# String Matching

String matching: definition of the problem (text,pattern)

- Exact matching: depends on what we have: text or patterns
  - The patterns ---> Data structures for the patterns
    - 1 pattern ---> The algorithm depends on |p| and |Σ|
    - k patterns ---> The algorithm depends on k, |p| and |Σ|
  - Extensions
  - Regular Expressions
  - The text ----> Data structure for the text (suffix tree, ...)
- Approximate matching:
  - Dynamic programming
  - Sequence alignment (pairwise and multiple)
  - Sequence assembly: hash algorithm
- Probabilistic search: Hidden Markov Models

# Approximate string matching

algorithmics genetics group

For instance, given the sequence

CTACTACTACGTGACTAATACTGATCGTAGCTAC…

search for the pattern ACTGA allowing one error…

… but what is the meaning of “one error”?

# Edit distance

We accept three types of errors:

1. Mismatch: ACCGTGAT ACCGAGAT
2. Insertion: ACCGTGAT ACCGATGAT
3. Deletion: ACCGTGAT ACCGGAT

} Indel

The edit distance d between two strings is the minimum number of substitutions,insertions and deletions needed to transform the first string into the second one

d(ACT,ACT)= d(ACT,AC)= d(ACT,C)=
d(ACT,)= d(AC,ATC)= d(ACTTG,ATCTG)=

# Edit distance

We accept three types of errors:

1. Mismatch: ACCGTGAT ACCGAGAT
2. Insertion: ACCGTGAT ACCGATGAT
3. Deletion: ACCGTGAT ACCGGAT

} Indel

The edit distance d between two strings is the minimum number of substitutions,insertions and deletions needed to transform the first string into the second one

d(ACT,ACT)= 0 d(ACT,AC)= 1 d(ACT,C)= 2
d(ACT,)= 3 d(AC,ATC)= 1 d(ACTTG,ATCTG)= 2

# Edit distance and alignment of strings

The Edit distance is related with the best alignment of strings

Given

d(ACT,ACT)=0 d(ACT,AC)=1 d(ACTTG,ATCTG)=2

which is the best alignment in every case?

- ACT and ACT : ACT
  ACT
- ACT and AT: ACT
  A -T
- ACTTG and ATCTG: ACTTG ACT - TG
  ATCTG A - TCTG

Then, the alignment suggest the substitutions, insertions and deletions to transform one string into the other

# Edit distance and alignment of strings

algorithmics genetics group .

But which is the distance between the strings

ACGCTATGCTATACG and ACGGTAGTGACGC?

… and the best alignment between them?

1966 was the first time this problem was discussed…

and the algorithm was proposed in 1968,1970,…

using the technique called “Dynamic programming”

