Fractales:
Application à l'analyse du génome

Alain Arneodo

Laboratoire Joliot-Curie / Laboratoire de Physique,
Ecole Normale Supérieure de Lyon
Alain.Arneodo@ens-lyon.fr

Benjamin Audit
Guillaume Chevereau
Julien Moukhtar
Leonor Palmeira
Philippe StJean
Cédric Vaillant
Lamia Zaghloul

Françoise Argoul
Zofia Haftek-Terreau
Monique Marilley
Pascale Milani
Cendrine Moskalenko

Yves d'Aubenton-Carafa
Claude Thermes

CGM, Gif-sur-Yvette, France

http://www.ens-lyon.fr/PHYSIQUE/index.php?page=equipe5
http://www.ens-lyon.fr/Joliot-Curie/

DESOXYRIBONUCLEIC ACID

A FEW HISTORICAL LANDMARKS

1869 Miescher isolates DNA

1944 DNA carries the genetic information (Avery)

1953 The double helix structure of DNA is discovered by Watson and Crick

A T G C
T A C G

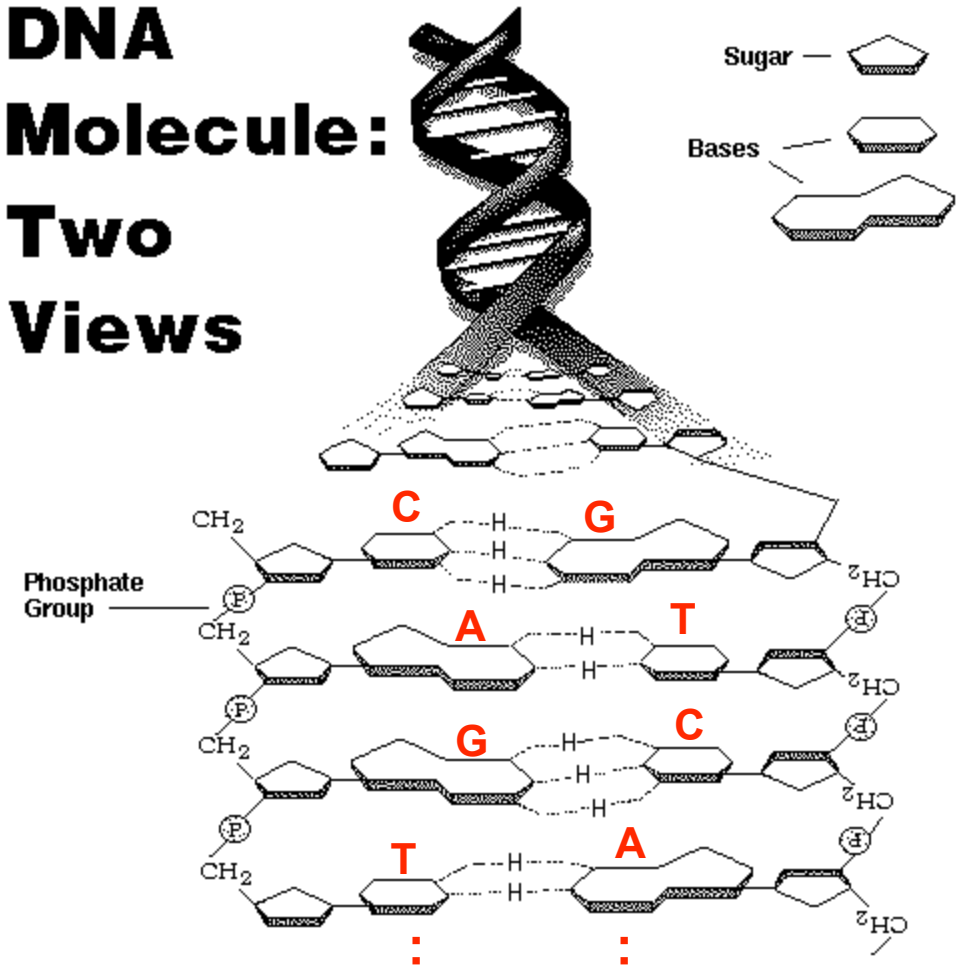
$\rightarrow$ a simple model for the transmission of the genetic information

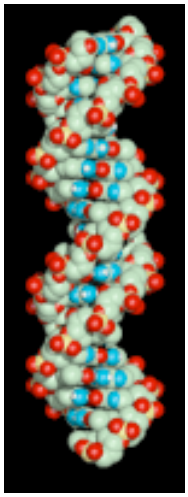
1966 Niremberg, Ochoa and Khorana elucidate the genetic code

$\rightarrow$ DNA codes for proteins

codon	ATG	GCG	ACG	...	GCC	GTG	TAA
amino acid	Met	Ala	Thr	...	Ala	Val	
	start						stop

DeoxyriboNucleic Acid



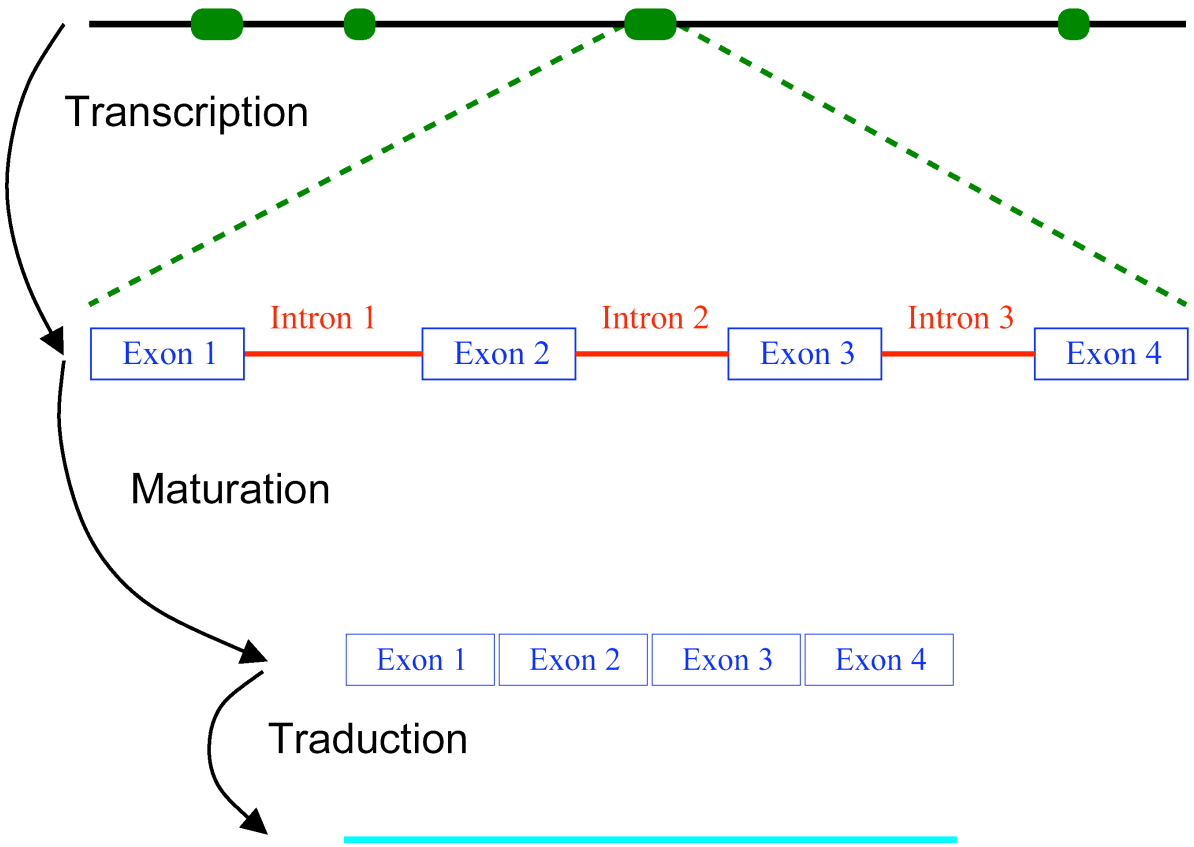


- Double helix macromolecule
- Each strand consists of an oriented sequence of four possible nucleotides:
  Adenine, Thymine, Guanine & Cytosine
- Complementary strands:
  [A]=[T] & [G]=[C] over the sum of both strands

Sequencing projects result in 4 letter texts :

gtcagtttcctgaggcgggtcgggacccaggcgtgagactggagtctgcc
caggggcccagctgagccagcctcctcgtcagctgcttgggccgccagga
cgccgccgggggtgcgccgcgcttccctggatggggtgcccccactcccc
tcggagccccagggagacccccgaactcagctcctctcaggggtgccag
ggggacccctcaaactccactccccgcaggttcctggggagacgccccct
gctcgattcccctcagggtcccagggagacccccctaattcagctcctctc
aggggtactgggggacctctcgagctccactcccatcagggtcccaggga
gacccccccaactatgctcaggggtcccagggagatgccagcaccccaact
ccgcttccctgggggccccctcccccttacagctcaacttccctcgagagt
ctggggctggggctccgttcagttcttgagtccccttccctcggggtgtc
ccggggccgcccacccccacactgtctgtgattccccaaggcgcgggtct
cgggccgcagcctgttccacgttctgctgctcgttcttttctggctcctt
gctttcgaaggagagaaggaggccttcgtttccagtctttttgccttttc
taatggagccctgcttttccttccgtgtcccttcaggctacttctgccag
gtttctatttttcattctttattatgacttcgcccaaaatattcttgact
tctattgagaaggattcgggggtctatttcttattcggaggcgtgtgctt
aagttccaaacagatgaggattttccagttaatccttctggggtgactta
ttgcttaatgccaccatagccagaaaatggactctcagtgtccgaaactg
cattcggctctgaagtgtctgtccttgtcacctcttgcaatgtttcgcgg
cgggaagcctgcactcgccgacgctgacgtaactgtttctgtctttcagg
tctacagcctcctgtgggtgggcgatattgacatatactttatttctata
tatgttatgaactcaatatttcttgcagcgggtctgctgataataagata
tgcctactctgcgagtctggaagccatcttaagcttaccctgtatgtgcc
ccatgcatctcttccgttacacggctcctgagttgacacctgtgtgataa
actggtaatagcaagtaaactgttttcttgtgctctgtaagctgctctag
caaattatctaggaggaggtggtcttggaaacccctgatttataagcggg
cagtcagcagtacacgtggcccagaatcgtgattggcatttgaagtgggg
gcagtagggtgggactgagcccttcacctgtggggtctgccctgctcaag
gcagtgtcagaattgaagtgaaatgttggacggtcggtgtccagagagtt
ggagaactggtttgtgtgtaaaaactnacatatttagggtcagaagtatg
...

