The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2585-1

Joint ICTP-TWAS School on Coherent State Transforms, Time-Frequency and Time-Scale Analysis, Applications

2 - 20 June 2014

What can we learn on genome structure and function from a wavelet-based multi-scale analysis of genomic and epigenetic data

A. Arneodo
ENS, Lyon
France

What can we learn on genome structure and function from a wavelet-based multi-scale analysis of genomic and epigenetic data ?

Alain Arneodo

Laboratoire de Physique,
Ecole Normale Supérieure de Lyon

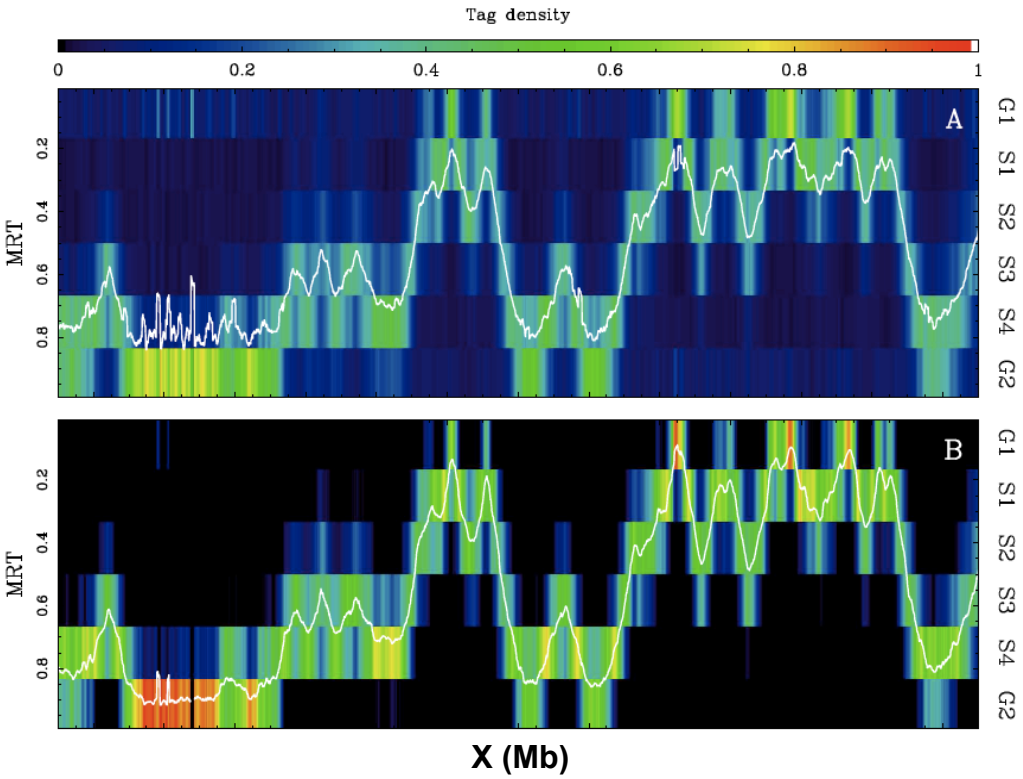
Alain.Arneodo@ens-lyon.fr

DNA sequencing projects result in 4 letter texts

```
gtcagttcctgaggcgggtcgggacccaggcgtgagactggagtctgcc
caggggcccagctgagccagcctcctcgtcagctgcttgggccgccagga
cgccgccgggggtgcgccgcgcttccctggatggggtgccccccactcccc
tcggagccccagggagacccccccgaactcagctcctctcaggggtgccag
ggggacccctcaaactccactccccgcaggttcctggggagacgcccccct
gctcgattcccctcagggtcccagggagacccccctaattcagctcctctc
aggggtactgggggacctctcgagctccactcccatcagggtcccaggga
```

Computing Mean Replication Timing profiles from RepliSeq data

(data from Chen, Genome Research (2010) and Hansen, PNAS (2010))



Long-range correlations in eukaryotic DNA

A footprint of nucleosome packaging

Alain Arneodo

Laboratoire de Physique
Ecole Normale Supérieure de Lyon
46 allée d'Italie, 69364 Lyon Cedex 07, FRANCE

Françoise Argoul
Benjamin Audit
Julien Moukhtar — ENS de Lyon, France
Jean-François Muzy
Cédric Vaillant

Yves d'Aubenton-Carafa — CGM, Gif-sur-Yvette, France
Claude Thermes

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

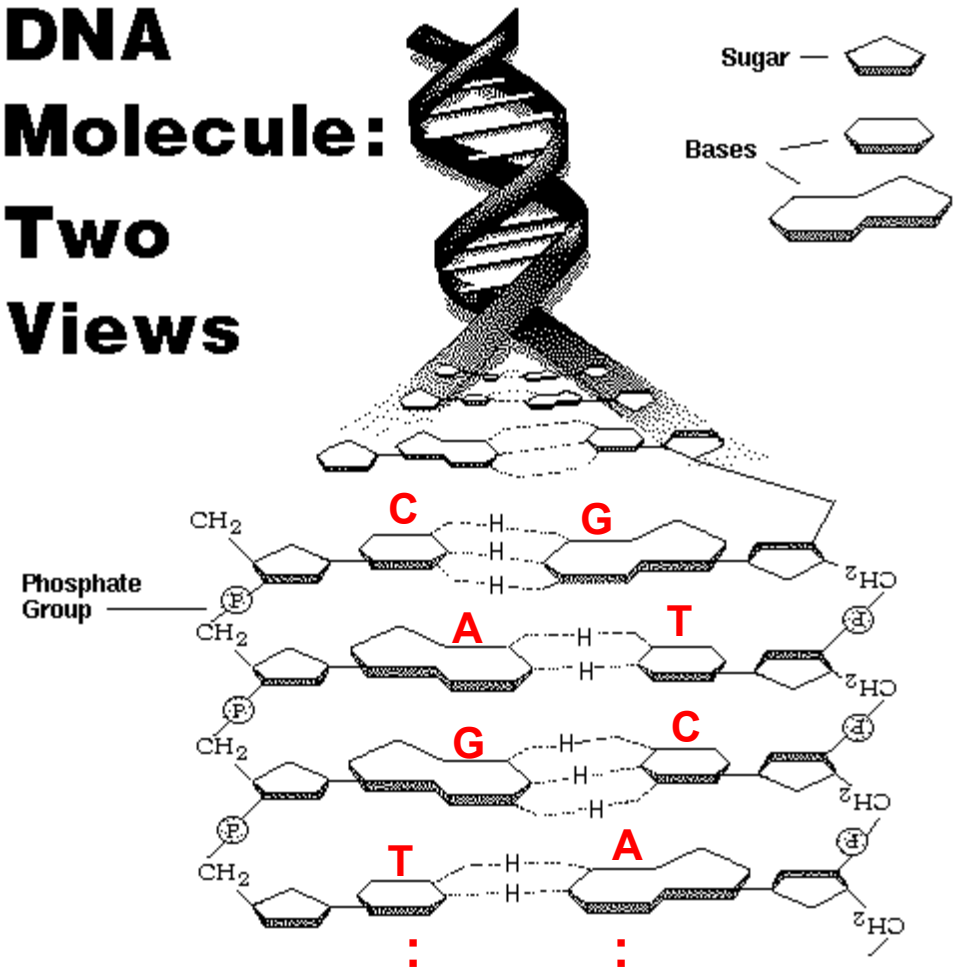
$\rightarrow$ a simple model for the transmission of the genetic information

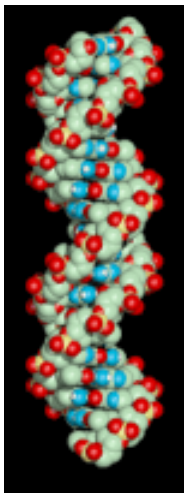
1966 Niremberg, Ochoa and Khorana elucidate the genetic code