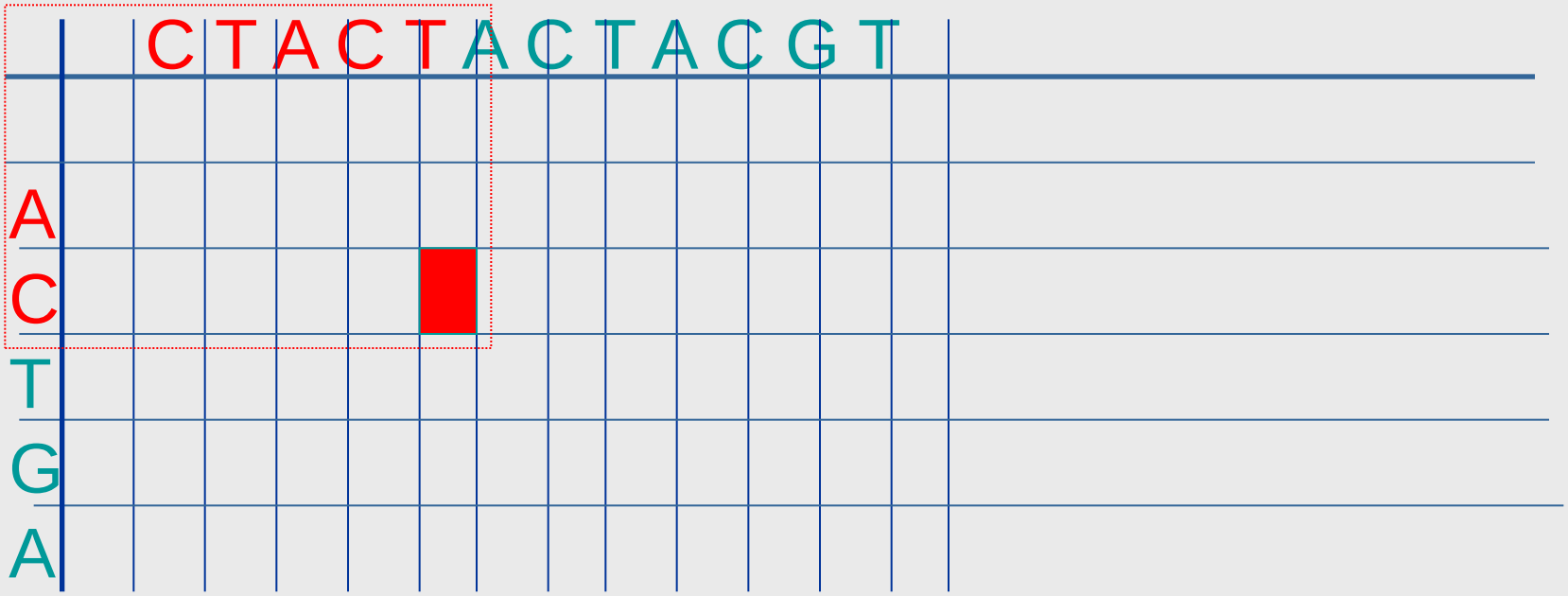
# Edit distance and alignment of strings

algorithmics genetics group .

# Edit distance and alignment of strings

algorithmics genetics group .

# Edit distance and alignment of strings




The cell contains the distance between AC and CTACT.

# Edit distance and alignment of strings

algorithmics genetics group .

# Edit distance and alignment of strings

algorithmics genetics group .

| | | C | T | A | C | T | A | C | T | A | C | G | T | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | ? | | | | | | | | | | | | |
| A | | | | | | | | | | | | | | |
| C | | | | | | | | | | | | | | |
| T | | | | | | | | | | | | | | |
| G | | | | | | | | | | | | | | |
| A | | | | | | | | | | | | | | |

# Edit distance and alignment of strings


-

C

# Edit distance and alignment of strings

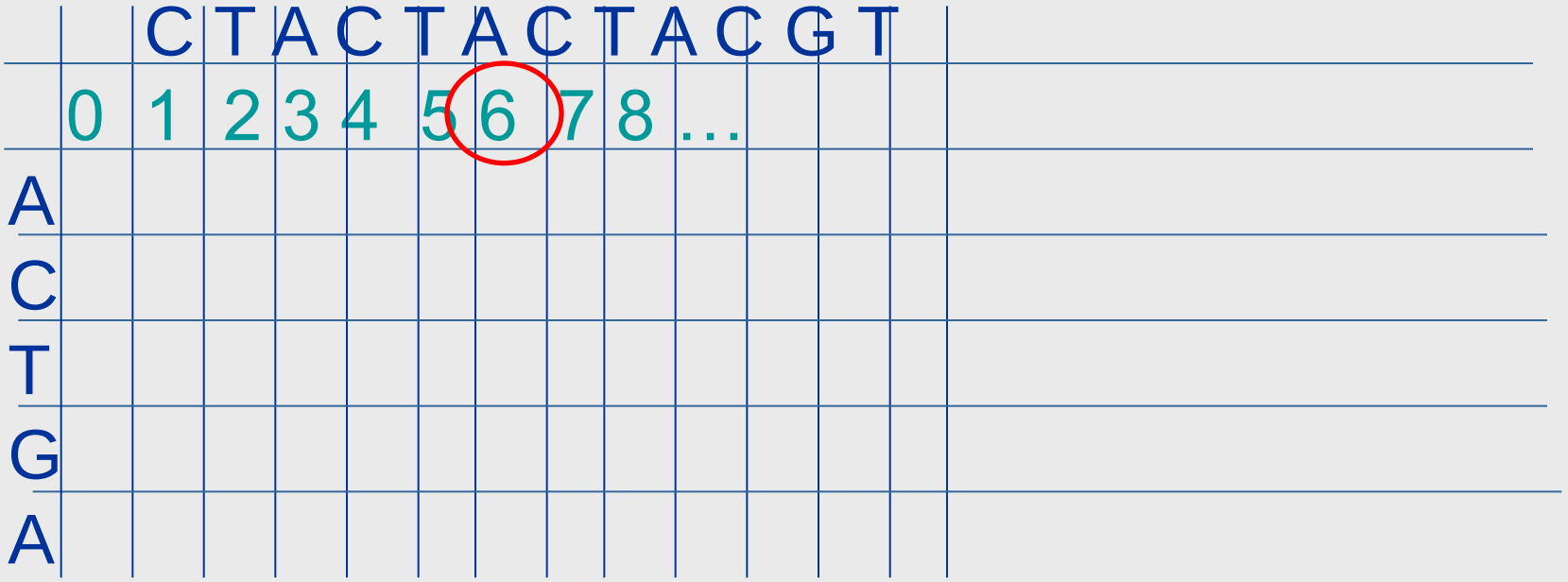
algorithmics genetics group .