ORGANIZATION OF THE HUMAN GENOME

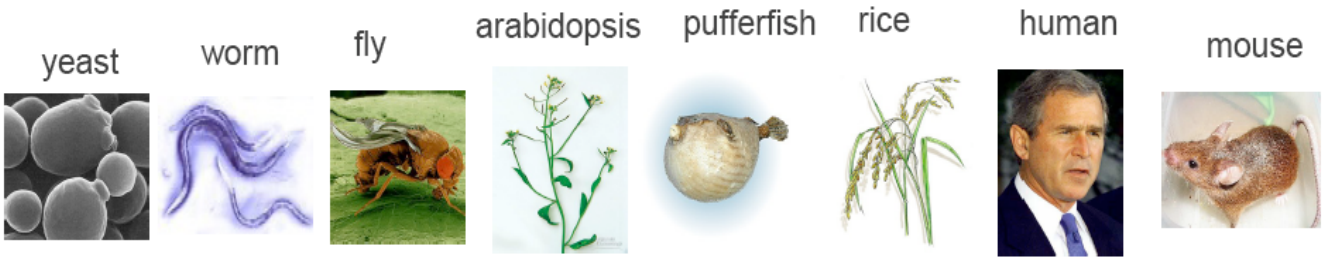


23 Chromosomes
$L \sim$ 100Mbp.

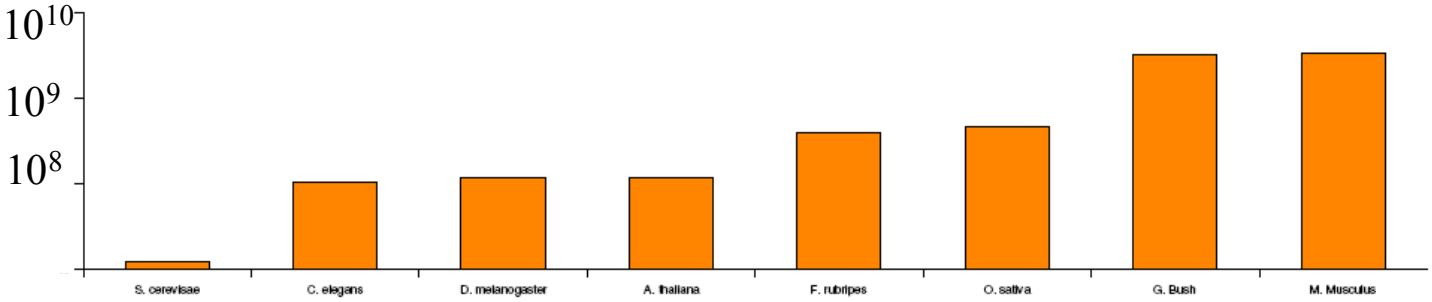
- Genes ($\sim$ 20%)
  $L \sim$ 10kbp.
  - Introns
    (INTervening seq.)
    $L \sim$ 1kbp.
  - Exons
    (EXpressed seq.) $\Rightarrow$ Proteins
    $L \sim$ 150bp.
    $L \sim$ 500AA.
- Non genic DNA

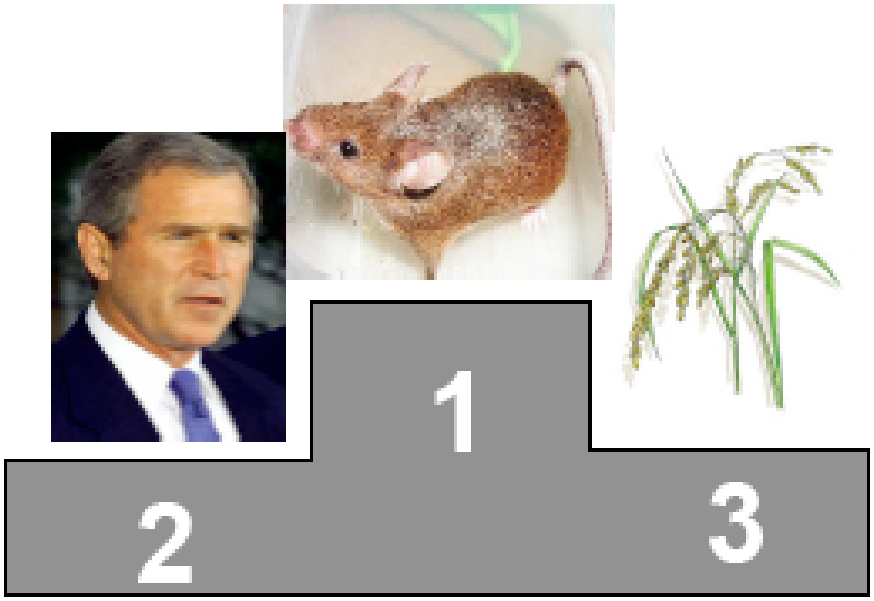
Eukaryotic genome context

(C. Hermann)



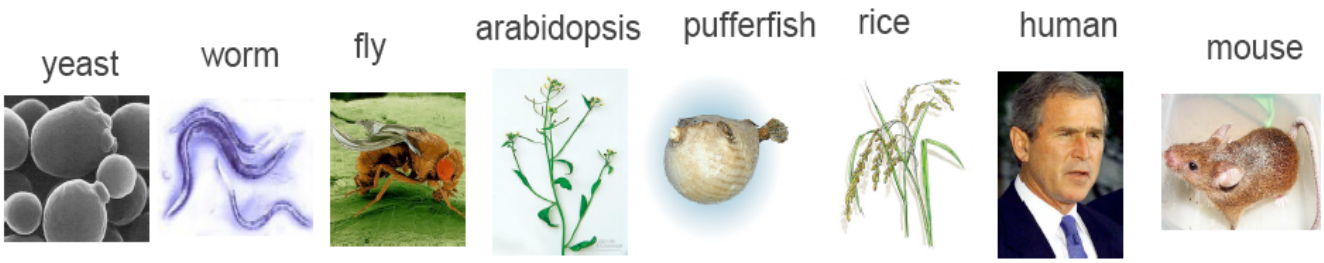
Genome size



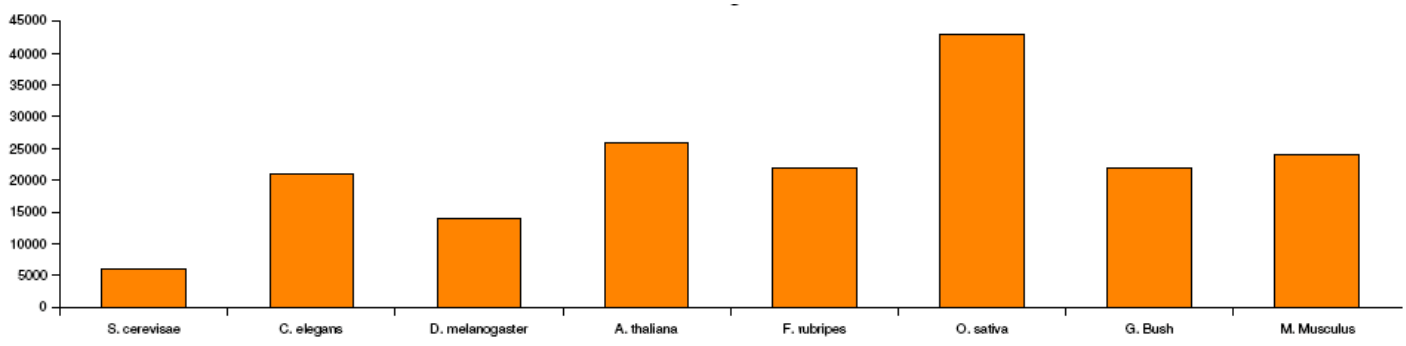


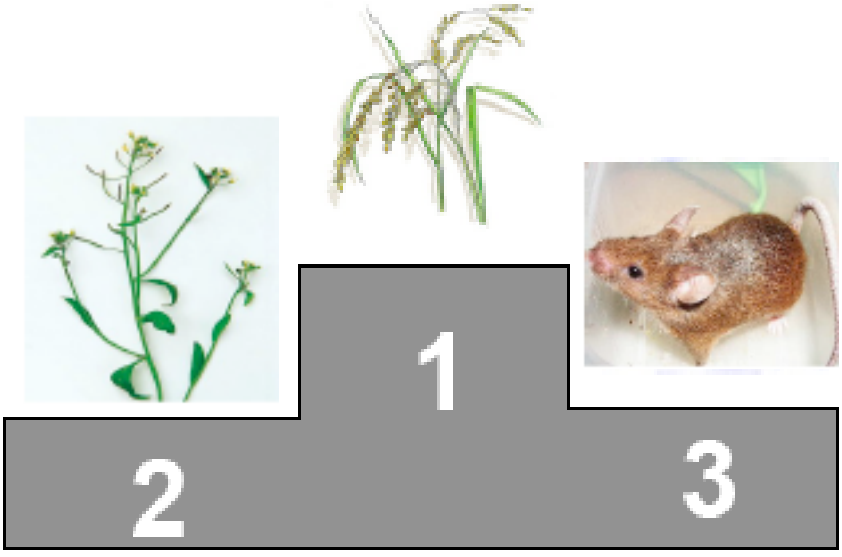
Eukaryotic genome context

(C. Hermann)



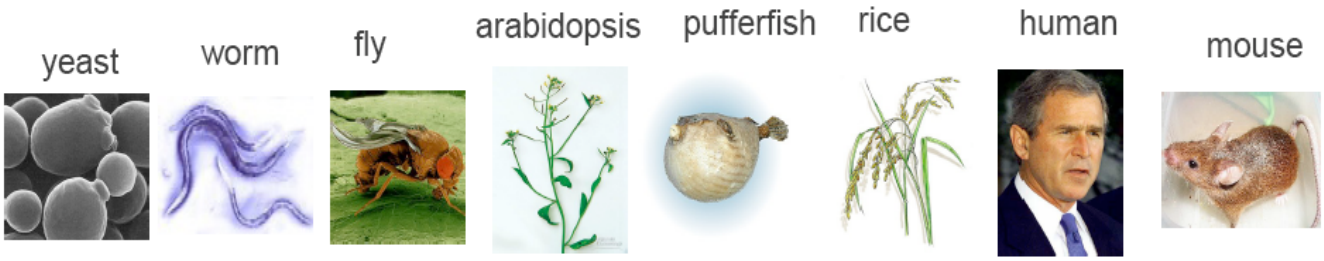
Number of genes



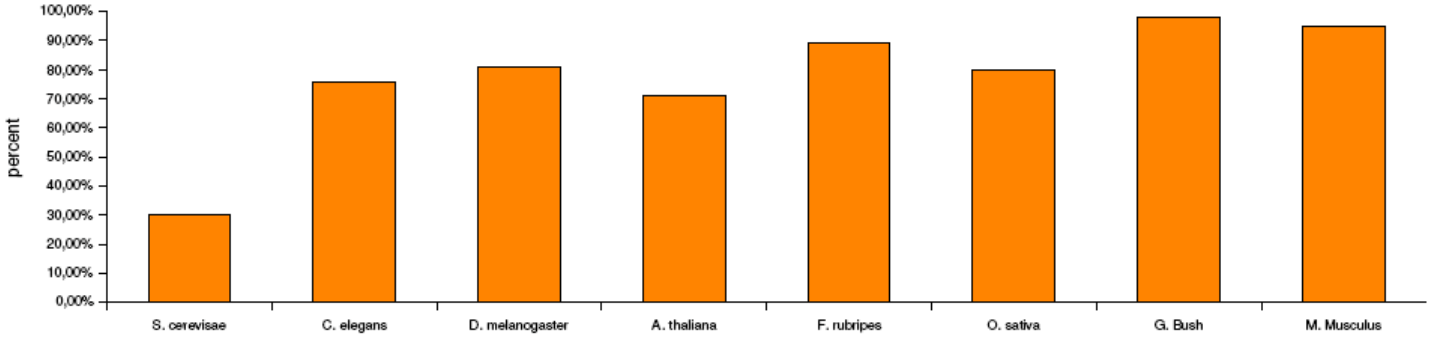


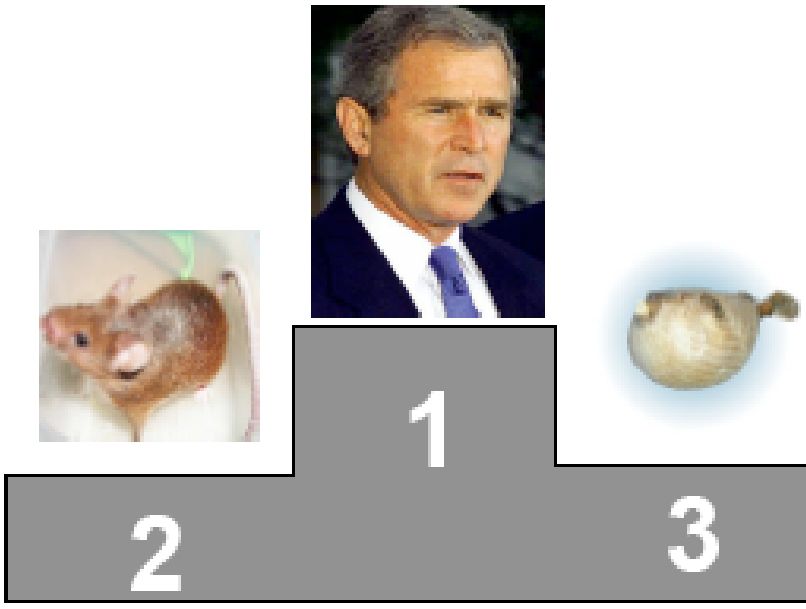
Eukaryotic genome context

(C. Hermann)



