

1. Genomic texts

- The cell, atom of the living world
- At the heart of the cell: the DNA macromolecule
- DNA codes for genetic information
- What is an algorithm?
- Counting nucleotides
- GC and AT contents of DNA sequence
- DNA walk
- **Compressing the DNA walk**
- Predicting the origin of DNA replication?
- Overlapping sliding window

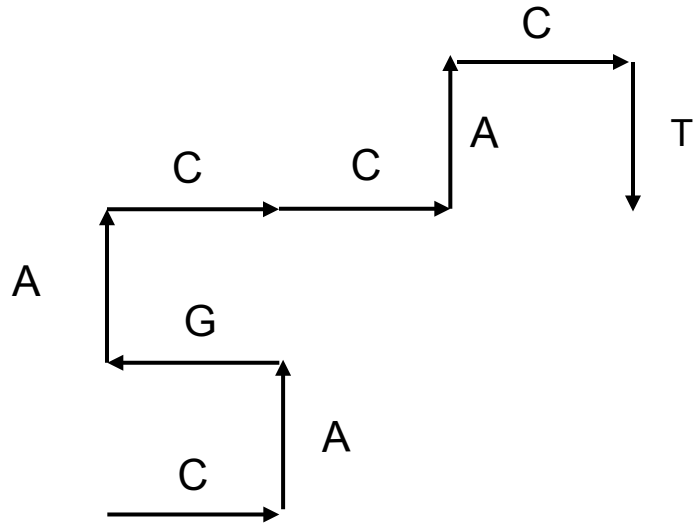
What about the screen size?

- Resolution of a screen
 - The number of distinct pixels in each dimension that can be displayed
 - For example: 1024 x 768

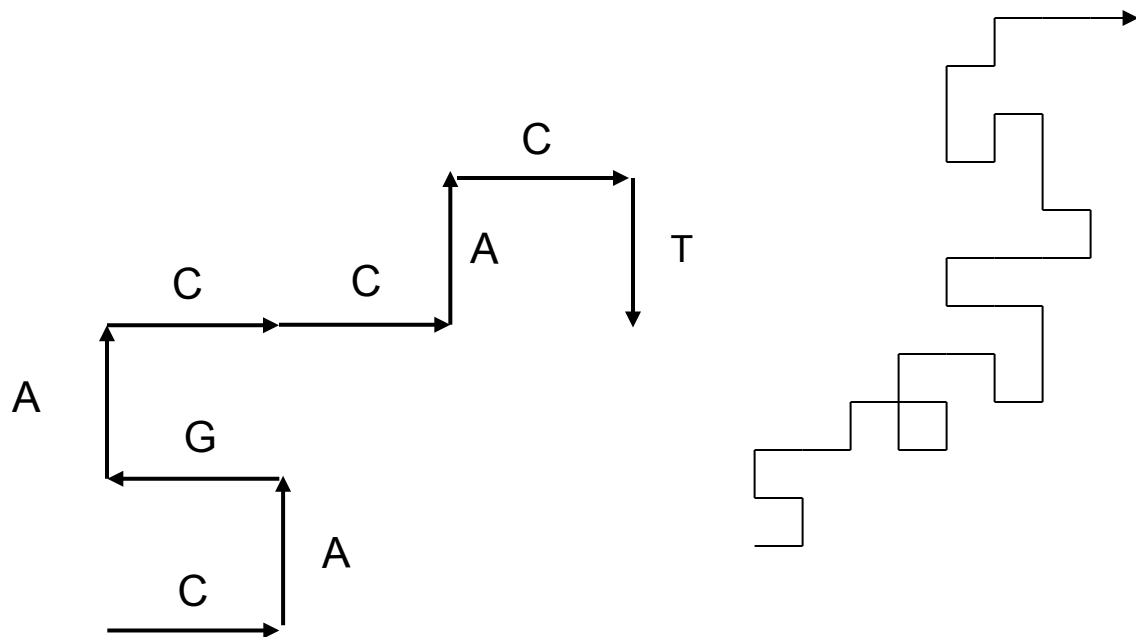
- Problem:

How to fit a series of several millions or billions segments in one screen?

- Compression is the answer



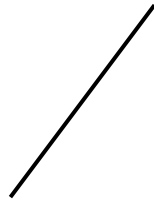
CAGACCACTCAGACCTCAAGGACCCAGAAGTGAACACC...



CAGACCACTCAGACCTCAAGGACCCAGAAAGTGAACAC

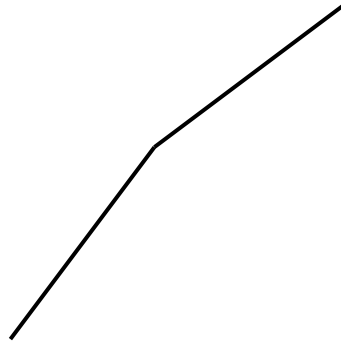
CAGACCACTCAGACCTCAAGGACCCAGAAGTGAACAC

- compute nbA, nbC, nbG, nbT in the current window of length L
- compute the coordinates of the end point of the segment
- draw the segment
- move the window forward



CAGACCACTCAGACCTCAAGGACCCAGAAAGTGAACAC

- compute nbA, nbC, nbG, nbT in the current window of length L
- compute the coordinates of the end point of the segment
- draw the segment
- move the window forward



```
L, nbA,nbC,nbG,nbT: integer
sequence: character string [1:*]
nbA,nbC,nbG,nbT  $\leftarrow$  0
for i from 1 to L do
  case sequence [i] of
    "A": nbA  $\leftarrow$  nbA + 1
    "C": nbC  $\leftarrow$  nbC + 1
    "G": nbG  $\leftarrow$  nbG + 1
    "T": nbT  $\leftarrow$  nbT + 1
  endcase
endfor
```


SeqLength, L, InitW, nbA,nbC,nbG,nbT, NbStepsRight, NbStepsUp: **integer**

XEndSegment, YEndSegment, Step: **real**

sequence: **character string** [1:*

InitW ← 1

repeat

nbA,nbC,nbG,nbT ← 0

for i **from** InitW **to** InitW + L - 1 **do**

case sequence [i] **of**

"A": nbA ← nbA + 1

"C": nbC ← nbC + 1

"G": nbG ← nbG + 1

"T": nbT ← nbT + 1

endcase

endfor

NbStepsRight ← nbC - NbG

NbStepsUp ← nbA - nbT

XEndSegment ← NbStepsRights * Step

YEndSegment ← NbStepsUp * Step

DrawTill (XEndSegment, YEndSegment)

InitW ← InitW + L

until InitW > SeqLength

CAGACCACTCAGACCTCAAGGACCCAGAAGTGAACAC