- -

CT

# Edit distance and alignment of strings

algorithmics genetics group




- - - - - -

CTACTA

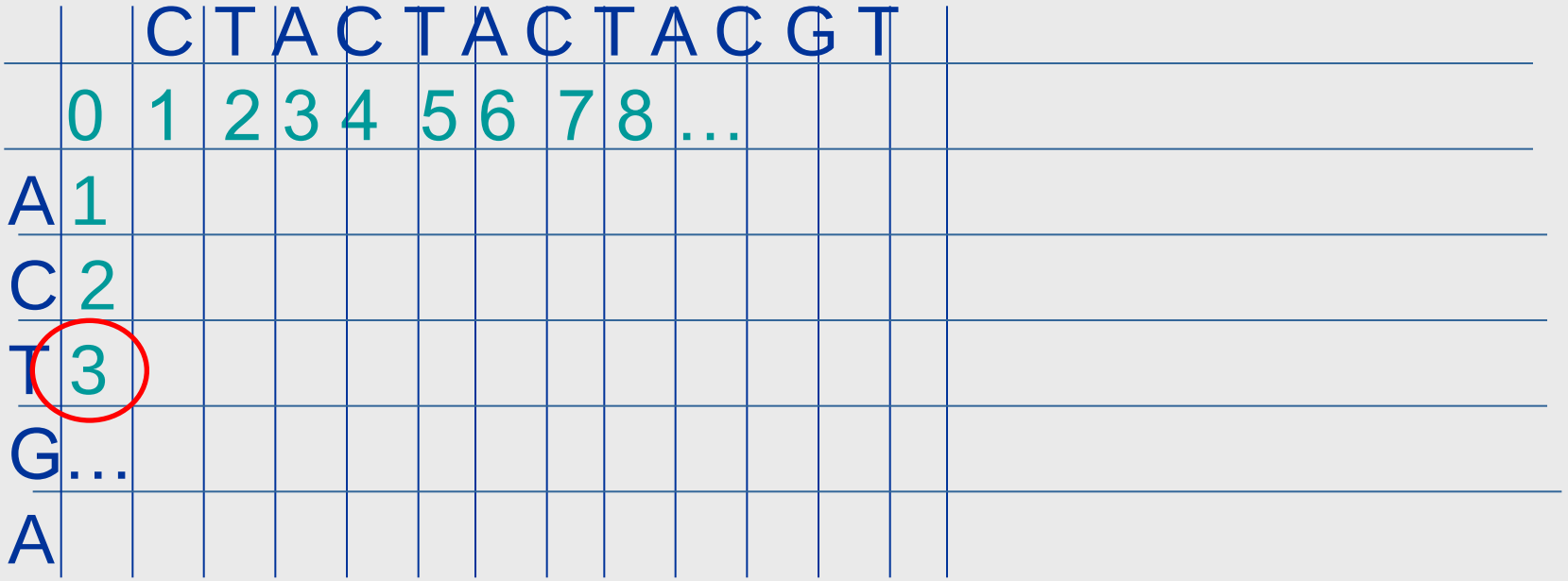
# Edit distance and alignment of strings