Non-coding DNA

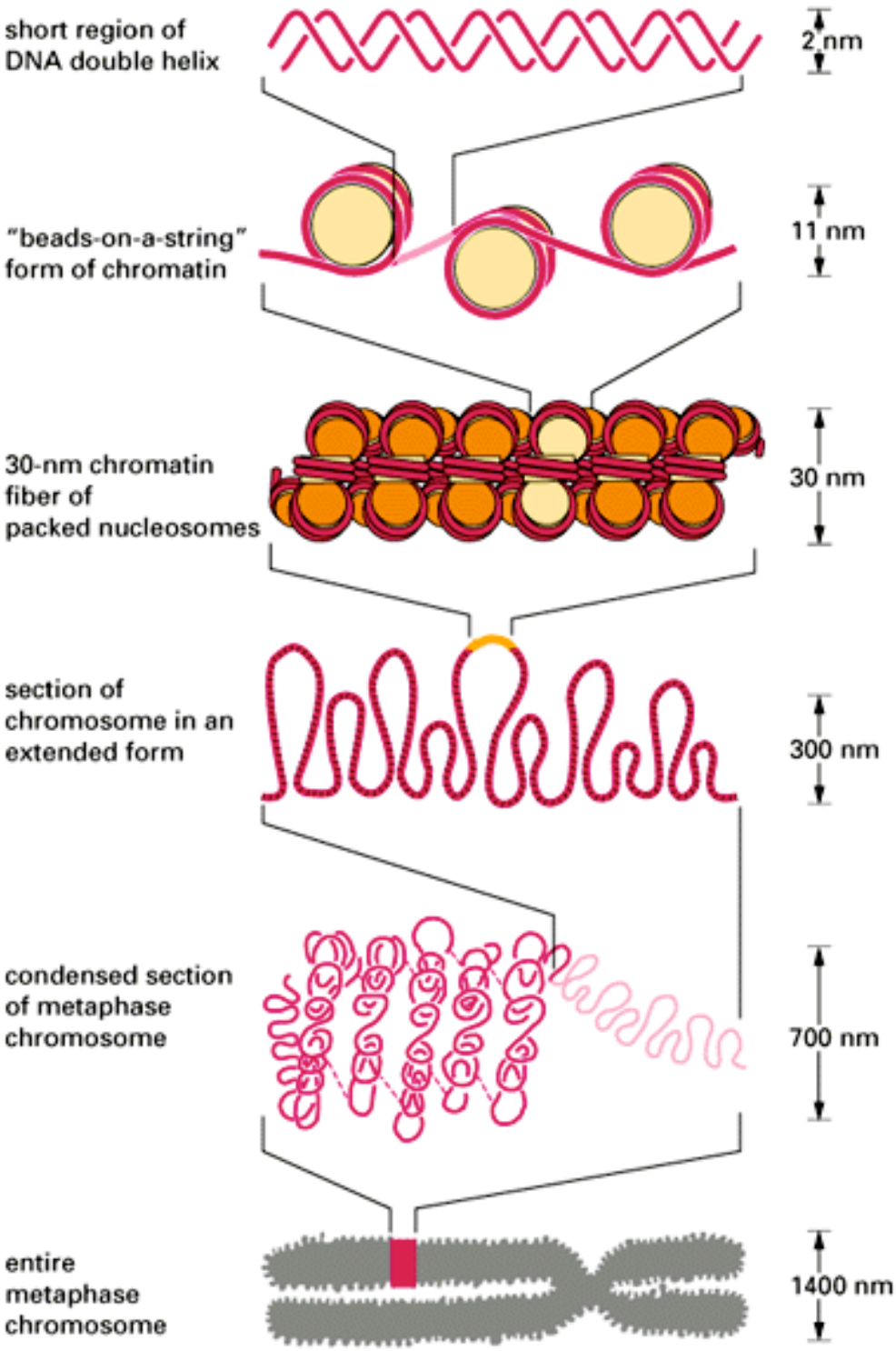




Eukaryotic genome context

(C. Hermann)

HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA



NET RESULT : EACH DNA MOLECULE HAS BEEN PACKAGED INTO A MITOTIC CHROMOSOME THAT IS 50.000x SHORTER THAN ITS EXTENDED LENGTH

DIFFERENT WAYS TO READ THE TEXT

I. "Classical" reading

- Looking for patterns
  - Genes, introns, exons detection
  - Splicing sites, promoters, replication origins recognition
- Characterizing repetitions
  - Tandem, interspersed repeats
  - Oligonucleotide usage
- Using methods such as
  - Hidden Markov chains
  - Fourier transform
  - Dot-plot matrices and recurrence plots

II. The physicist reading

- Hypothesis: The DNA text results from a stochastic process :

  ACGTTCGAT ?

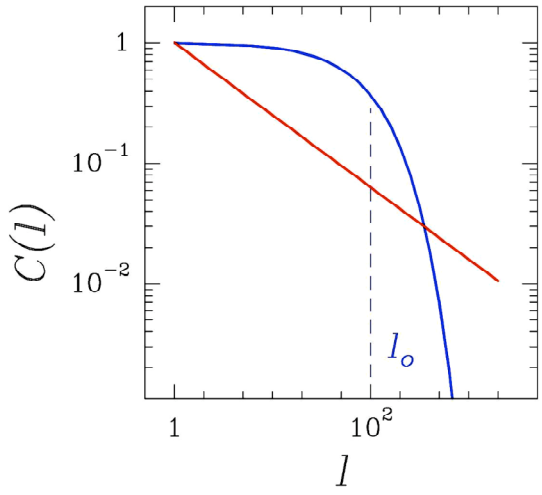
- Question: The choice of the next nucleotide :

  i. Depends on a finite number (l_o) of the previous trials
  $\rightarrow$ Short range correlations and exponential decay of the correlation function:

  $$C(l) \propto \exp(-l/l_o)$$

  ii. Depends on all the previous nucleotides
  $\rightarrow$ Long range correlations and power law decay of the correlation function:

  $$C(l) \propto l^{-\kappa}$$



INVARIANCE UNDER DILATATION

DNA WALK REPRESENTATION (PENG *et al.* 92)

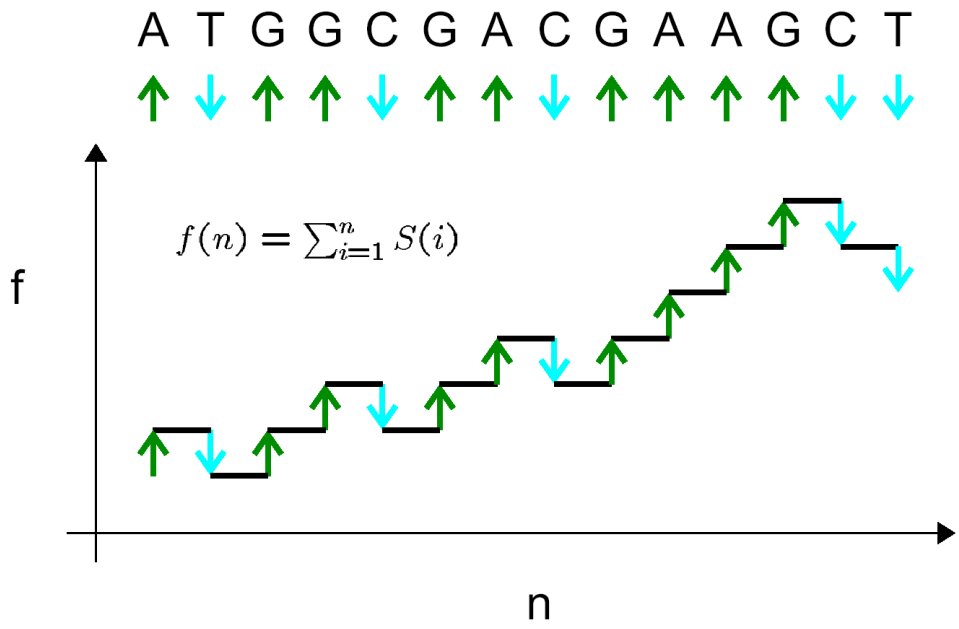
1. Each nucleotide is associated to a numerical value
   (A to a, T to t, G to g and C to c).

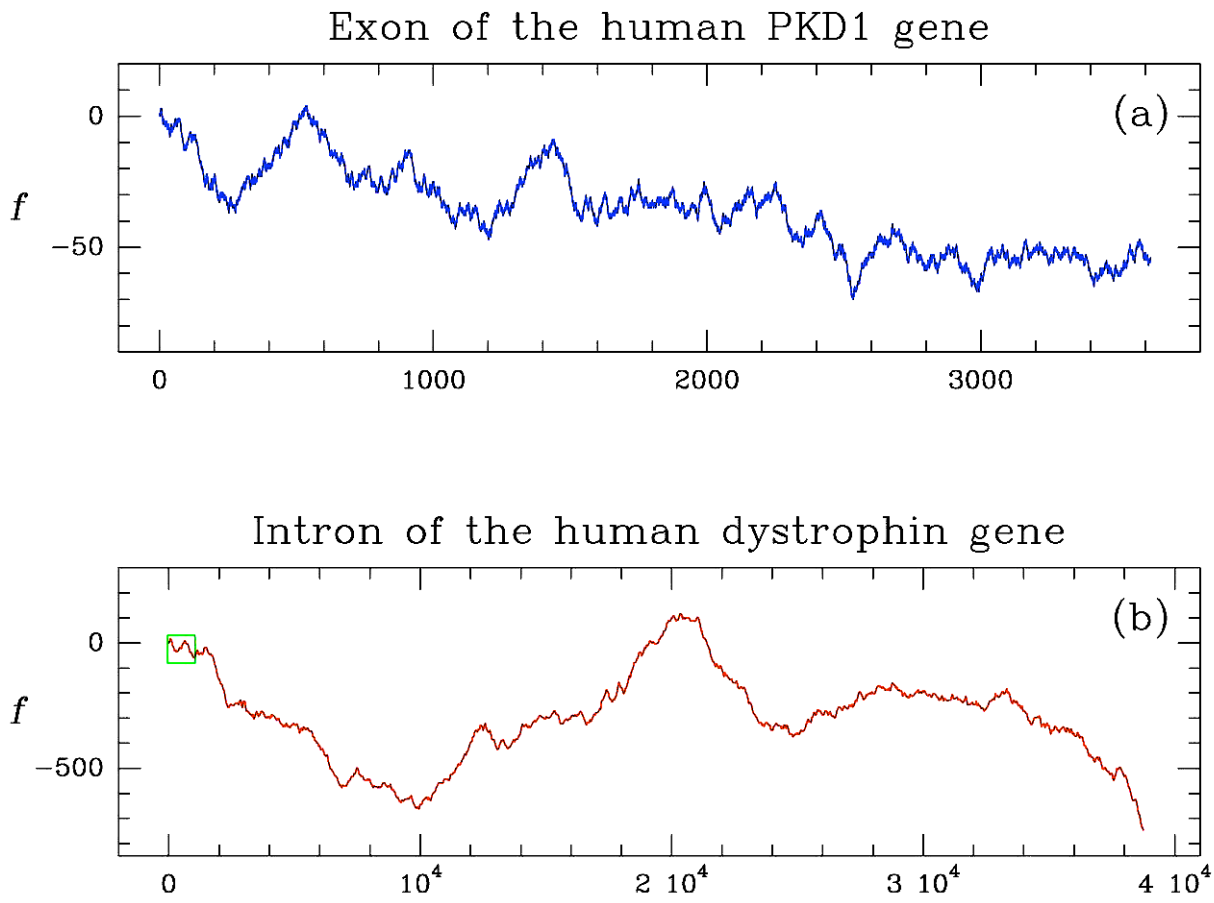
   purine-pyrimidine : $a = g = 1$ and $t = c = -1$
   weak-strong : $a = t = 1$ and $g = c = -1$
   amino-keto : $a = c = 1$ and $t = g = -1$

   A-non A : $a = 1$ and $t = g = c = -1/3$
   T-non T : $t = 1$ and $a = g = c = -1/3$
   G-non G : $g = 1$ and $a = t = c = -1/3$
   C-non C : $c = 1$ and $a = t = g = -1/3$