$\rightarrow$ DNA codes for proteins

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

DeoxyriboNucleic Acid



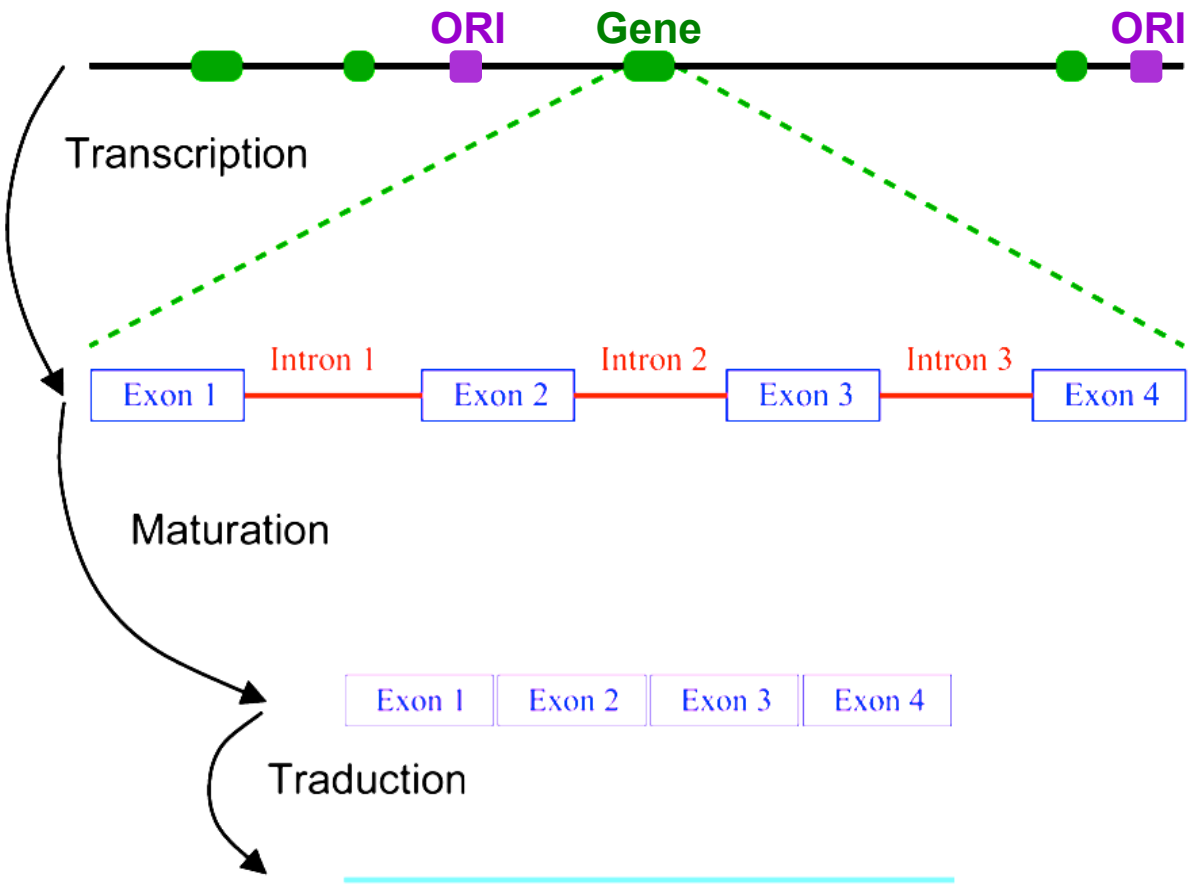


- Double helix macromolecule
- Each strand consists of an oriented sequence of four possible nucleotides:
  Adenine, Thymine, Guanine & Cytosine
- Complementary strands:
  [A]=[T] & [G]=[C] over the sum of both strands

Sequencing projects result in 4 letter texts :

gtcagttcctgaggcgggtcgggacccaggcgtgagactggagtctgcc
caggggcccagctgagccagcctcctcgtcagctgcttgggccgccagga
cgccgccgggggtgcgccgcgcttccctggatggggtgcccccactcccc
tcggagccccagggagacccccgaactcagctcctctcaggggtgccag
ggggacccctcaaactccactccccgcaggttcctggggagacgccccct
gctcgattcccctcagggtcccagggagacccccctaattcagctcctctc
aggggtactgggggacctctcgagctccactcccatcagggtcccaggga
gacccccccaactatgctcaggggtcccagggagatgccagcaccccaact
ccgcttccctggggcccccctcccccttacagctcaacttccctcgagagt
ctggggctggggctccgttcagttcttgagtcccccttccctcggggtgtc
ccggggccgcccacccccacactgtctgtgattccccaaggcgcgggtct
cgggccgcagcctgttccacgttctgctgctcgttctttttctggctcctt
gctttcgaaggagagaaggaggccttcgtttccagtctttttgccttttc
taatggagccctgcttttccttccgtgtcccttcaggctacttctgccag
gtttctatttttcattctttattatgacttcgcccaaaatattcttgact
tctattgagaaggattcgggggtctatttcttattcggaggcgtgtgctt
aagttccaaacagatgaggattttccagttaatccttctggggtgactta
ttgcttaatgccaccatagccagaaaatggactctcagtgtccgaaactg
cattcggctctgaagtgtctgtccttgtcacctcttgcaatgtttcgcgg
cgggaagcctgcactcgccgacgctgacgtaactgtttctgtctttcagg
tctacagcctcctgtgggtgggcgatattgacatatactttatttctata
tatgttatgaactcaatatttcttgcagcgggtctgctgataataagata
tgcctactctgcgagtctggaagccatcttaagcttaccctgtatgtgcc
ccatgcatctcttccgttacacggctcctgagttgacacctgtgtgataa
actggtaatagcaagtaaactgttttcttgtgctctgtaagctgctctag
caaattatctaggaggaggtggtcttggaaaccccctgatttataagcggg
cagtcagcagtacacgtggcccagaatcgtgattggcatttgaagtgggg
gcagtagggtgggactgagcccttcacctgtggggtctgccctgctcaag
gcagtgtcagaattgaagtgaaatgttggacggtcggtgtccagagagtt
ggagaactggtttgtgtgtaaaaactnacatatttagggtcagaagtatg
...

ORGANIZATION OF THE HUMAN GENOME

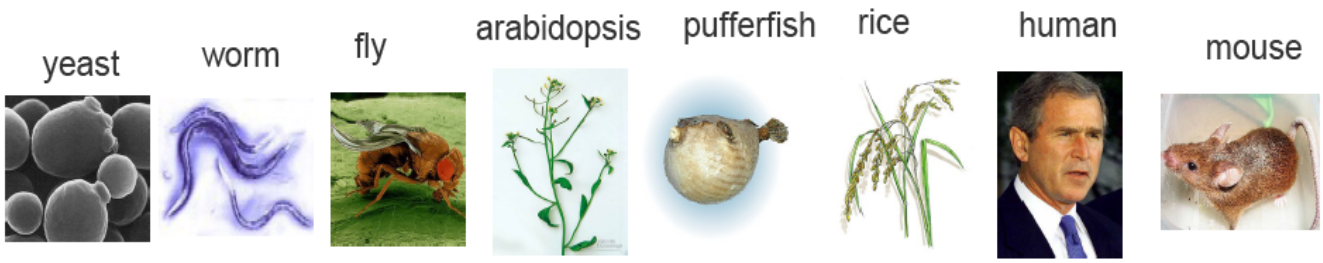


23 Chromosomes
$L \sim$ 100Mbp.

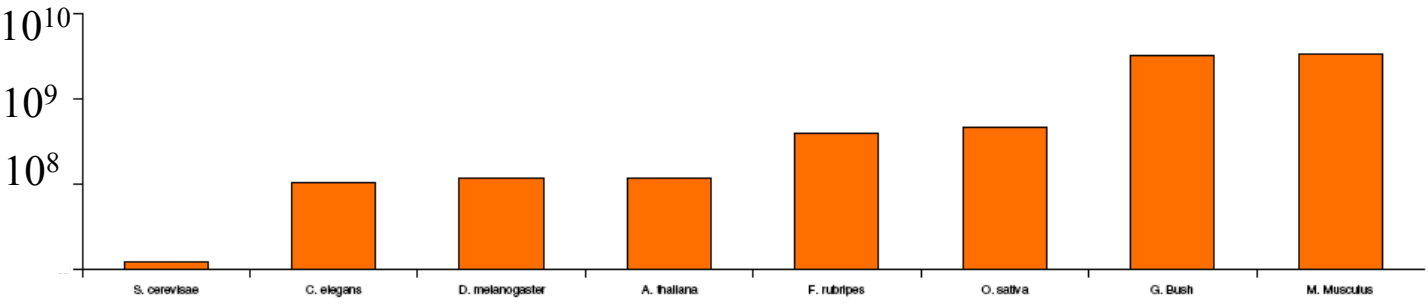
- Genes ($\sim$ 20%)
  $L \sim$ 10kbp.
  - Introns
    (INTervening seq.)
    $L \sim$ 1kbp.
  - Exons
    (EXpressed seq.) $\Rightarrow$ Proteins
    $L \sim$ 150bp. $L \sim$ 500AA.
- Non genic DNA