algorithmics genetics group

| | | C | T | A | C | T | A | C | T | A | C | G | T | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ... | | | | |
| A | ? | | | | | | | | | | | | | |
| C | ? | | | | | | | | | | | | | |
| T | ? | | | | | | | | | | | | | |
| G | | | | | | | | | | | | | | |
| A | | | | | | | | | | | | | | |

# Edit distance and alignment of strings

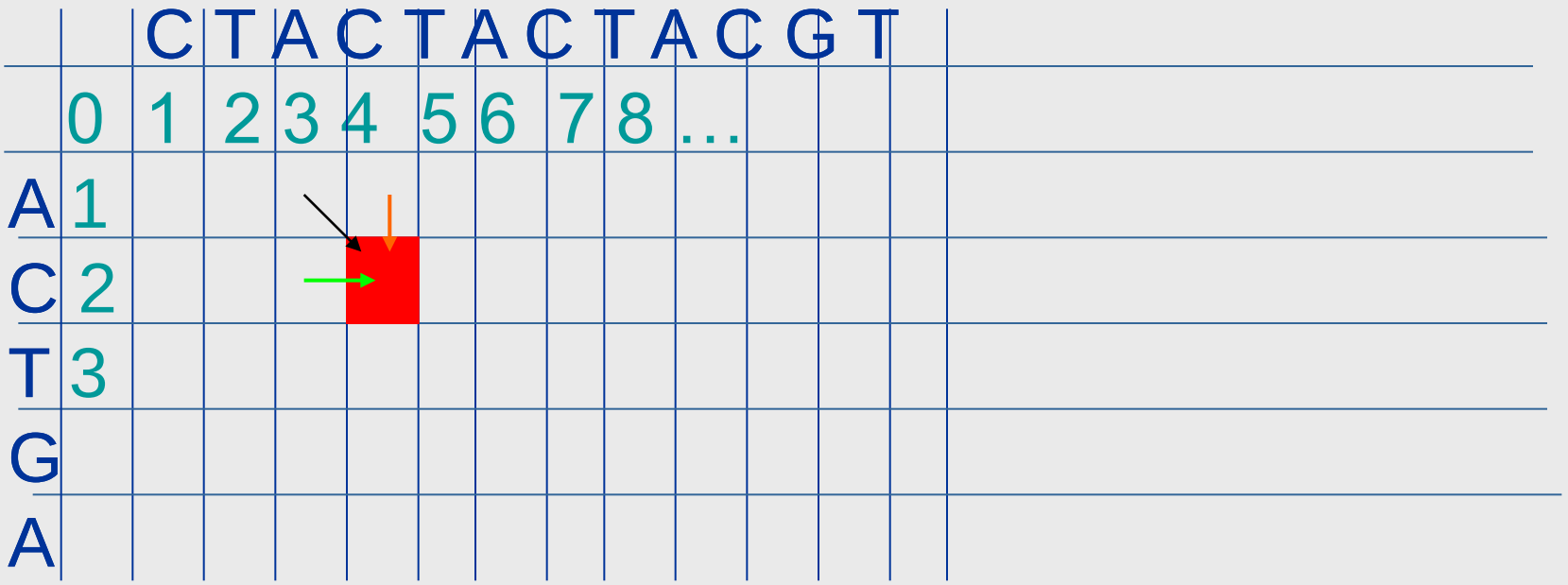
algorithmics genetics group .




ACT

- - -

# Edit distance and alignment of strings




BA(AC,CTAC)= best

d(AC,CTAC)=min

- BA(AC,CTA) + C → d(AC,CTA)+1
- BA(A,CTA) + C C → d(A,CTA)
- BA(A,CTAC) + C - → d(A,CTAC)+1

# Edit distance and alignment of strings

algorithmics genetics group .

Connect to

http://alggen.lsi.upc.es/docencia/ember/leed/Tfc1.htm

and use the global method.