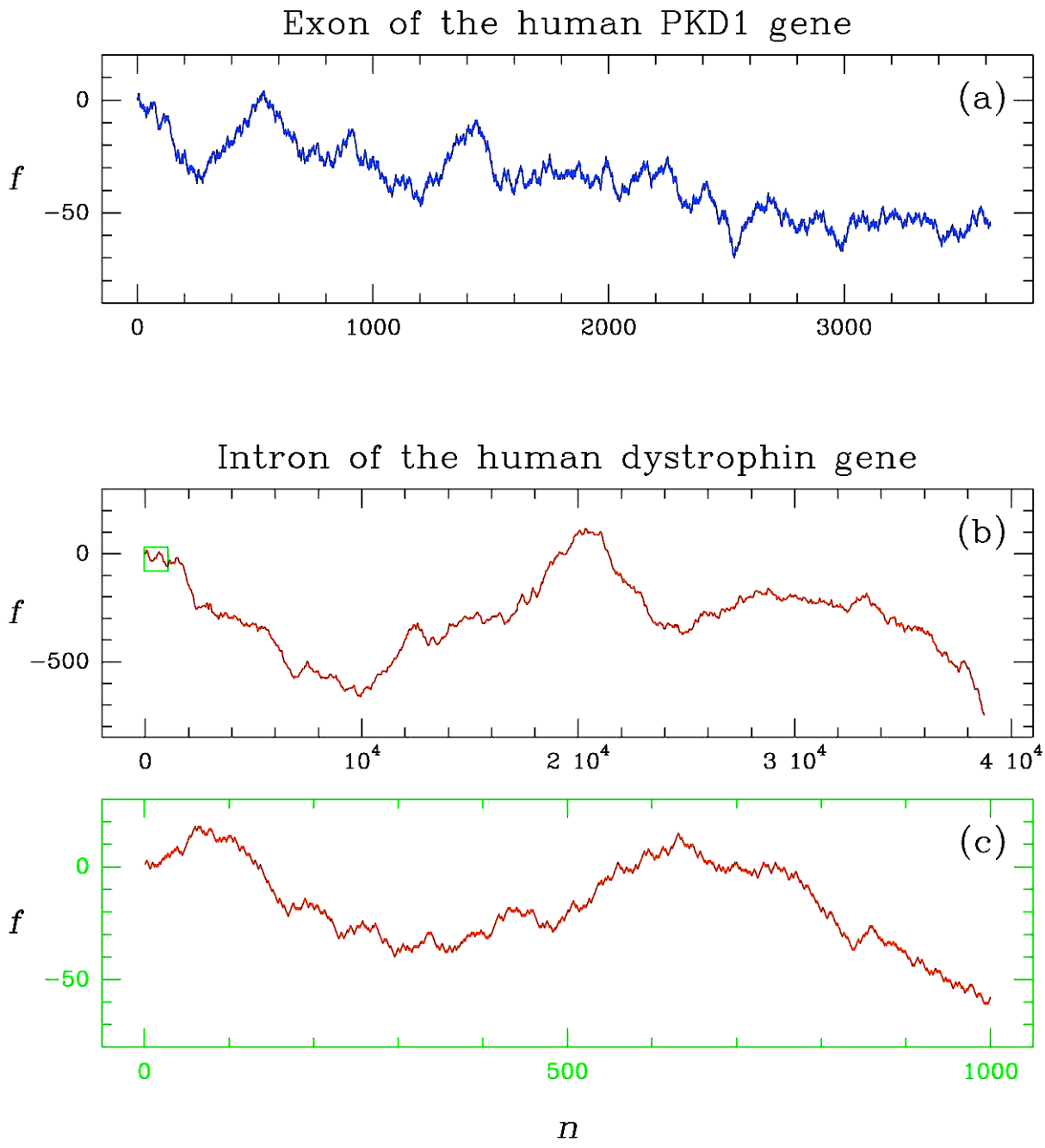
2. Suppose you have a walker on the line. The value associated to the i^{th} nucleotide defines the i^{th} step $S(i)$ of the walker

Example using the purine (↑) pyrimidine (↓) distinction :





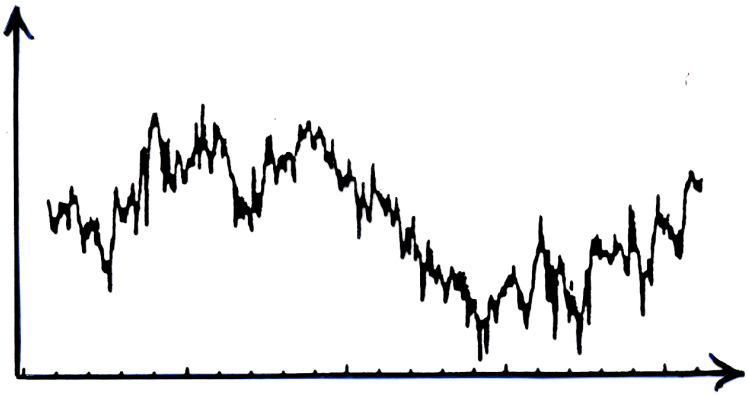
Most of the physicist works amount to characterizing the roughness of a DNA walk landscape



Most of the physicist works amount to characterizing the roughness of a DNA walk landscape

FRACTAL SIGNALS

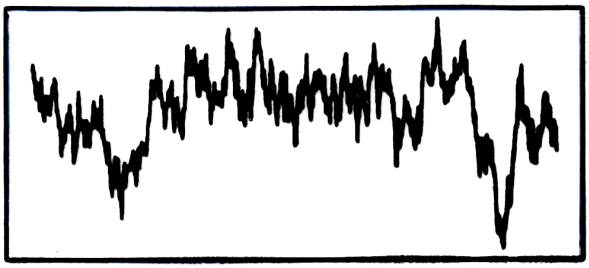
V(t)



time

Turbulent velocity signal

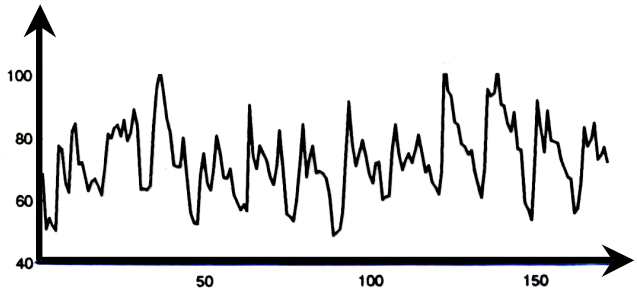
S(t)



time

Brownian signal
“ 1/f noise”

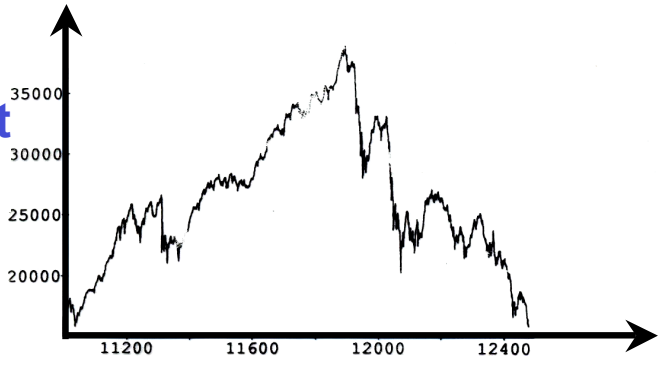
Heart rate



time

Medical signal

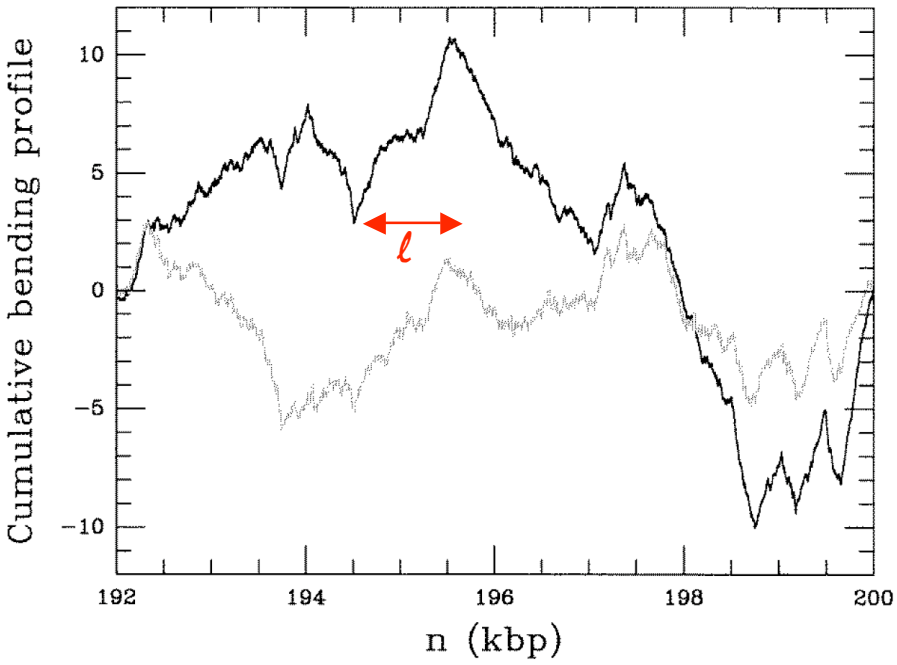
Market prices



days

Financial time series

Roughness exponent



• Root-mean square of the height fluctuations

$$W(\ell) = \text{rms}\,[f(n+\ell)-f(n)] \sim \ell^{H}$$

H = roughness exponent $D_f = 2 - H$

• Power spectrum

$$S_f(k) \sim k^{-(2H+1)}$$

• Correlation function

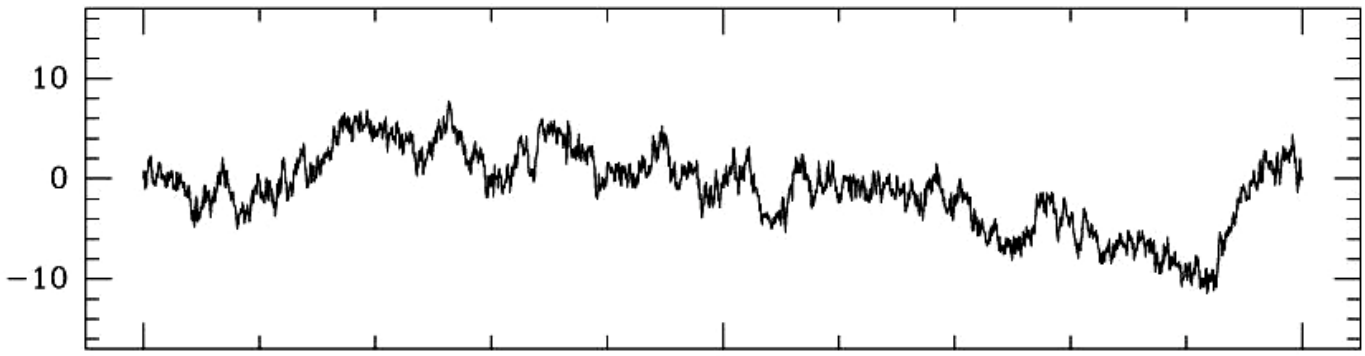
$$C_f(\tau) = \langle \Delta_1 f(n)\, \Delta_1 f(n+\tau) \rangle - \langle \Delta_1 f(n) \rangle^2 \sim \tau^{2H-2}$$