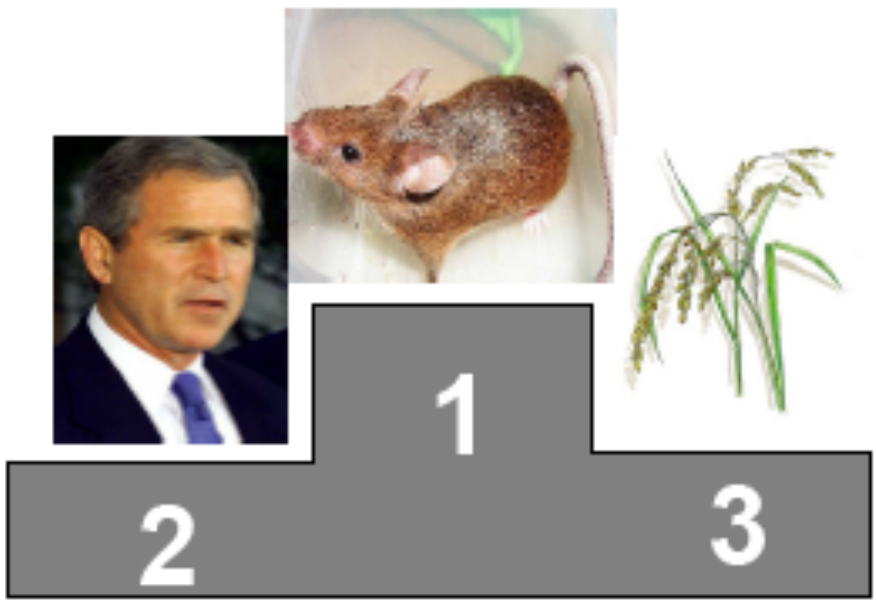
Eukaryotic genome context

(C. Hermann)



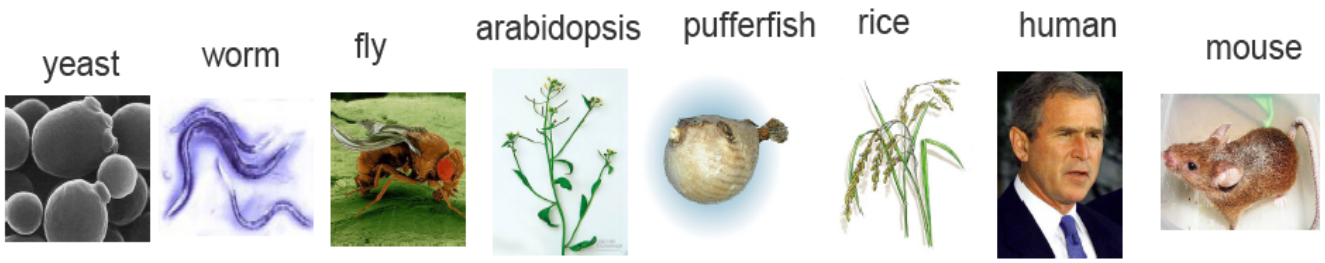
Genome size



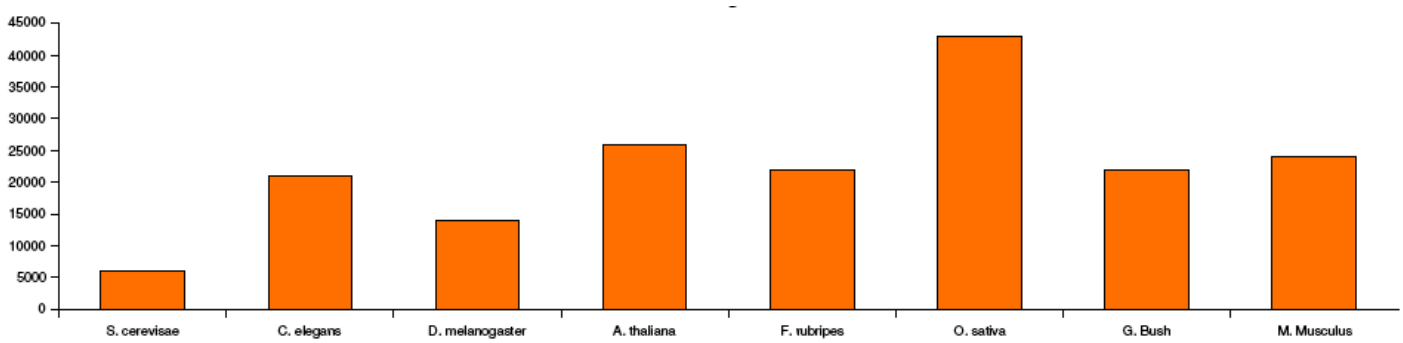


Eukaryotic genome context

(C. Hermann)

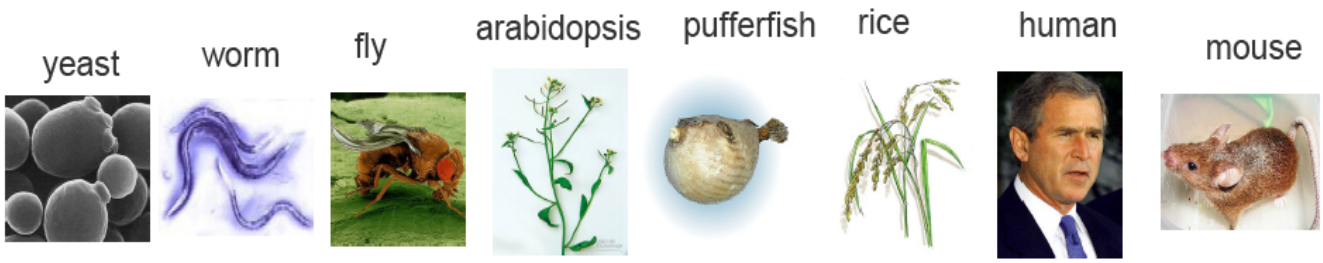


Number of genes

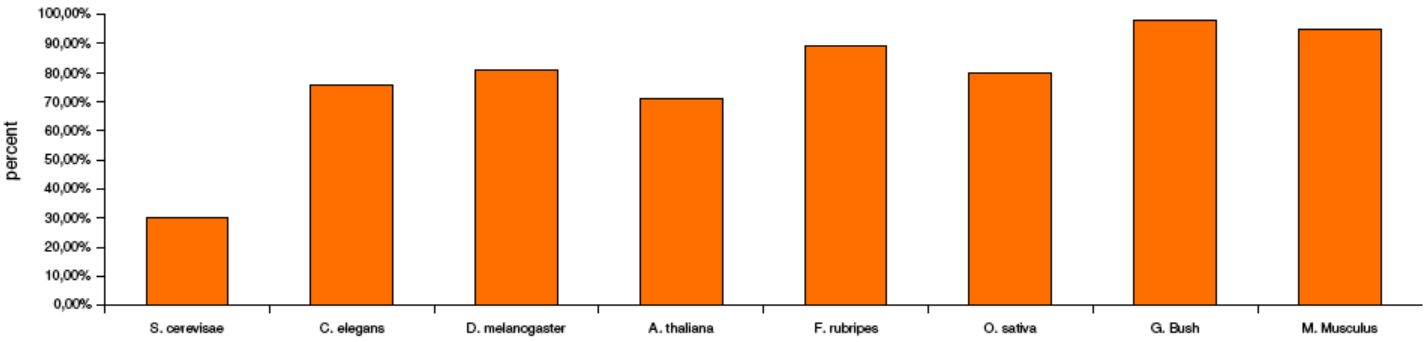


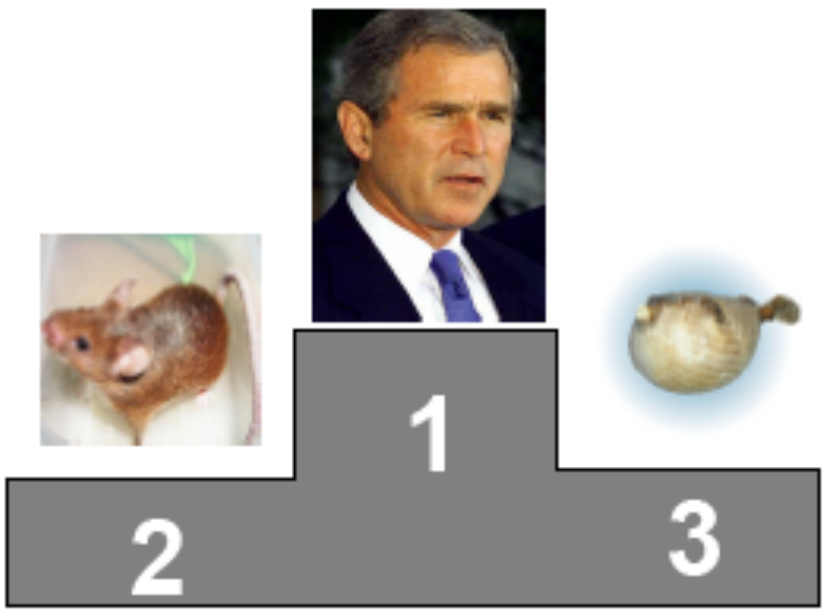
Eukaryotic genome context

(C. Hermann)