# Edit distance and alignment of strings

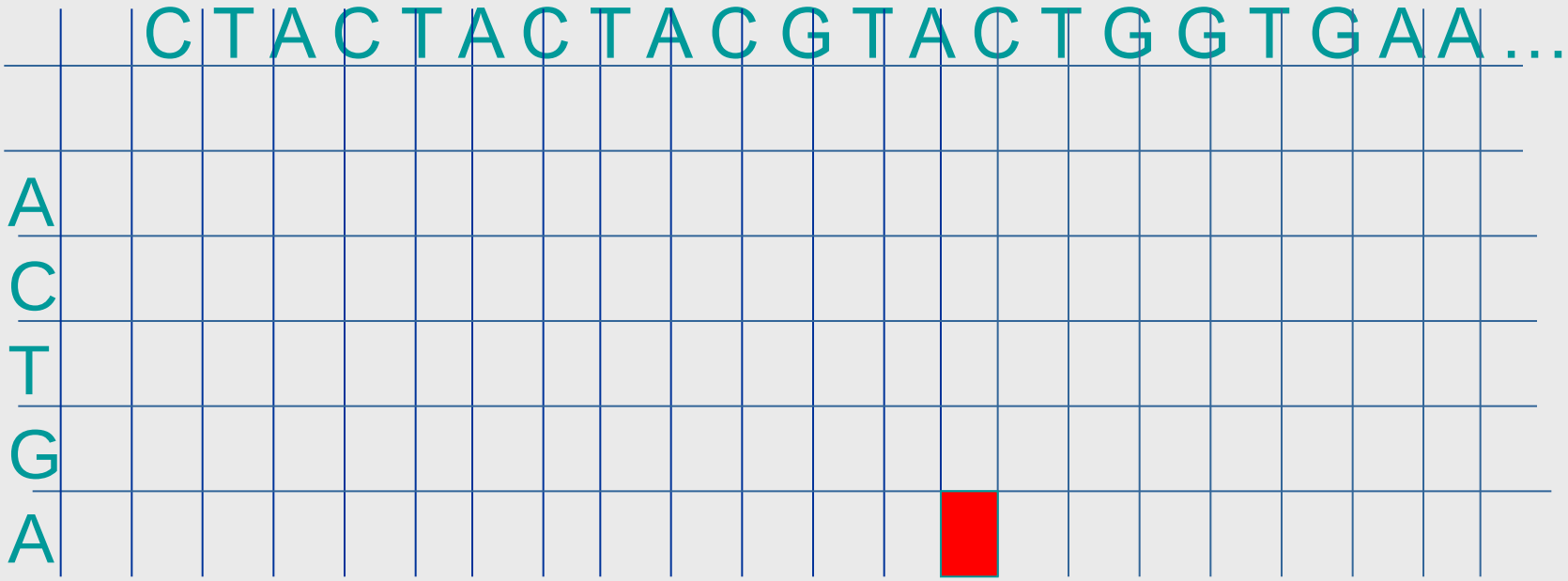
algorithmics genetics group .

How this algorithm can be applied

to the approximate search?