SYNTHETIC DNA WALKS

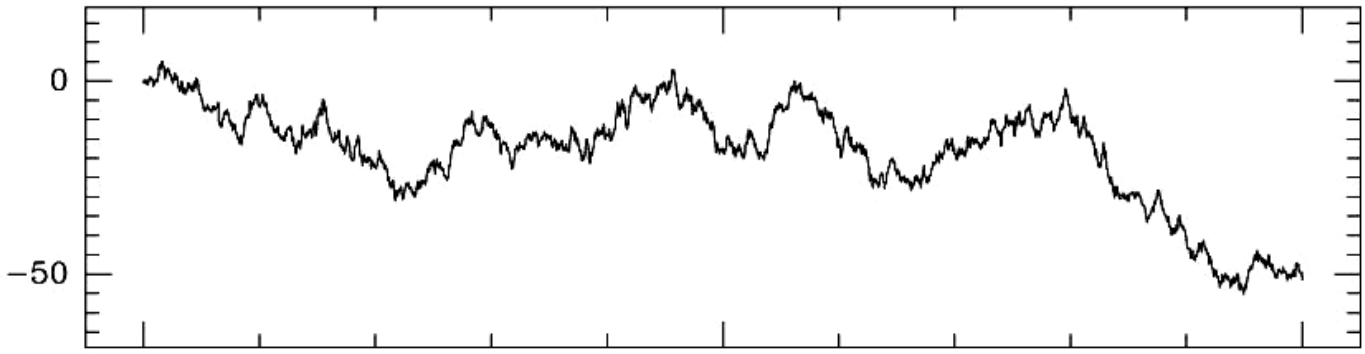
Fractional Brownian motions : B_H

Fractal dimension: $D_f = 2 - H$
H = roughness exponent

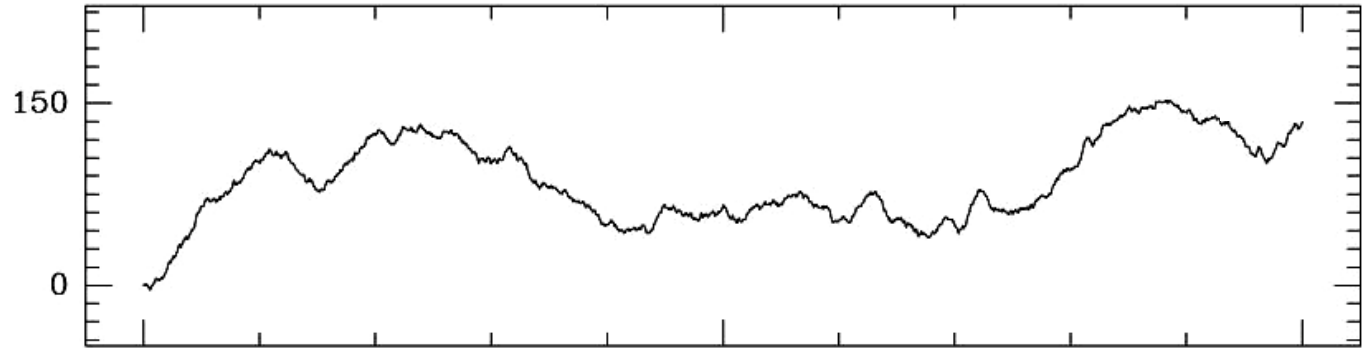
H = 0.3 anti-correlated



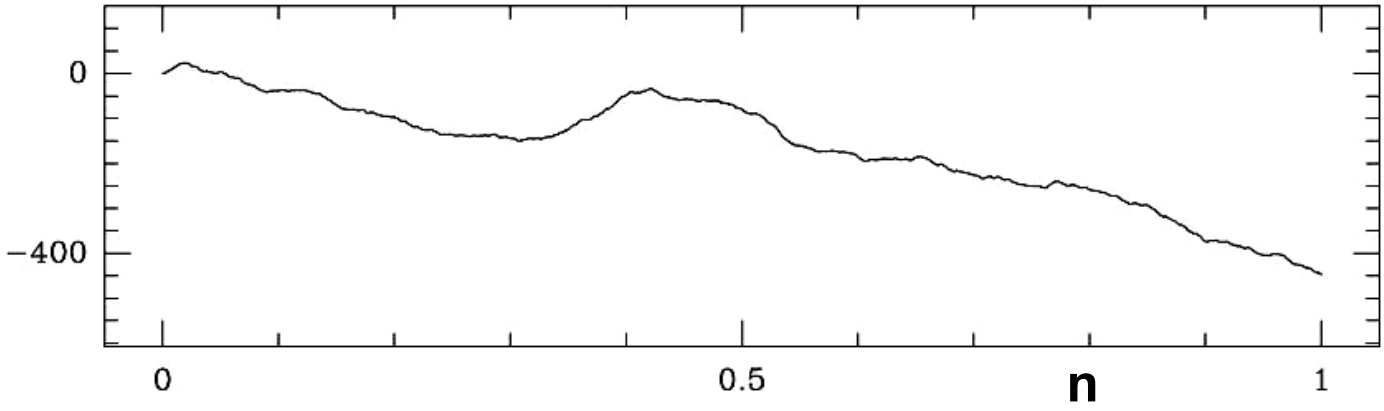
H = 0.5 uncorrelated



H = 0.7 long-range correlated



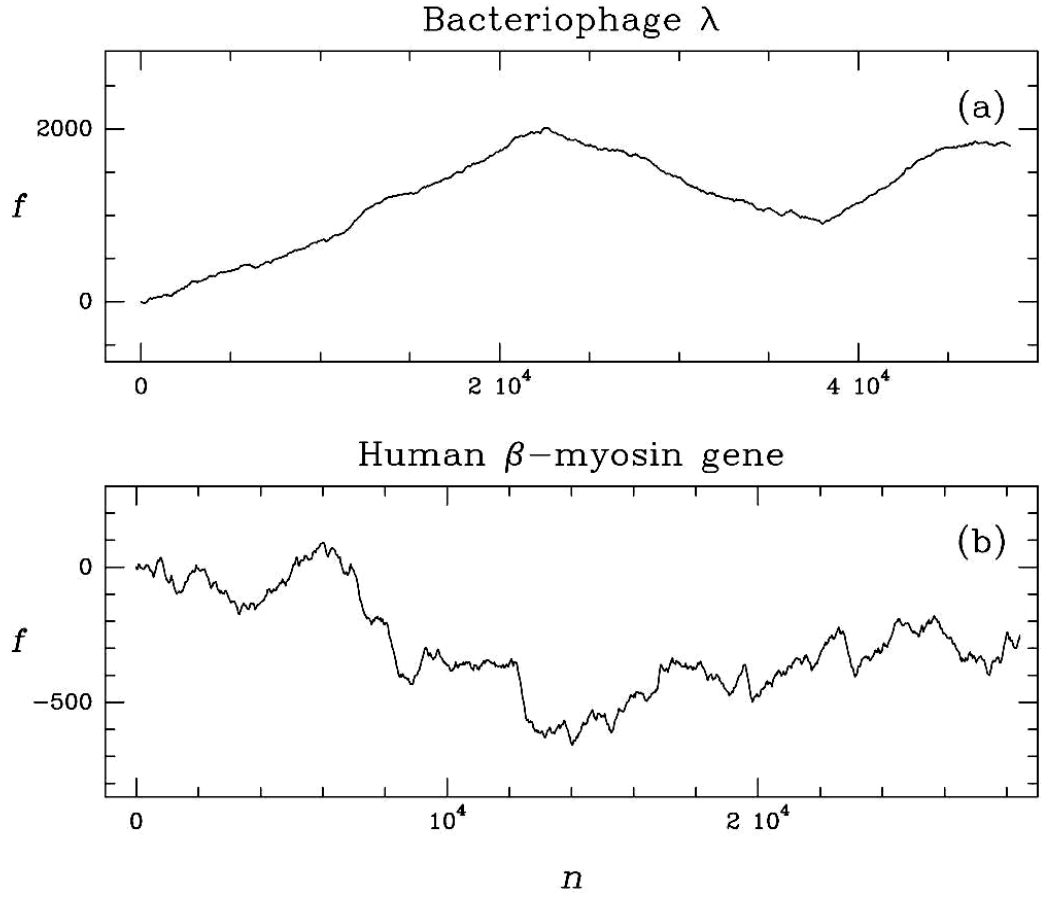
H = 0.9 long-range correlated



Are the observed LRC a bias in the measurement ?

Is the mosaic structure of DNA enough to account for the observed misleading LRC in DNA sequences ?

Karlin and Brendel 93 :

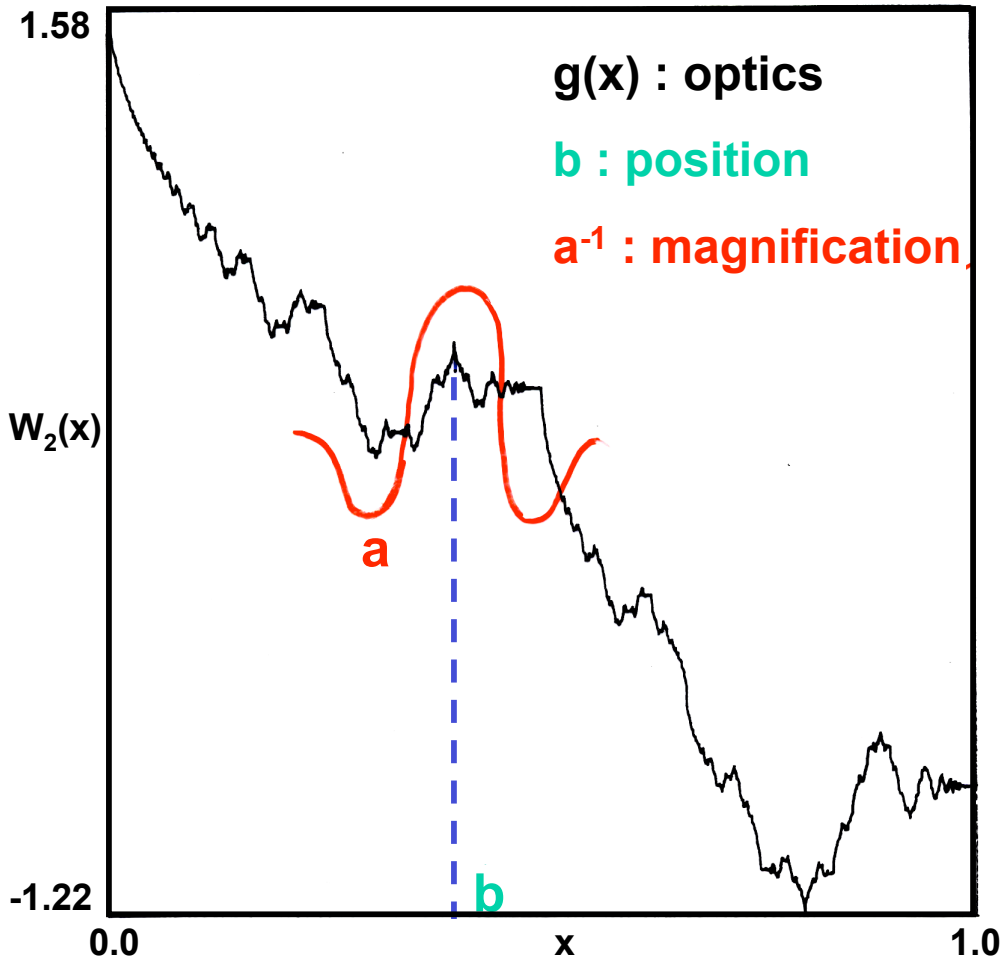


A specific analysing tool is needed to avoid confusing a biased uncorrelated random walk with an unbiased correlated random walk