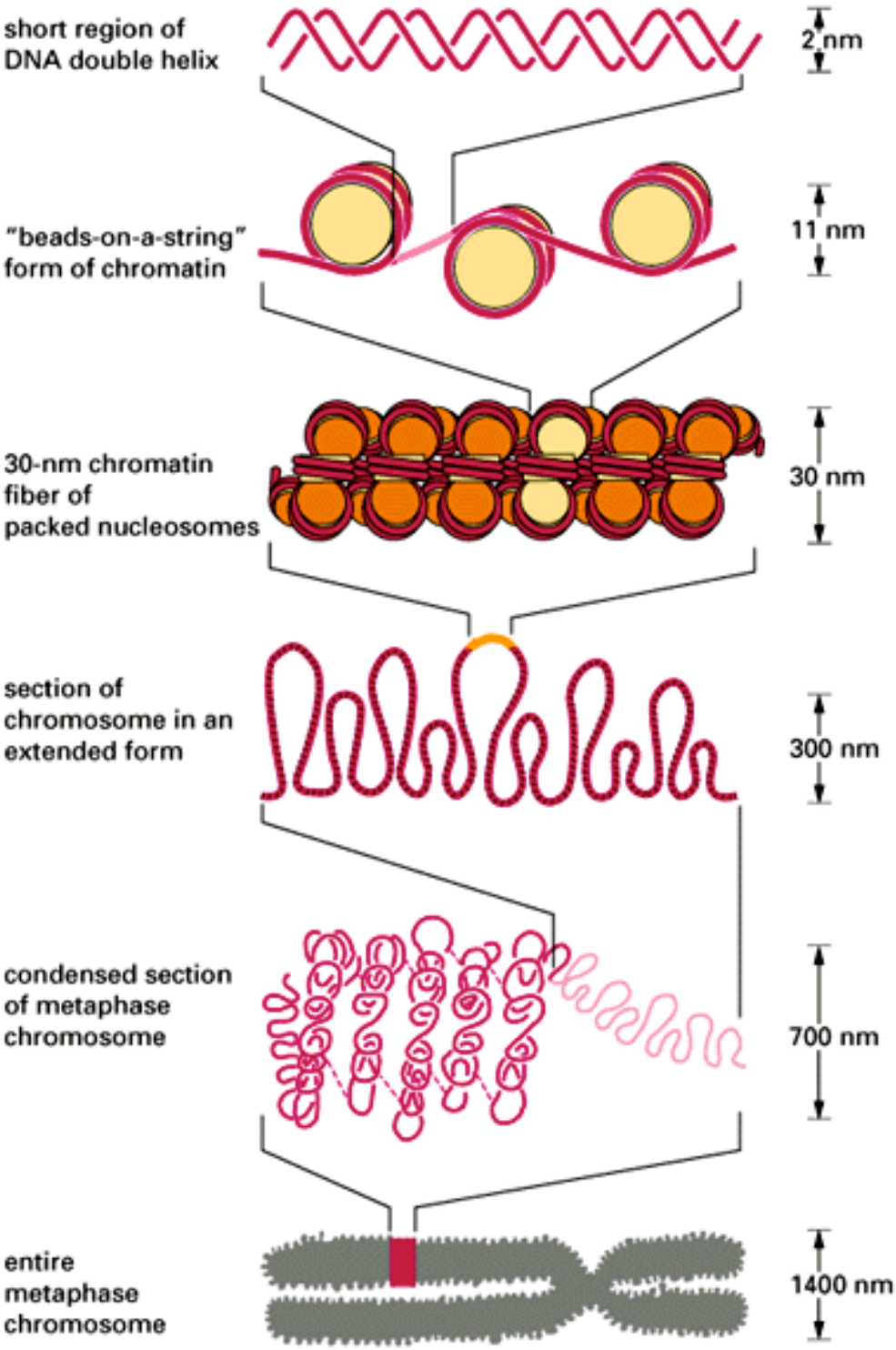


Non-coding DNA



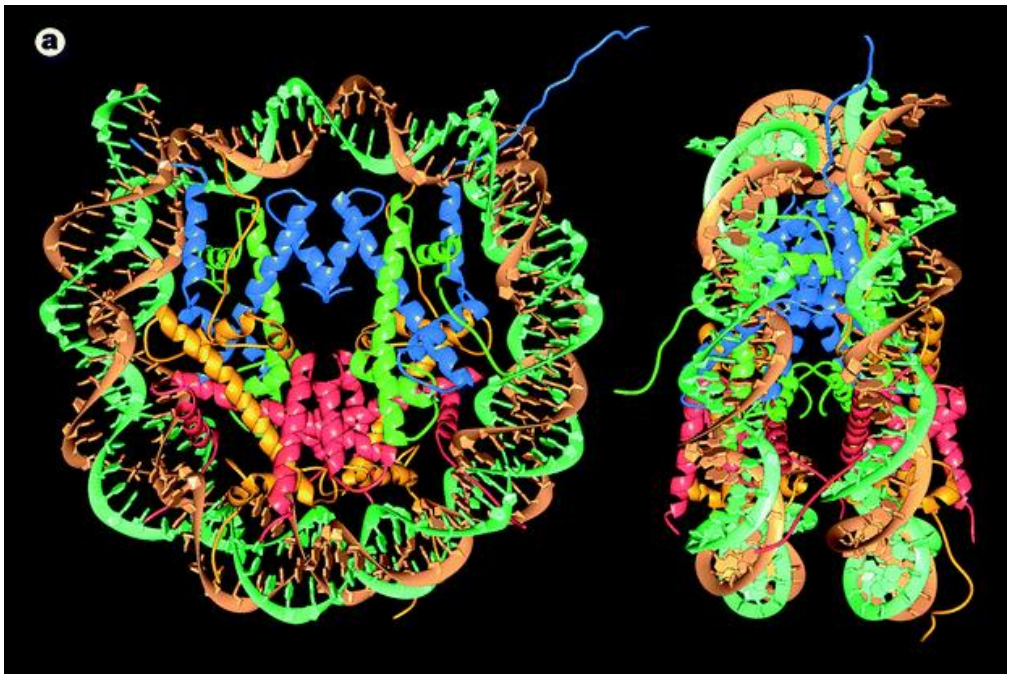


HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA



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

Long-range correlations in genomic DNA: A signature of the nucleosomal structure



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 ?

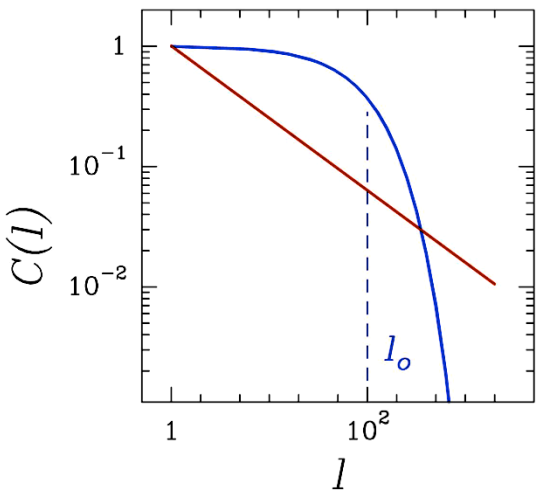
- Question: The choice of the next nucleotide :

  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)