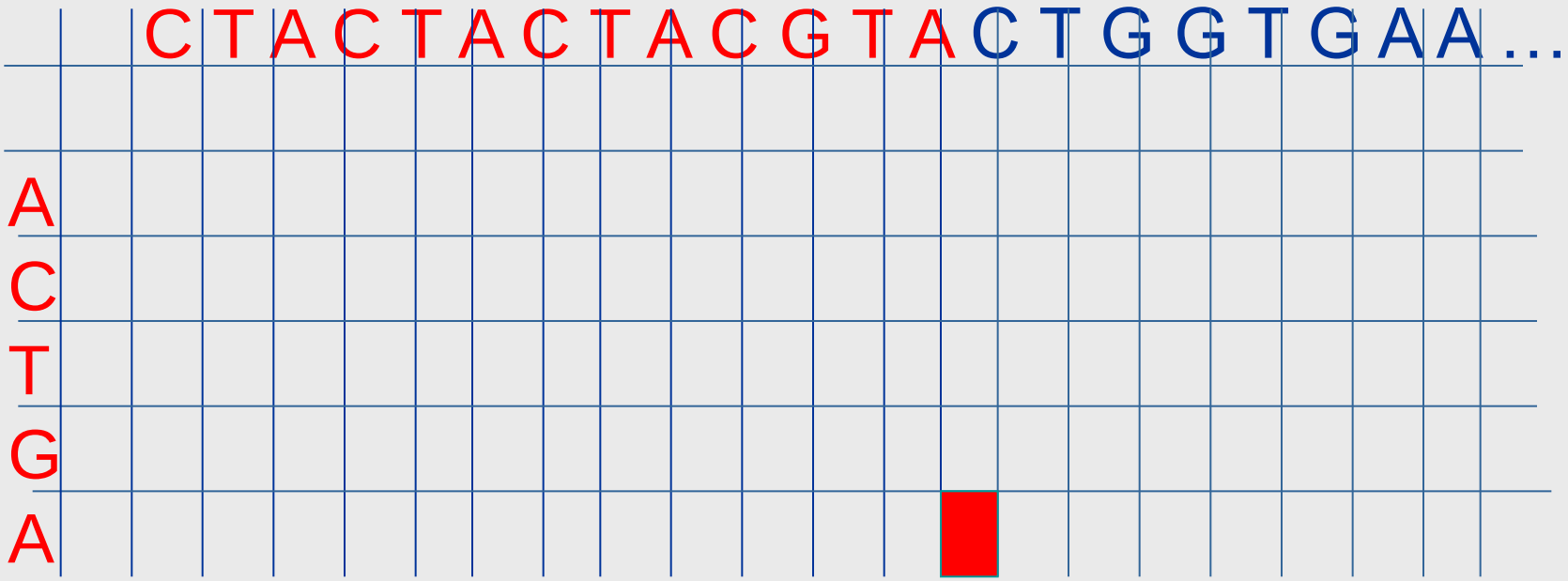
to the K-approximate string searching?