WAVELET ANALYSIS OF FRACTAL SIGNALS

$$T_g(a,b) = \frac{1}{a} \int g^* \left(\frac{x-b}{a} \right) f(x)\, dx$$

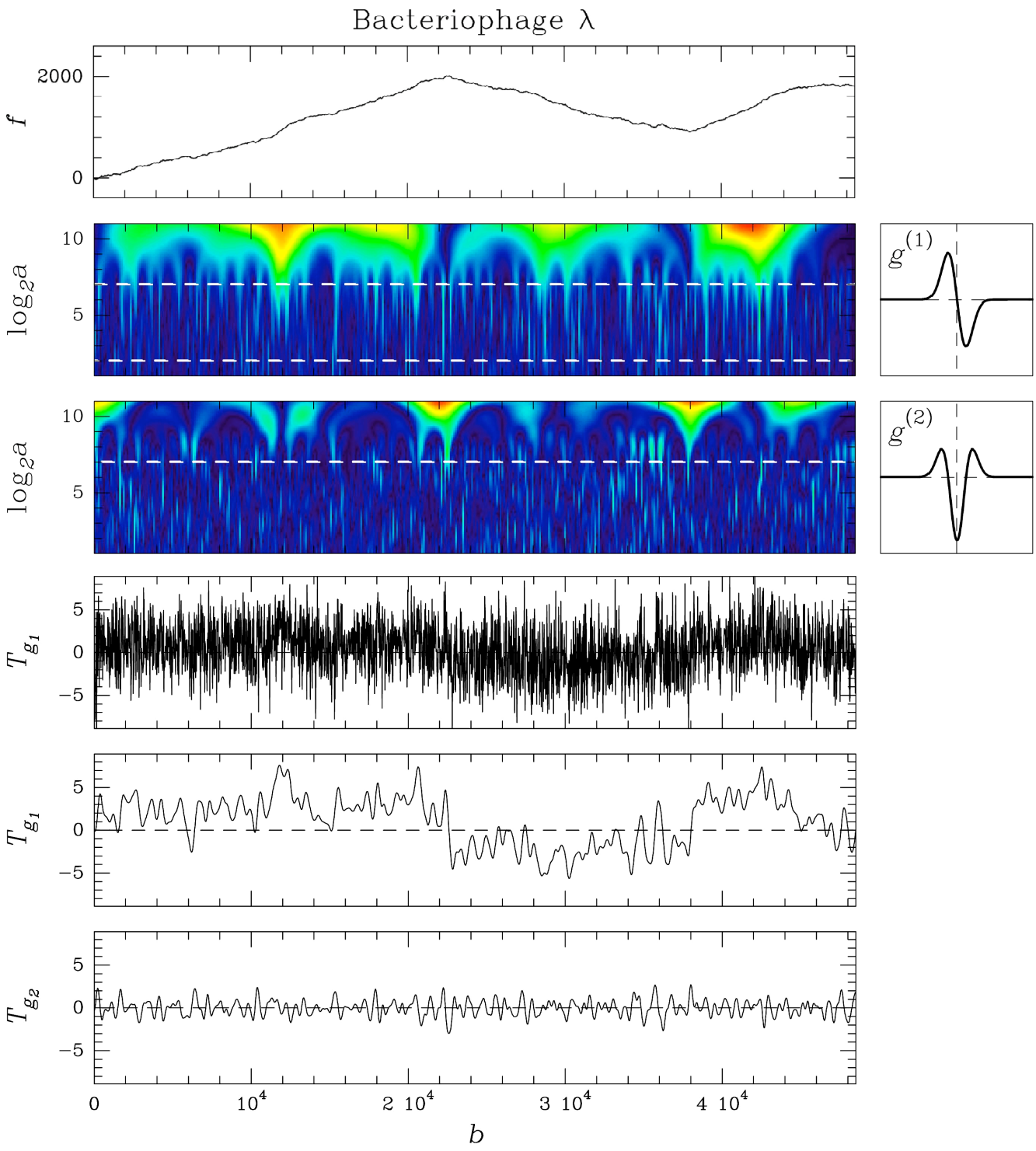
Mathematical microscope



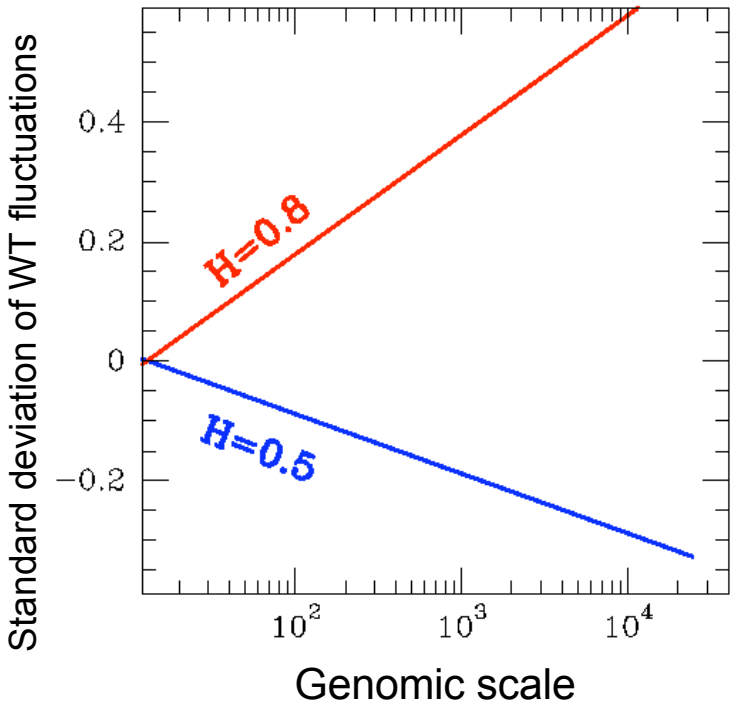
" Singularity scanner"

The wavelet transform allows us to **LOCATE** (**b**) the singularities of f and to **ESTIMATE** (**a**) their strength h(x) (Hölder exponent)

WAVELET ANALYSIS OF THE DNA SEQUENCE OF THE BACTERIOPHAGE λ



A UNIQUE WAY TO DISPLAY RESULTS

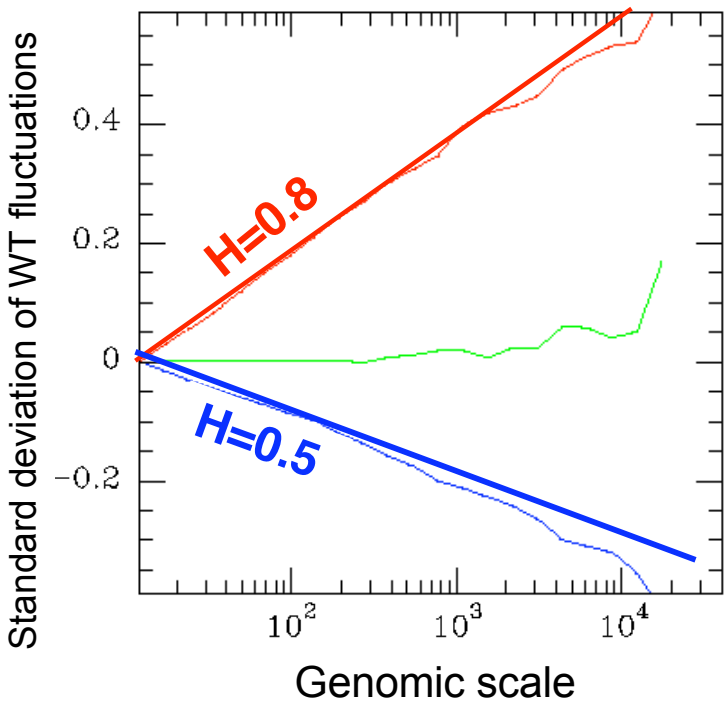


1. Straight line $\Leftrightarrow$ scale invariance properties

2. The slope of a linear behavior gives the roughness exponent H

$$\begin{cases} H = 0.5 & \text{No LRC} \\ H > 0.5 & \text{LRC} \end{cases}$$

A UNIQUE WAY TO DISPLAY RESULTS

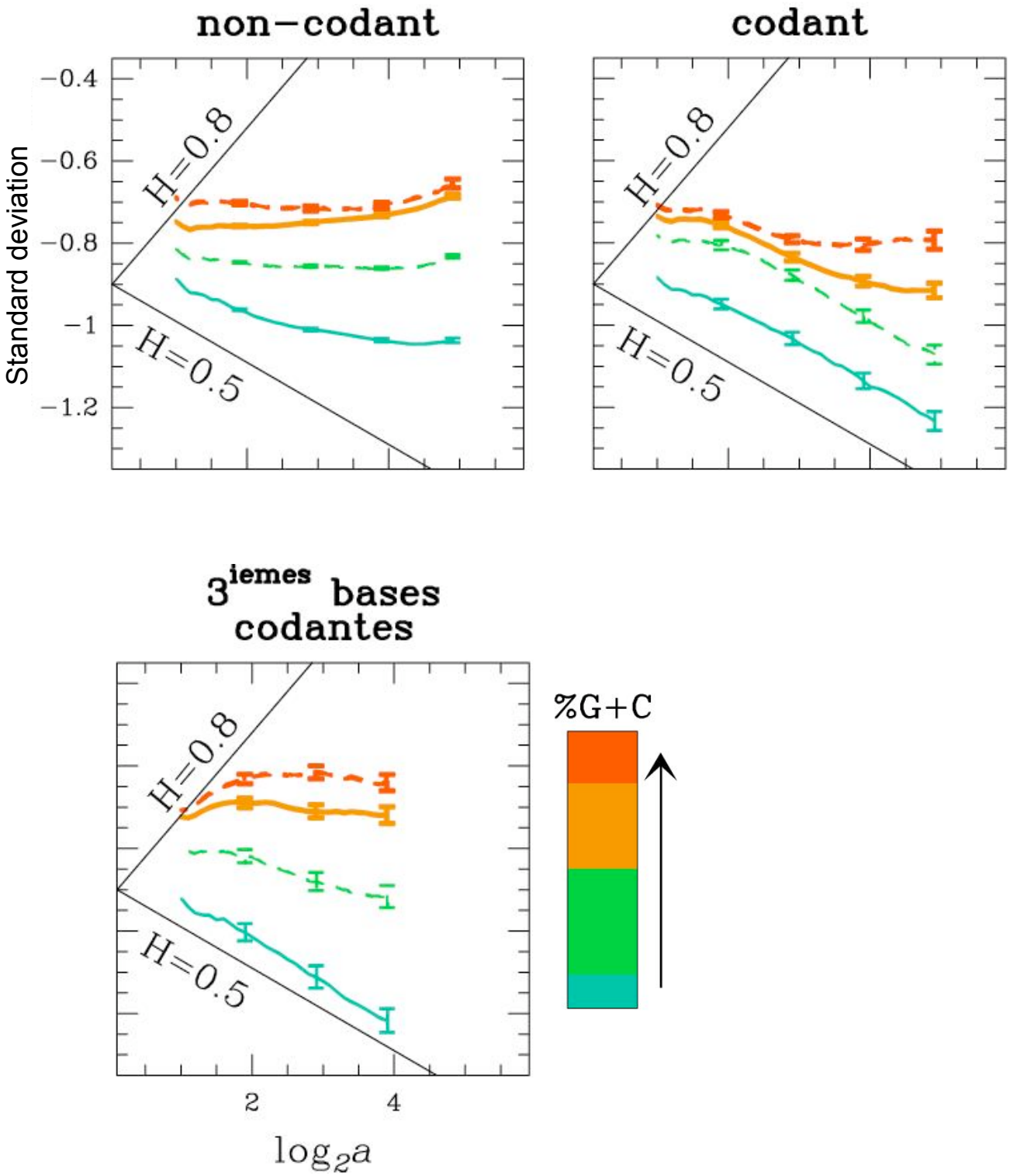


1. Straight line $\Leftrightarrow$ scale invariance properties

2. The slope of a linear behavior gives the roughness exponent H

$$\begin{cases} H = 0.5 & \text{No LRC} \\ H > 0.5 & \text{LRC} \end{cases}$$

Presence of LRC in human coding sequences



WHICH BIOLOGICAL MECANISMS CAN ACCOUNT FOR LRC IN DNA SEQUENCES

- Genomes dynamics and plasticity

  Point mutation

  Insertion, deletion

  Transposition

  Duplication of exons, genes or chromosomes

  Recombinaison

  Generalized Lévy walk model (Buldyrev *et al.* 93)

  Length distribution of protein coding segments (Herzel and Große 97)