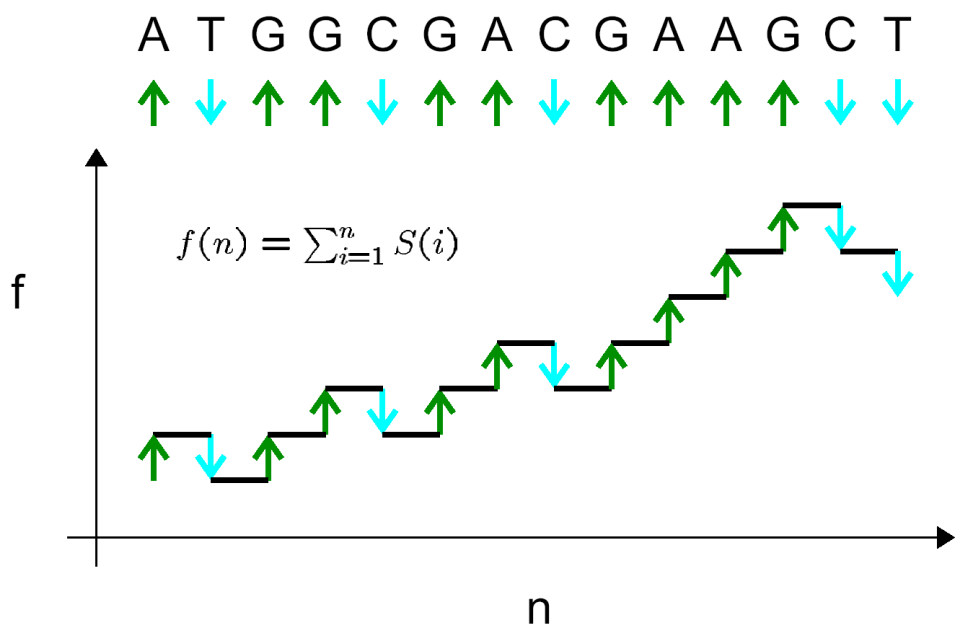
1. Each nucleotide is associated to a numerical value
   (A to a, T to t, G to g and C to c).

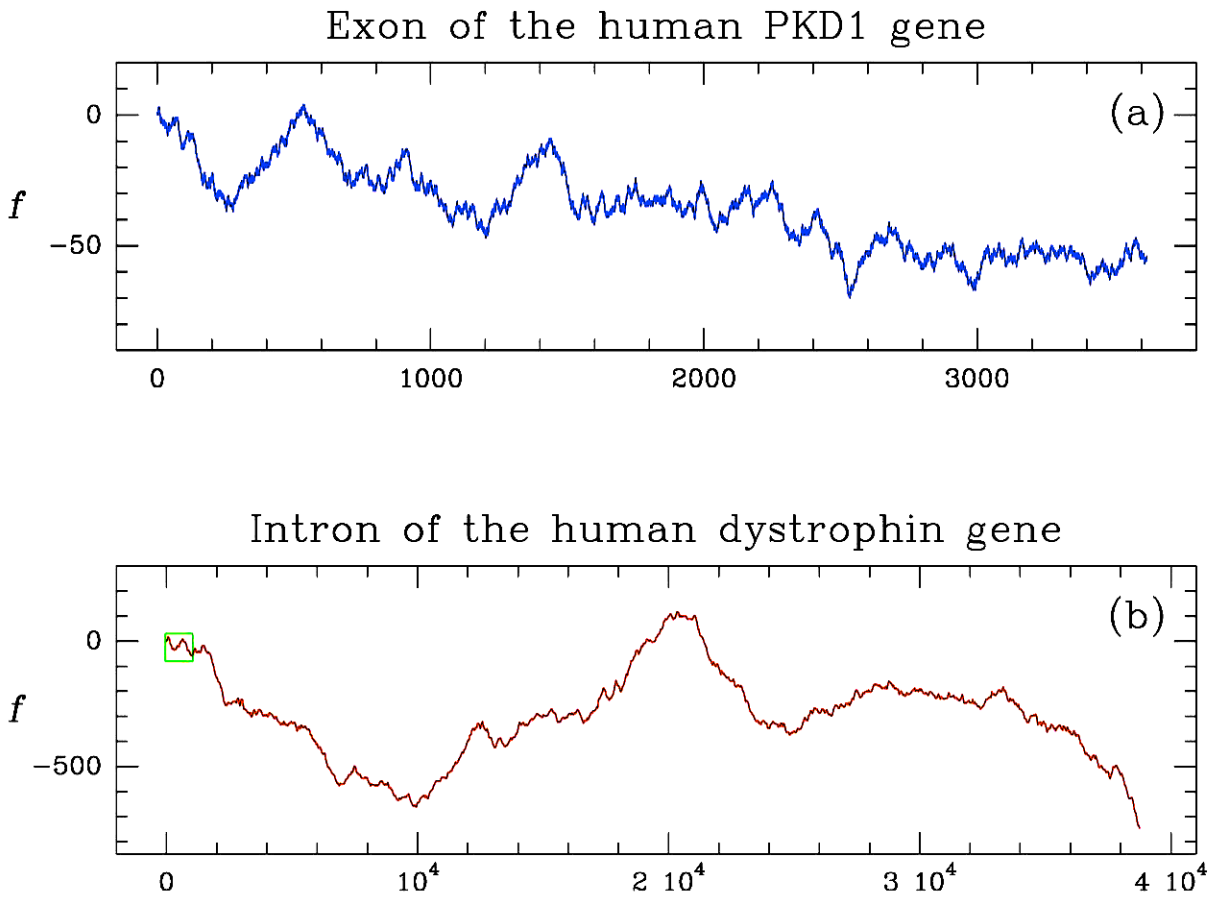
   purine-pyrimidine : $a = g = 1$ and $t = c = -1$
   weak-strong : $a = t = 1$ and $g = c = -1$
   amino-keto : $a = c = 1$ and $t = g = -1$

   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 $a = t = g = -1/3$

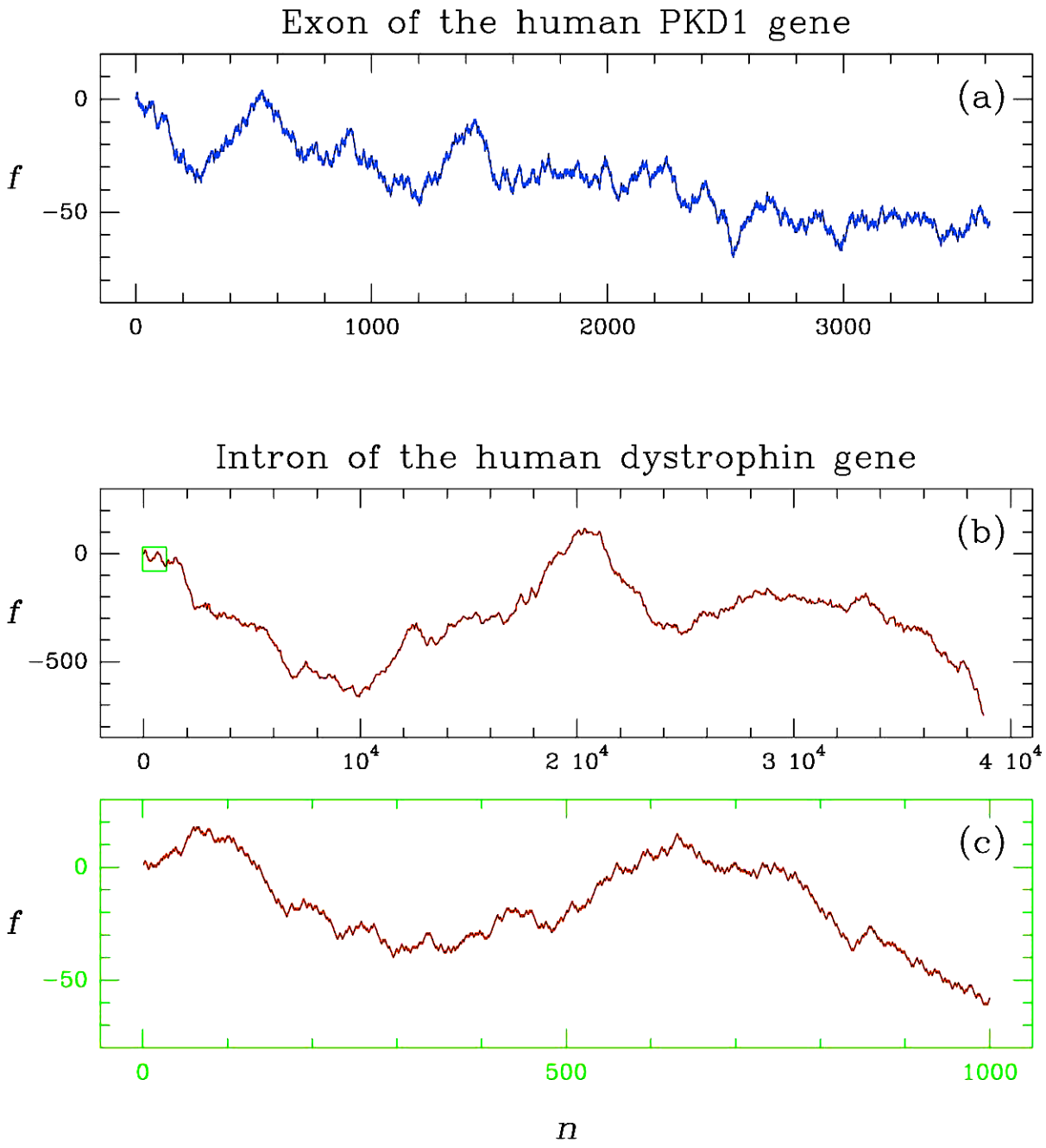
2. Suppose you have a walker on the line. The value associated to the i^{th} nucleotide defines the i^{th} step $S(i)$ of the walker