# K-approximate string searching




This cell …

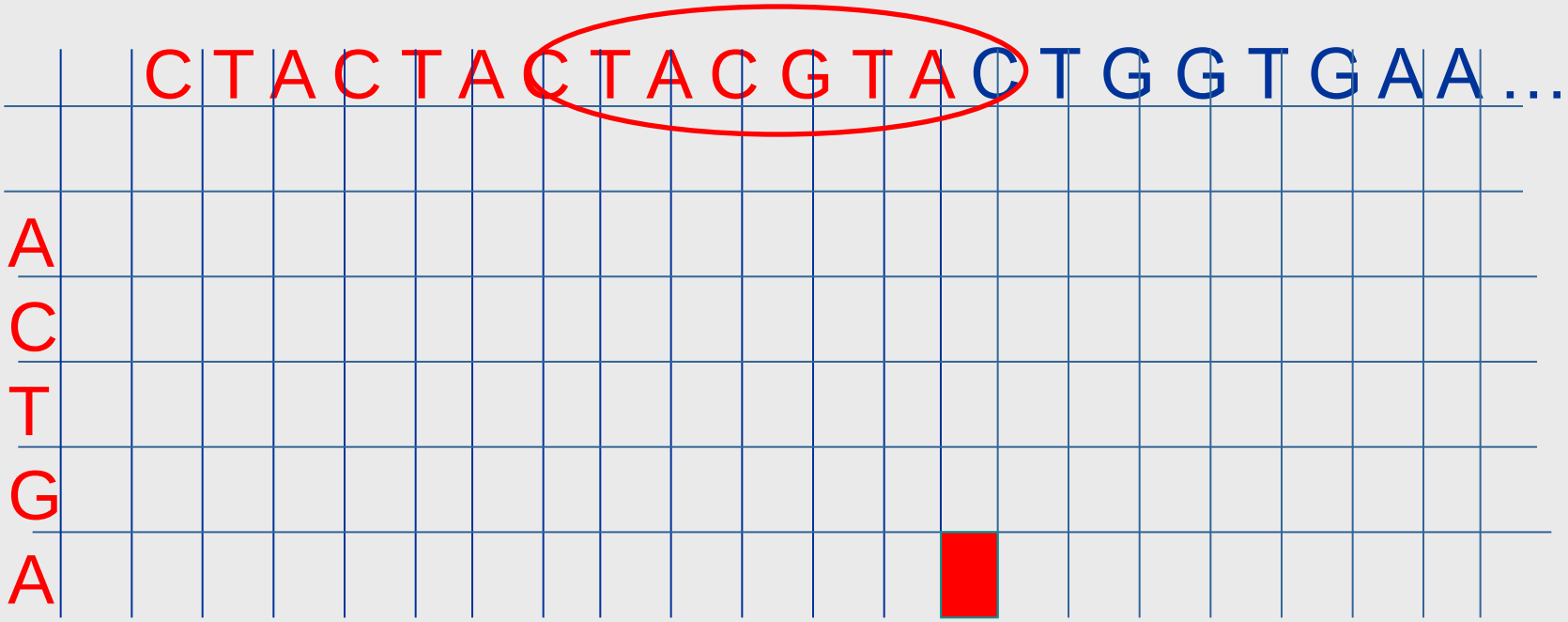
# K-approximate string searching




This cell gives the distance between (ACTGA, CT…GTA)…

…but we only are interested in the last characters

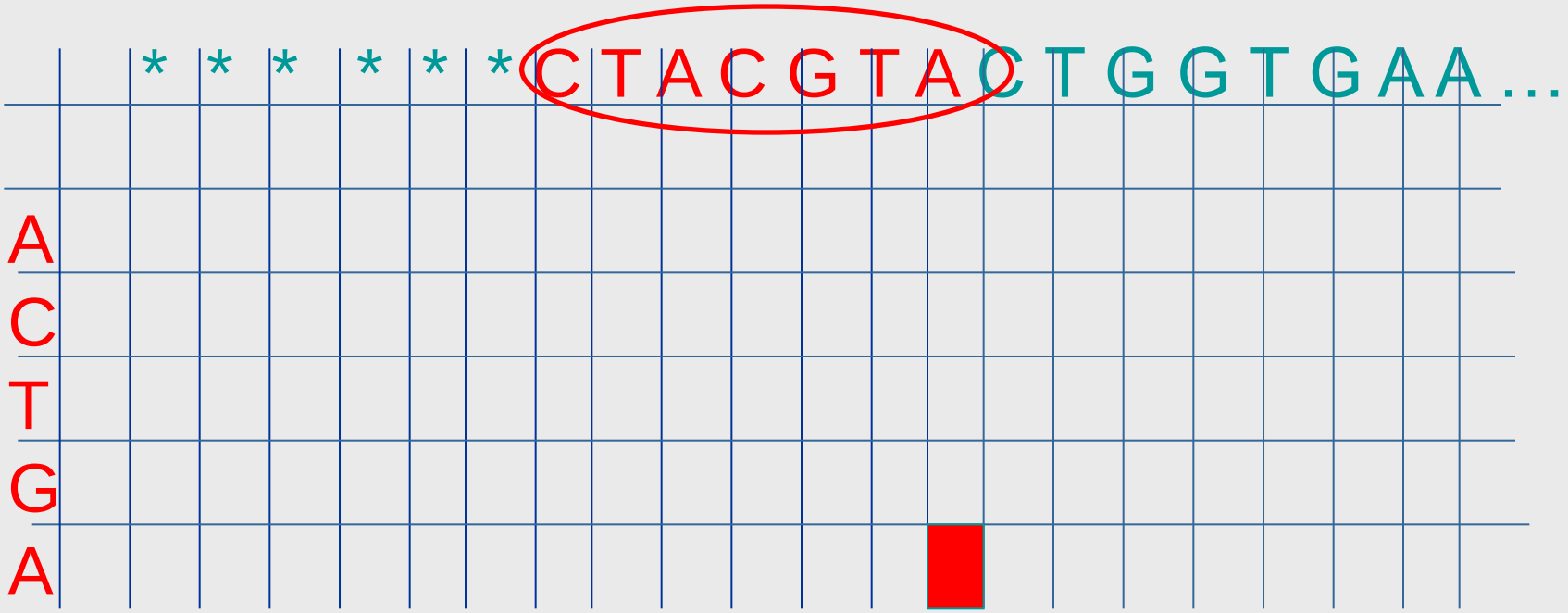
# K-approximate string searching




This cell gives the distance between (ACTGA, CT…GTA)…

…but we only are interested in the last characters

# K-approximate string searching




This cell gives the distance between (ACTGA, CT…GTA)…

…but we only are interested in the last characters…

…no matter where they appears in the text, then…

# K-approximate string searching


This cell gives the distance between (ACTGA, CT…GTA)…

