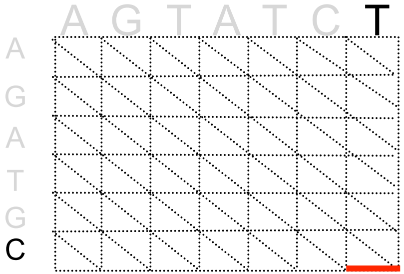
—
T

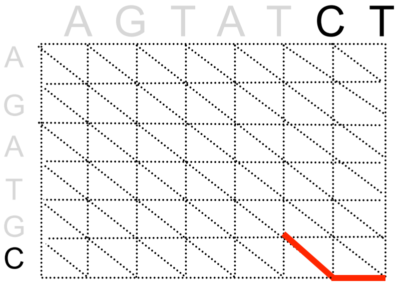
(a)



—
T

(a)



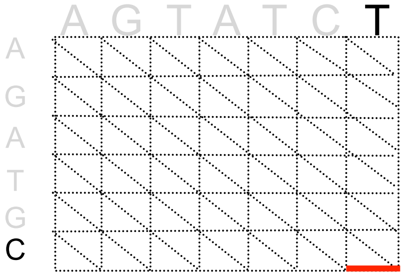


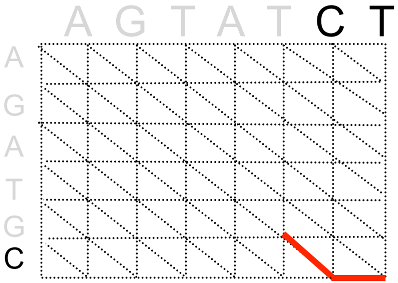
(b)

C —
C T

—
T

(a)





(b)

C —
C T

G C —
T C T

(c)