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





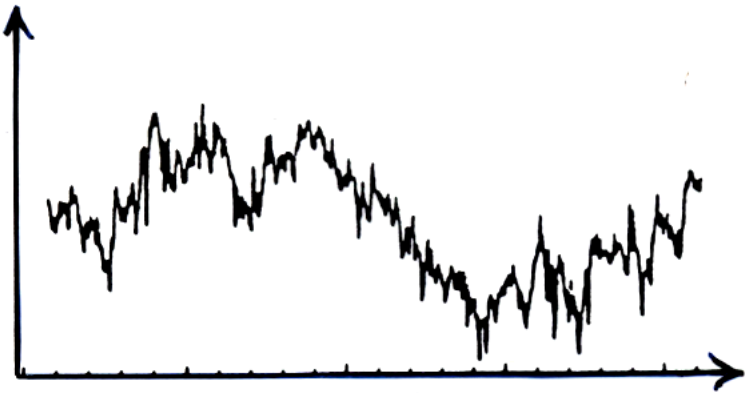
Most of the physicist works amount to characterizing the roughness of a DNA walk landscape



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

FRACTAL SIGNALS

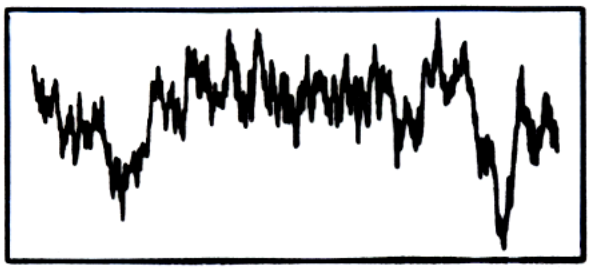
V(t)



time

Turbulent velocity signal

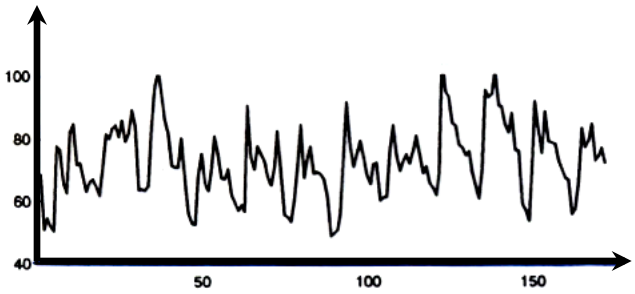
S(t)



time

Brownian signal
‘‘ 1/f noise’’

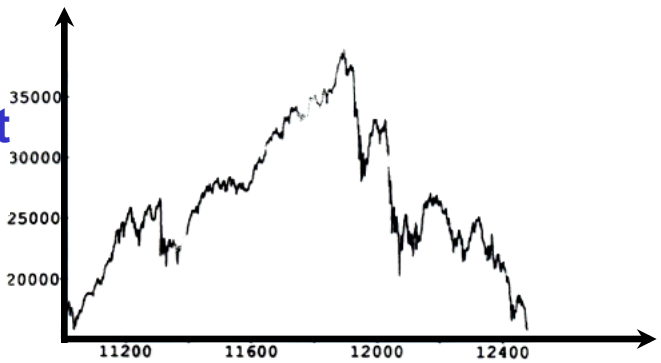
Heart rate



time

Medical signal

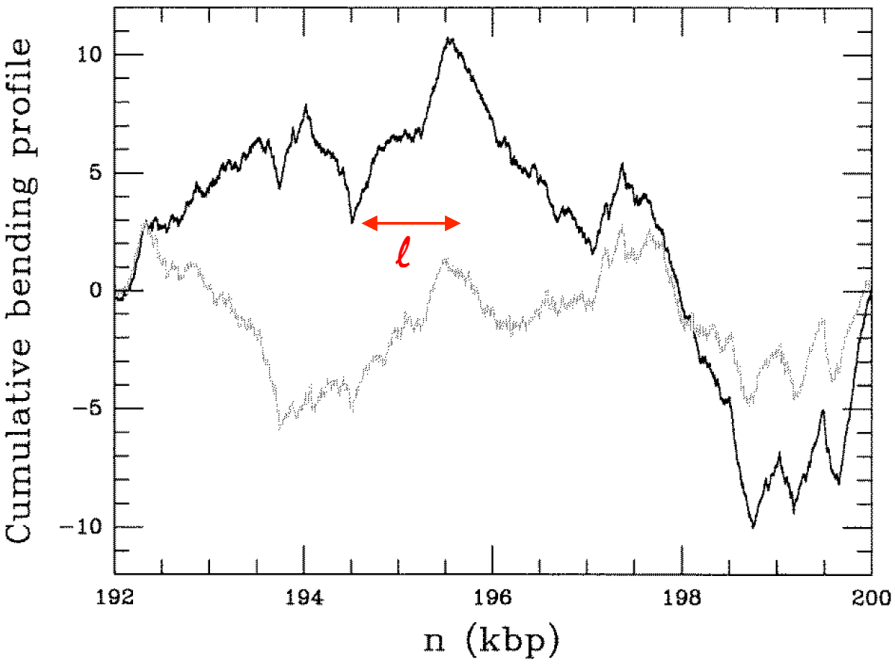
Market prices



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

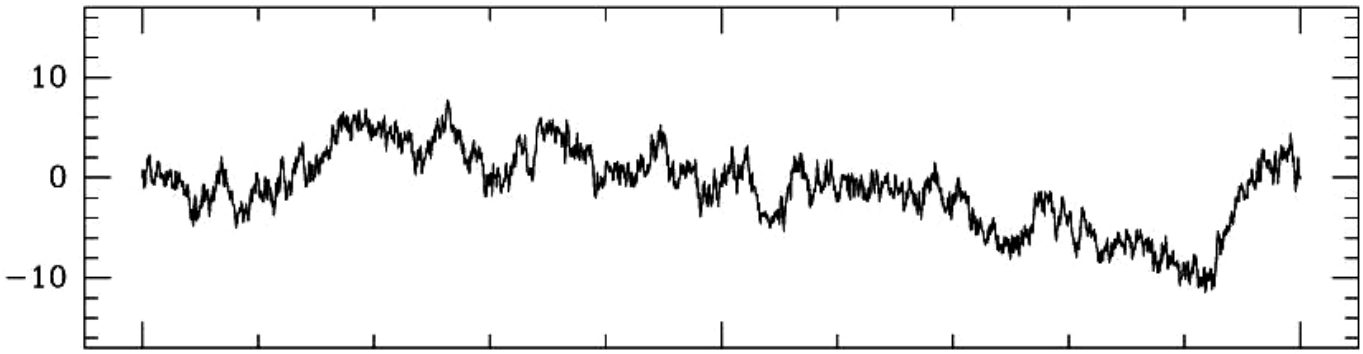
$$C_f(\tau) = \langle \Delta_1 f(n)\, \Delta_1 f(n+\tau) \rangle - \langle \Delta_1 f(n) \rangle^2$$