- Compaction constraints - Accession to information

  Nucleosome

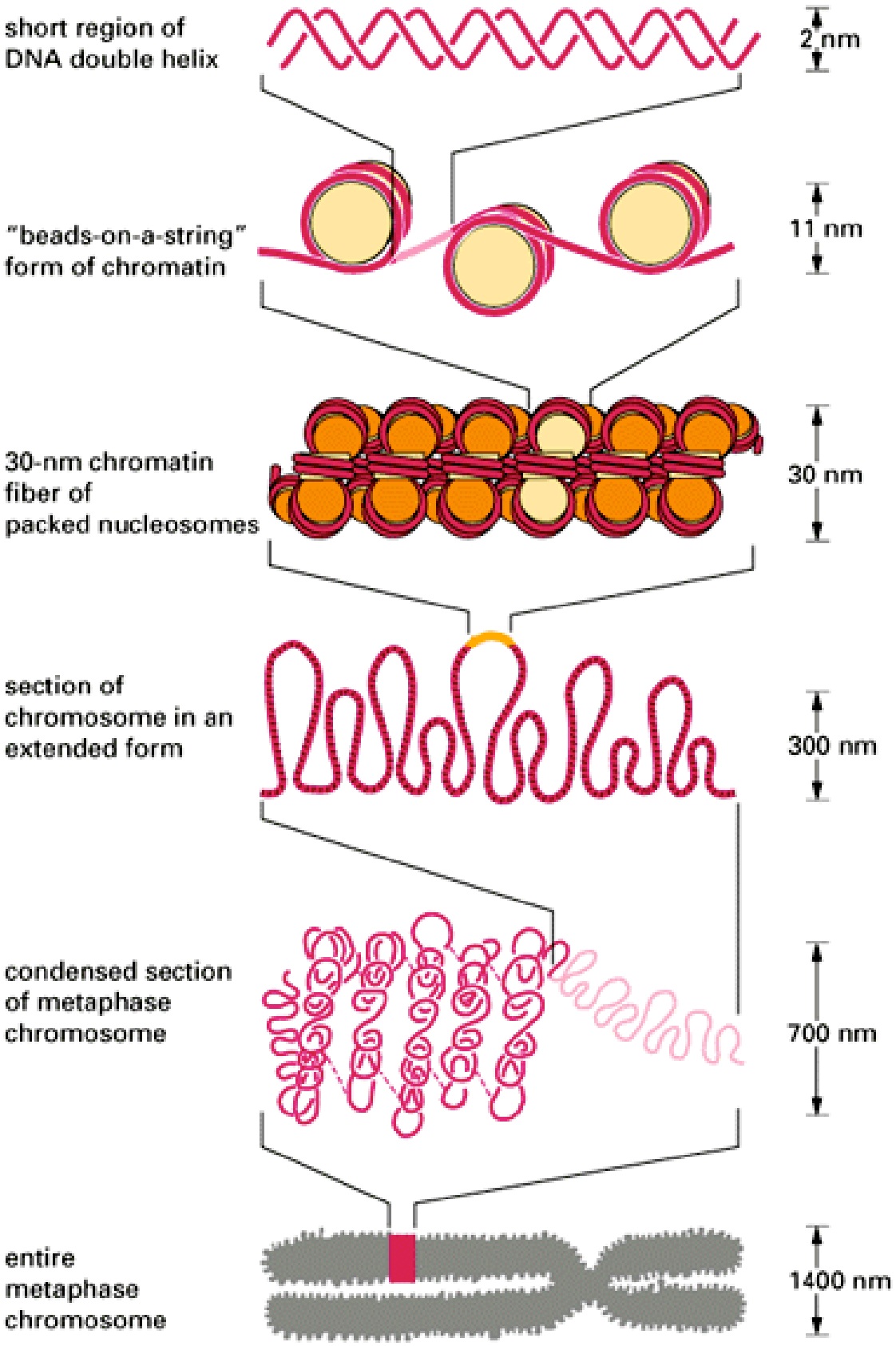
  Chromatine fiber

  Higher order folding up to the metaphase chromosome

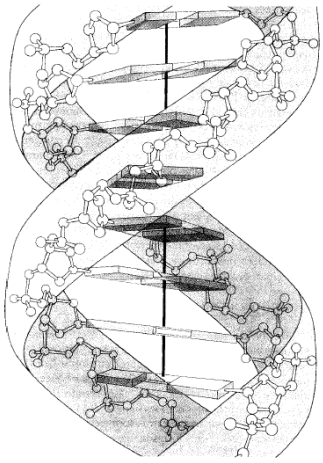
  Fractal model of chromosomes (Takahashi 89)

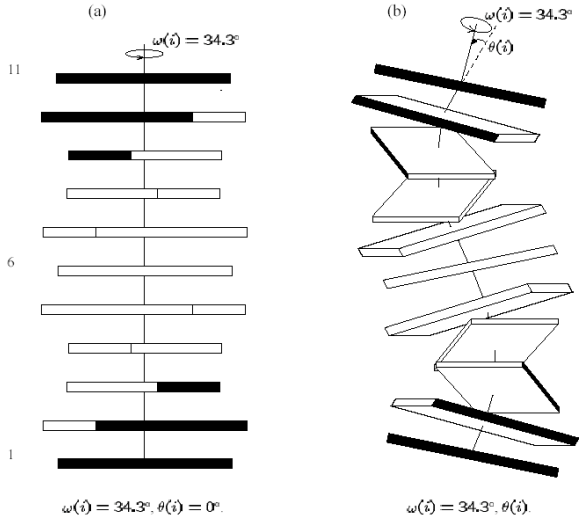
  Crumpled globule model (Grosberg *et al.* 93)

HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA



DNA WALKS THAT REFLECT THE STRUCTURE OF THE DNA POLYMER





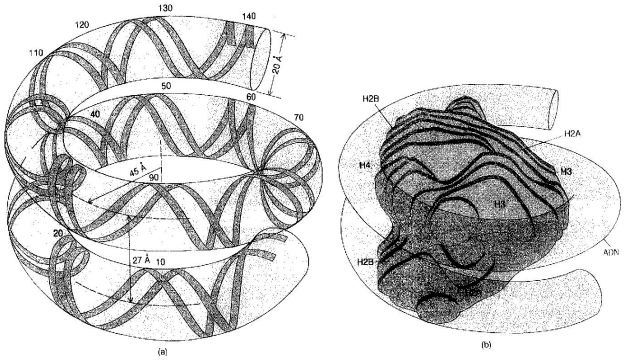
2 trinucleotide codings based on experiments :

Trinucleotide		
	PNuc	DNase I
AAA/TTT	0.0	0.1
AAC/GTT	3.7	1.6
AAG/CTT	5.2	4.2
AAT/ATT	0.7	0.0
ACA/TGT	5.2	5.8
ACC/GGT	5.4	5.2
ACG/CGT	5.4	5.2
ACT/AGT	5.8	2.0
AGA/TCT	3.3	6.5
AGC/GCT	7.5	6.3
AGG/CCT	5.4	4.7
ATA/TAT	2.8	9.7
ATC/GAT	5.3	3.6
ATG/CAT	6.7	8.7
CAA/TTG	3.3	6.2
CAC/GTG	6.5	6.8

Trinucleotide		
	PNuc	DNase I
CAG/CTG	4.2	9.6
CCA/TGG	5.4	0.7
CCC/GGG	6.0	5.7
CCG/CGG	4.7	3.0
CGA/TCG	8.3	5.8
CGC/GCG	7.5	4.3
CTA/TAG	2.2	7.8
CTC/GAG	5.4	6.6
GAA/TTC	3.0	5.1
GAC/GTC	5.4	5.6
GCA/TGC	6.0	7.5
GCC/GGC	10.0	8.2
GGA/TCC	3.8	6.2
GTA/TAC	3.7	6.4
TAA/TTA	2.0	7.3
TCA/TGA	5.4	10.0

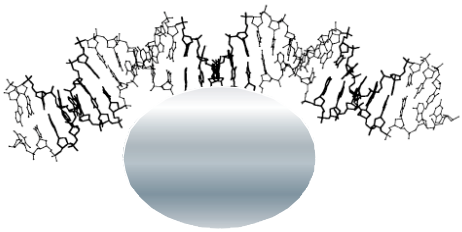
Nucleosome positioning

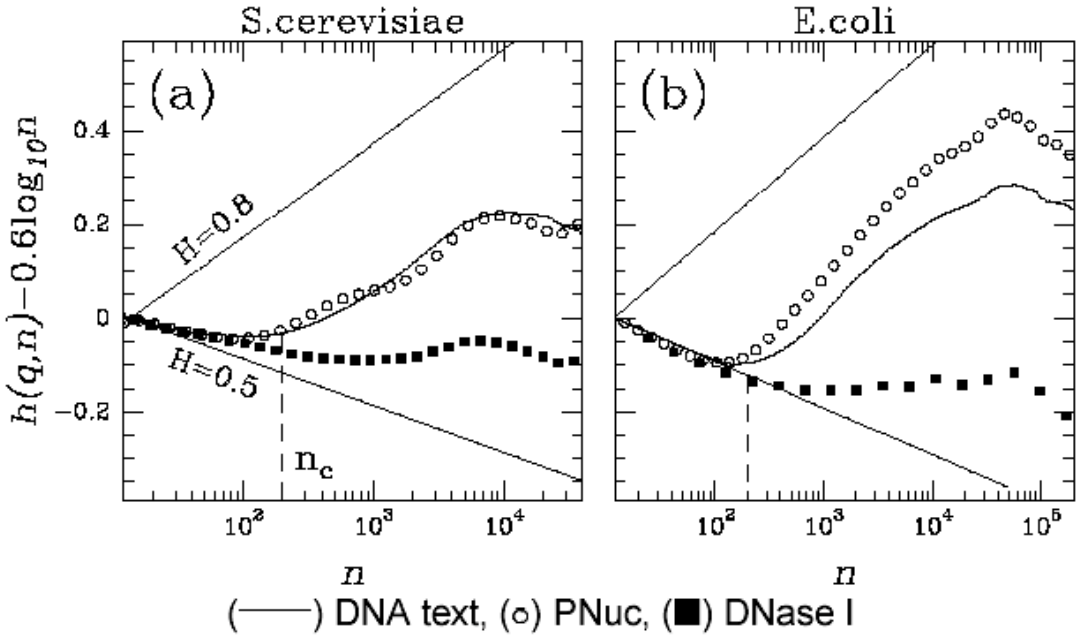
local curvature



Dnase I sensitivity

Local flexibility

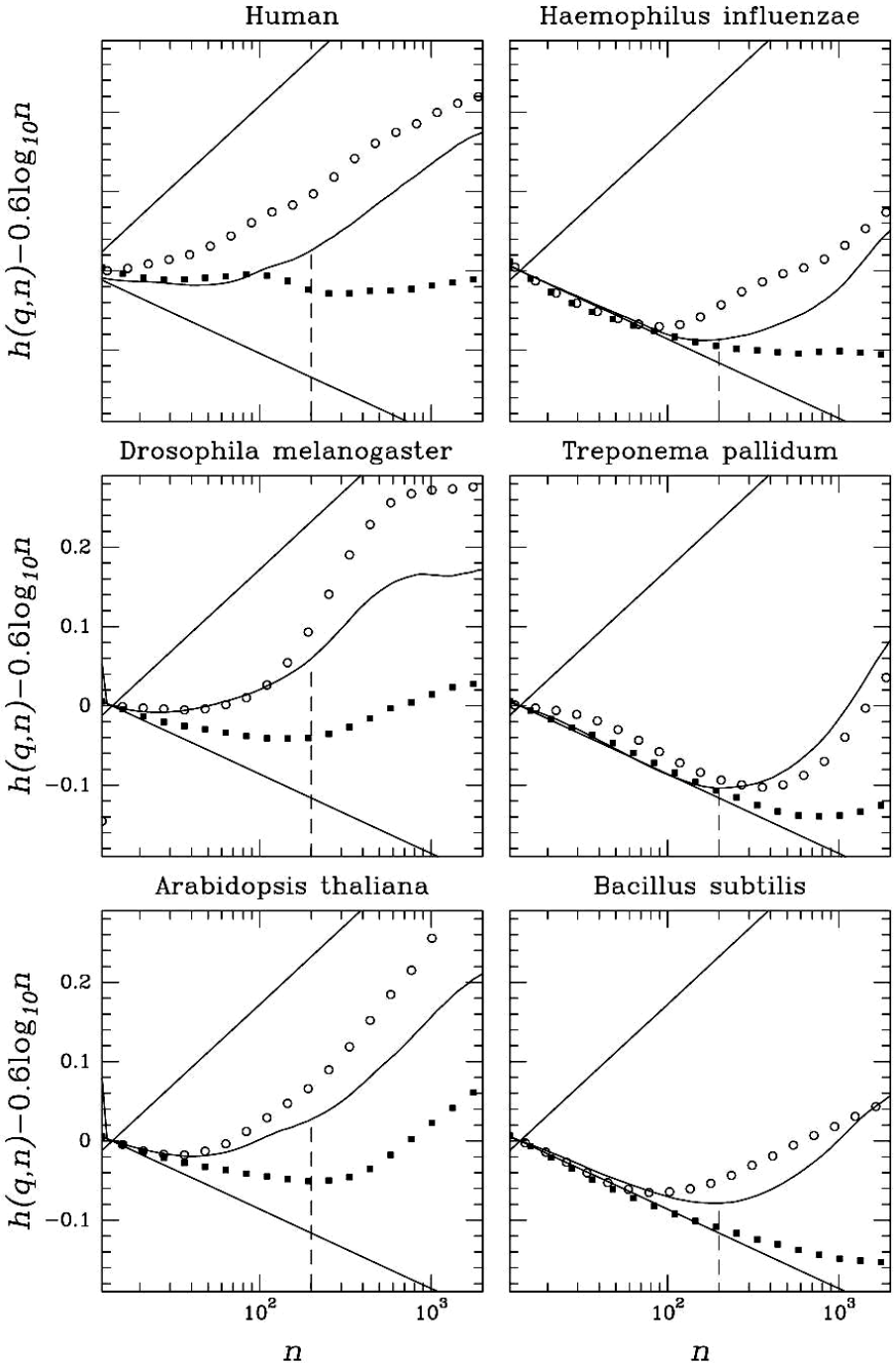




(——) DNA text, (○) PNuc, (■) DNase I

Hypothesis: LRC in the small scales regime is the signature of of the nucleosomal structure