…but we only are interested in the last characters…

…no matter where they appears in the text, then…

# K-approximate string searching

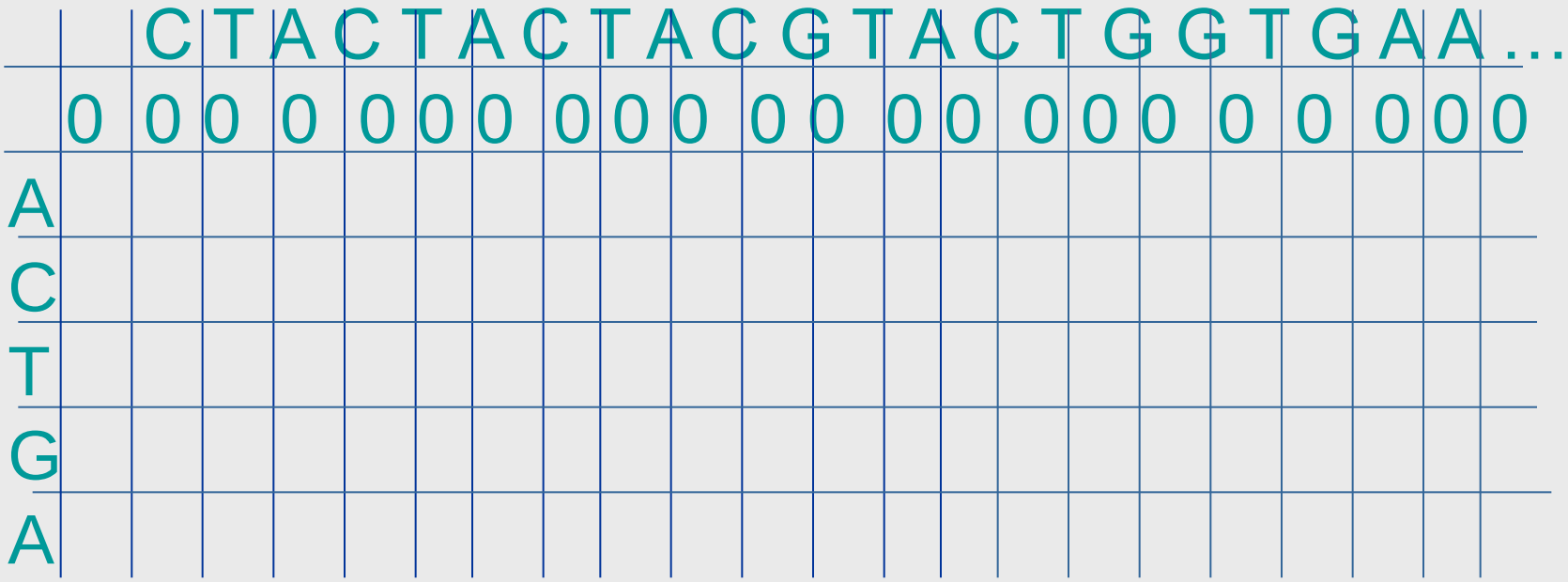

This cell gives the distance between (ACTGA, CT…GTA)…

…but we only are interested in the last characters…

…no matter where they appears in the text, then…

# K-approximate string searching

algorithmics genetics group




This cell gives the distance between (ACTGA, CT…GTA)…

…but we only are interested in the last characters…

…no matter where they appears in the text, then

# K-approximate string searching

Connect to

http://alggen.lsi.upc.es/docencia/ember/leed/Tfc1.htm

and use the semi-global method.