$$\sim \tau^{2H-2}$$

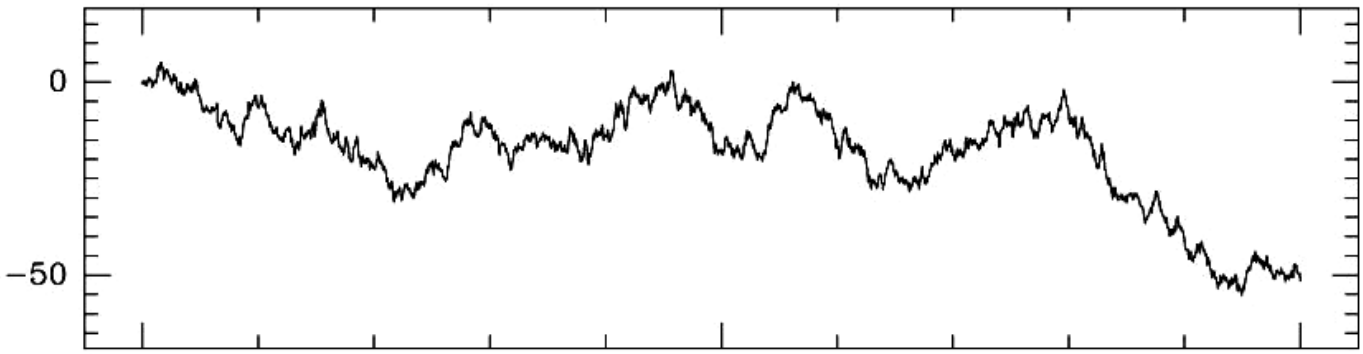
SYNTHETIC DNA WALKS

Fractional Brownian motions : B_H

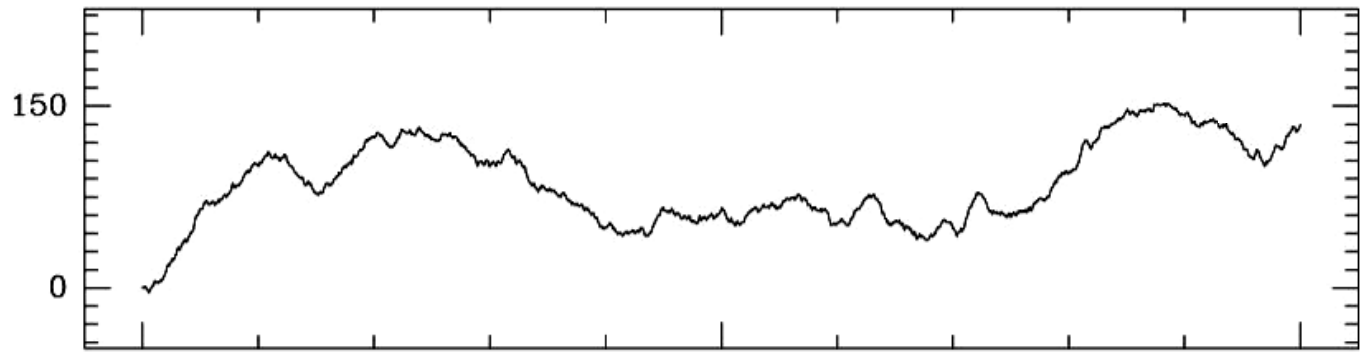
H = 0.3 anti-correlated



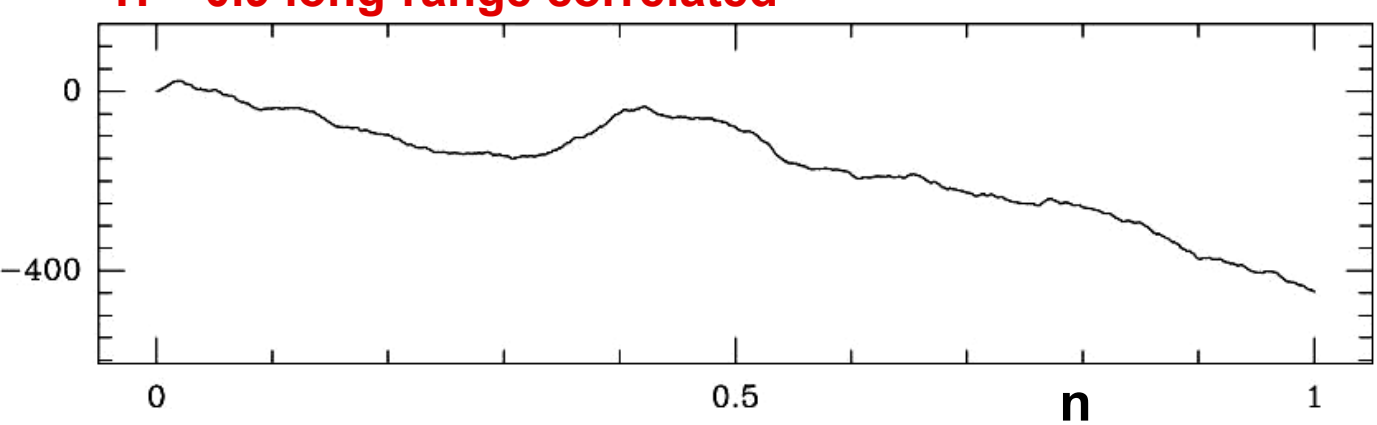
H = 0.5 uncorrelated



H = 0.7 long-range correlated



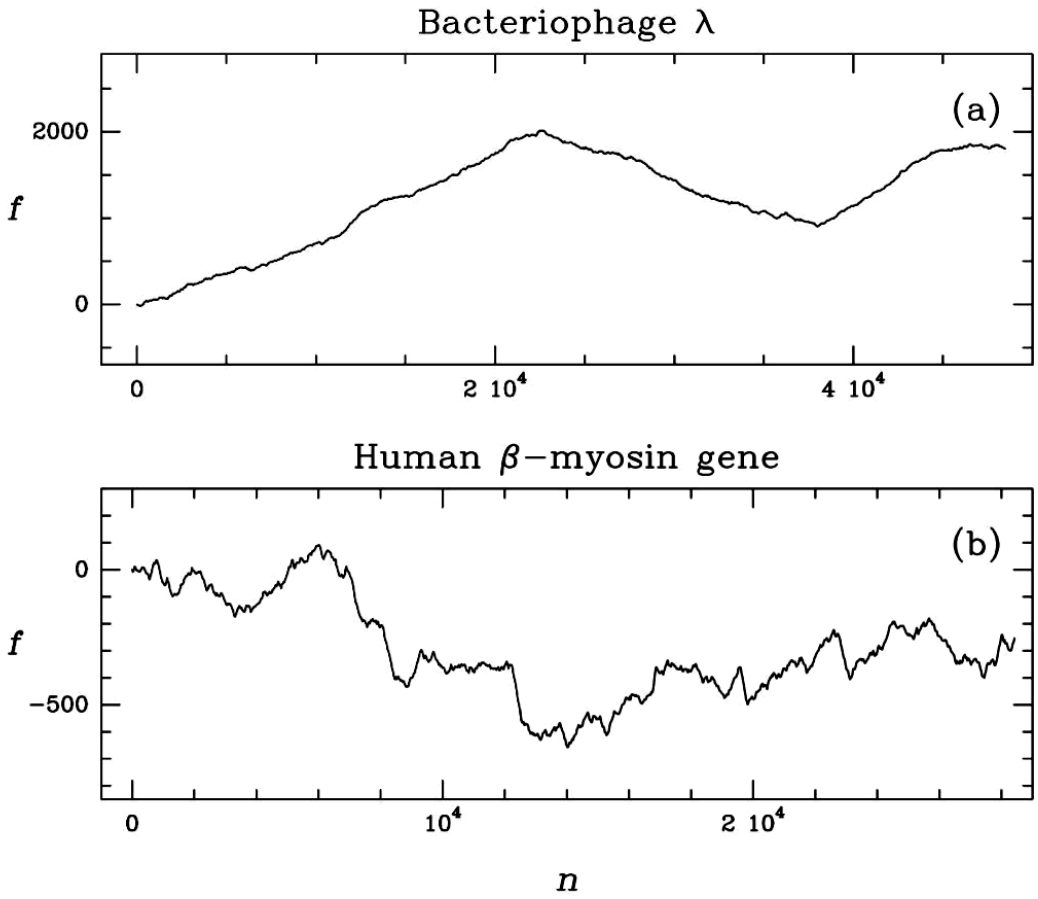
H = 0.9 long-range correlated



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 :

