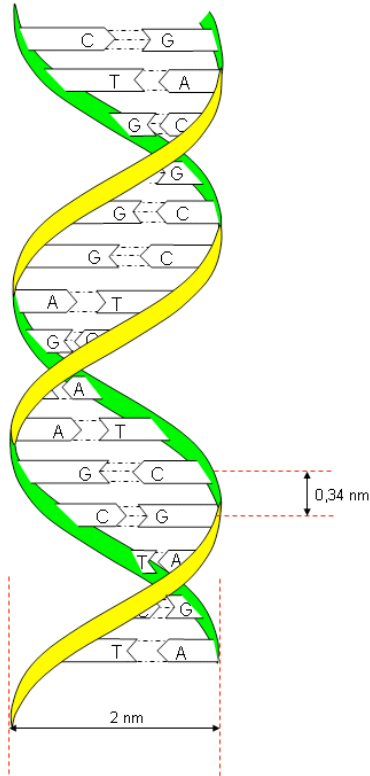# 1. Genomic texts

- 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

DNA codes for genetic information

The sequence of nucleotides codes for the genetic information

Sequencing → **AGCTAGAGGCCAGTTCG…**

Genome

- The DNA molecule, as the support of genetic information
  - may be organized into chromosomes, plasmids, segments…
  - "sequencing the Human genome", "annotating the *E. coli* genome"

- The set of genes of an organism
  - genes are regions of DNA which code for proteins
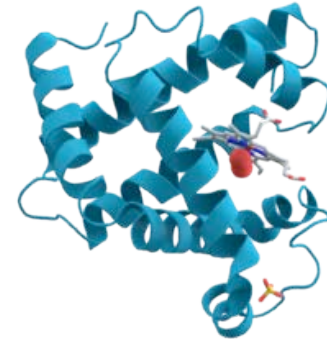 - "functions of a gene"

Did you say information?

- 0110111001011011100101011111001011…
  - numbers
  - pictures
  - videos
  - music

Did you say information?

- 01101110010110111001010111110011011…
  - numbers
  - pictures
  - videos
  - music

- GAAGGGAGGAAGTGCCTGA….
  - proteins
  - signals for protein expression

Good news!

Computer science, or better:

Informatics

deals with information processing

Bioinformatics

deals with genetic information processing

The nature of genetic information

- Texts in the 4-letter alphabet

  - A, C, G, T

  - The initials of the four types of nucleotides, or bases

- TTACGTATTGCTATTG….

  - no ponctuation

  - no space

  - no evident markers

AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCC
AGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTG
AAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTT
GCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGT
TACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCT
GAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCG
CCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGC
TGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGT
CAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCG
CTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATAC
CGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGC
ATTTCCAATCTGAATAACATGGCAATGTTCAGCGTTTCTGGTCCGGGGATGAAAGGGATGGTCGGCATGG
CGGCGCGCGTCTTTGCAGCGATGTCACGCGCCCGTATTTCCGTGGTGCTGATTACGCAATCATCTTCCGA
ATACAGCATCAGTTTCTGCGTTCCACAAAGCGACTGTGTGCGAGCTGAACGGGCAATGCAGGAAGAGTTC
TACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACGGCTGGCCATTATCTCGGTGG
TAGGTGATGGTATGCGCACCTTGCGTGGGATCTCGGCGAAATTCTTTGCCGCACTGGCCCGCGCCAATAT
CAACATTGTCGCCATTGCTCAGGGATCTTCTGAACGCTCAATCTCTGTCGTGGTAAATAACGATGATGCG
ACCACTGGCGTGCGCGTTACTCATCAGATGCTGTTCAATACCGATCAGGTTATCGAAGTGTTTGTGATTG
GCGTCGGTGGCGTTGGCGGTGCGCTGCTGGAGCAACTGAAGCGTCAGCAAAGCTGGCTGAAGAATAAACA
TATCGACTTACGTGTCTGCGGTGTTGCCAACTCGAAGGCTCTGCTCACCAATGTACATGGCCTTAATCTG
GAAAACTGGCAGGAAGAACTGGCGCAAGCCAAAGAGCCGTTTAATCTCGGGCGCTTAATTCGCCTCGTGA
AAGAATATCATCTGCTGAACCCGGTCATTGTTGACTGCACTTCCAGCCAGGCAGTGGCGGATCAATATGC
CGACTTCCTGCGCGAAGGTTTCCACGTTGTCACGCCGAACAA

Some orders of magnitude

- Mb "megabase": 10^6 bases
  - *E. coli* 4.5 Mb

Some orders of magnitude

- Mb "megabase": 10^6 bases
  - *E. coli* 4.5 Mb

- Gb "gigabase": 10^9 bases
  - *H. sapiens* 3.5 Gb

Some orders of magnitude

- Mb "megabase": 10^6 bases
  - *E. coli* 4.5 Mb

- Gb "gigabase": 10^9 bases
  - *H. sapiens* 3.5 Gb

- Tb "terabase": 10^{12} bases
  - *Amoeba dubia* 670 Gb = 0.67 Tb

Pictures & movies : material licensing

p. 3 : Nature Education / Conditions of use for academic research : users may view, print, copy, download and text and data-mine the content

p. 5, p. 6 : DanielZanetti (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

p. 6, p. 8 : Fotolia

p. 13 :
https://microbewiki.kenyon.edu/index.php/File:Amoeba_proteus_x_100.jpg#filelinks