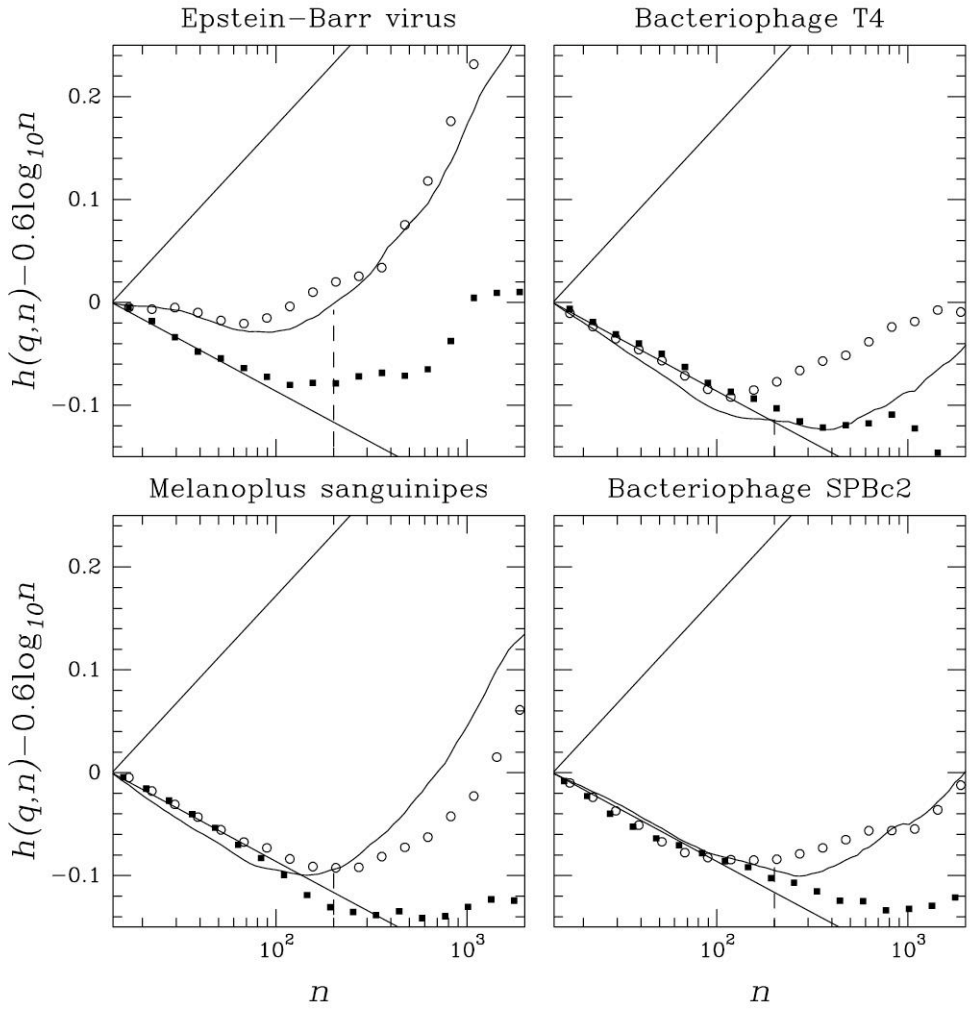
Eucaryotes Bacteria



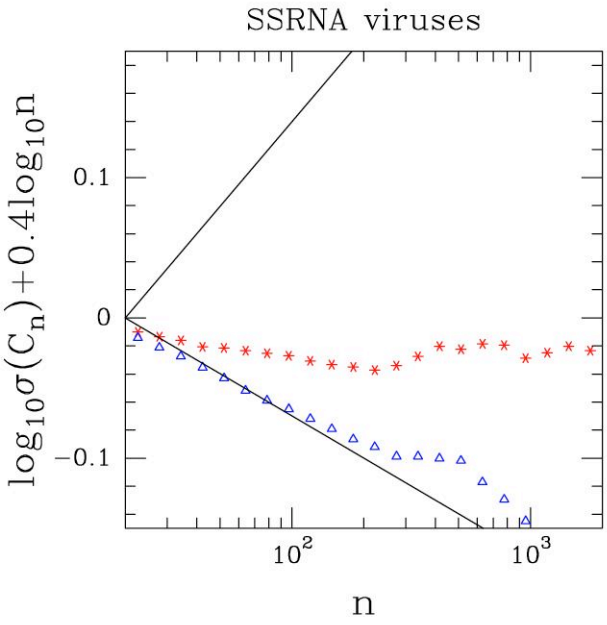
(——) DNA text, (○) PNuc, (■) DNase I

Nucleosomes No nucleosomes

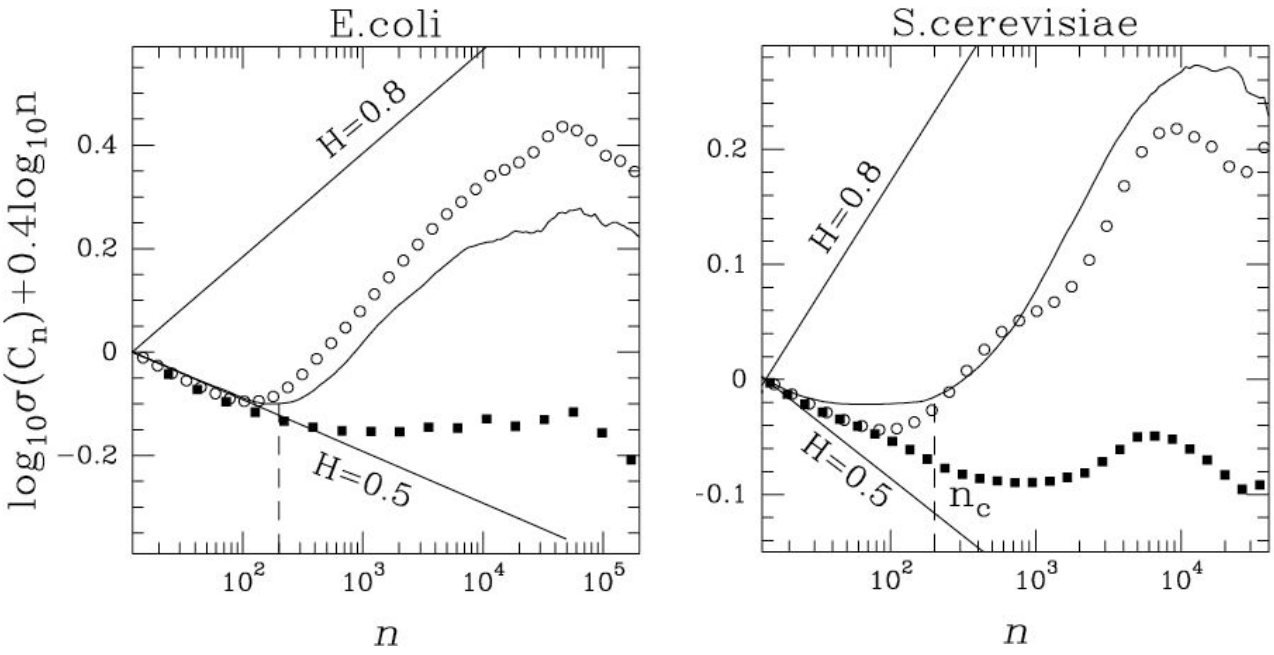
SMALL SCALES LRC ARE RELATED TO NUCLEOSOME LIKE STRUCTURES



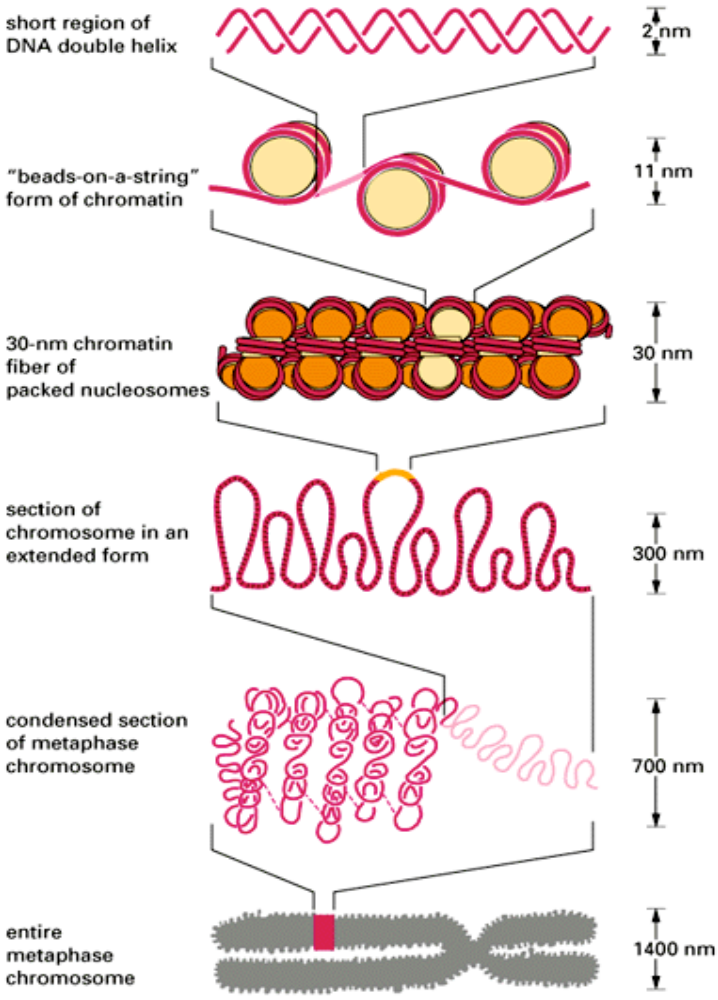
Pox virus don't display LRC in the small scale regime



Among the SSRNA viruses only retroviruses display LRC in the small scale regime

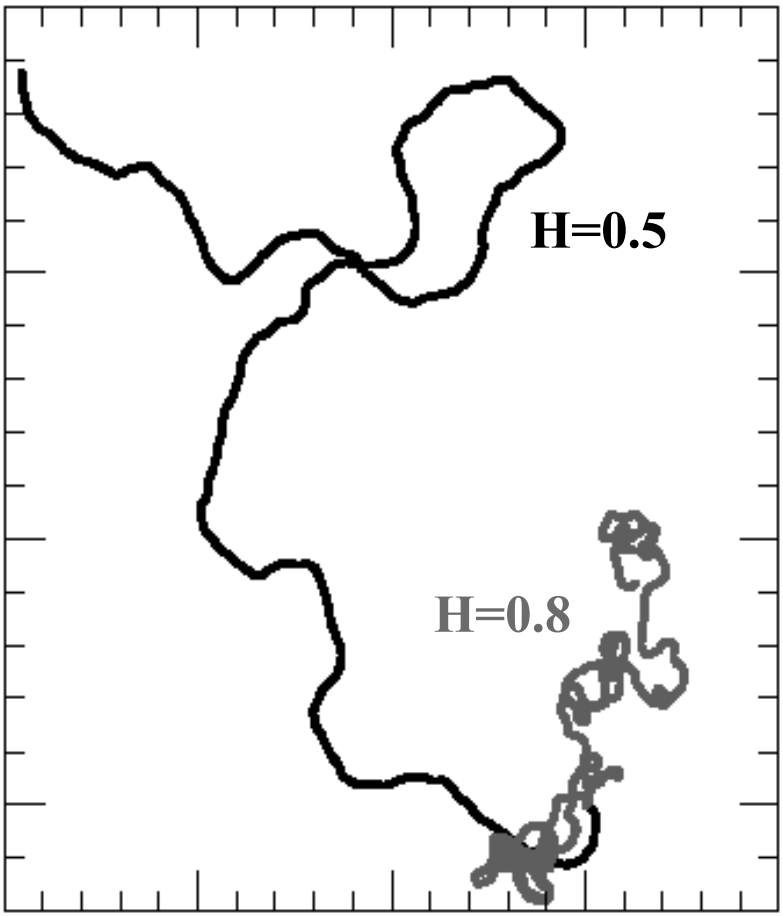


HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA



Influence of the DNA sequence on the formation and dynamics of nucleosomes

Uncorrelated DNA



Long-range correlated DNA

L_{DNA} = 3000 bp