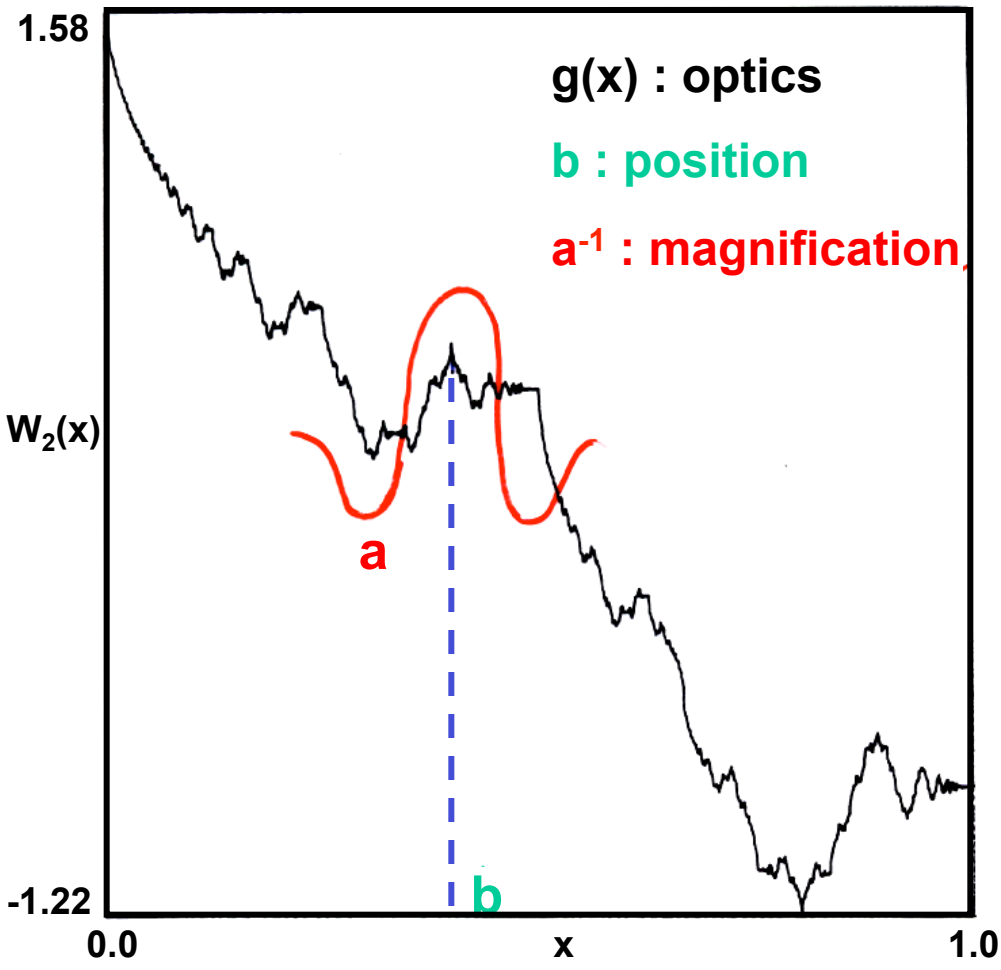


A specific analysing tool is needed to avoid confusing a biased uncorrelated random walk with an unbiased correlated random walk

WAVELET ANALYSIS OF FRACTAL SIGNALS

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

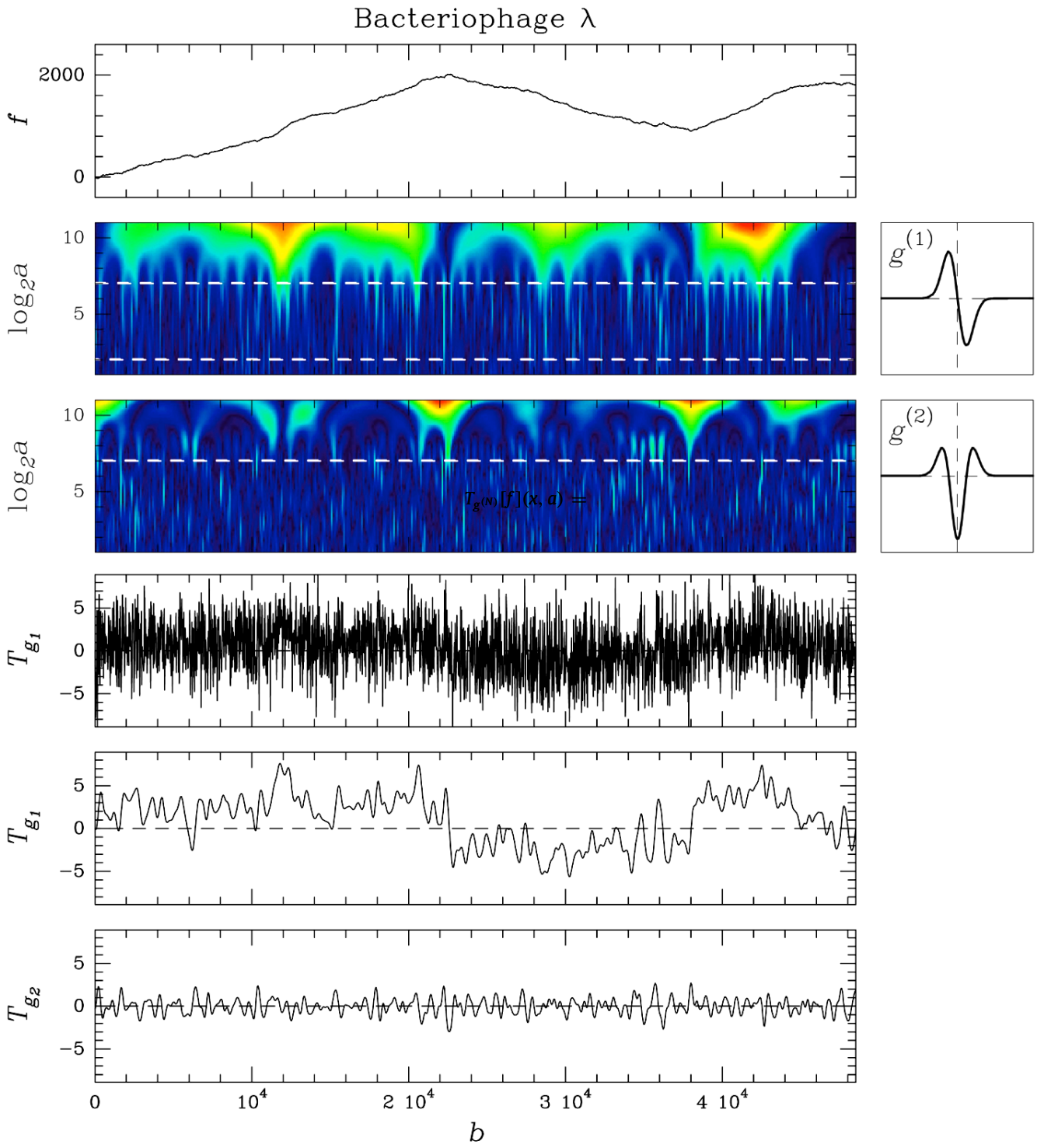
Mathematical microscope



‘ ‘ Singularity scanner’ ’

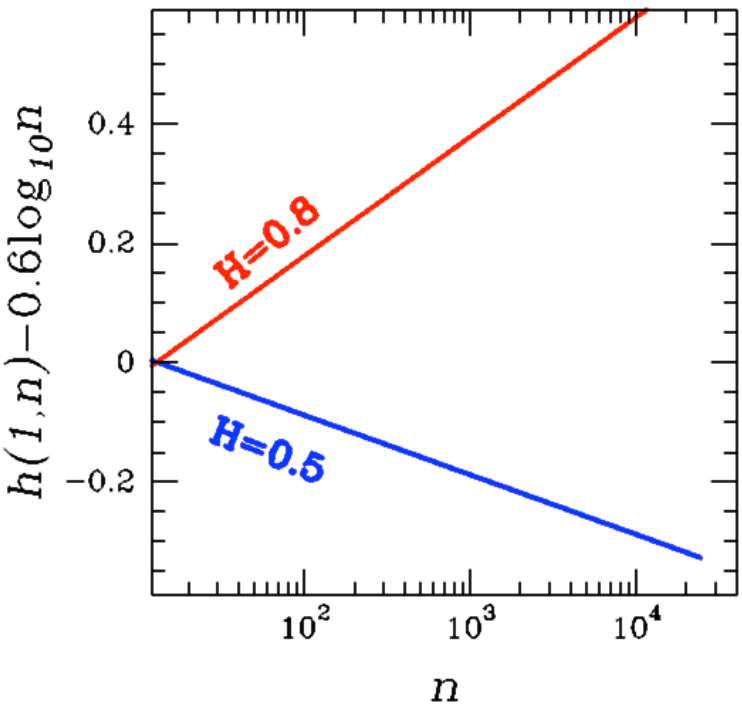
The wavelet transform allows us to **LOCATE** (**b**) the singularities of f and to **ESTIMATE** (**a**) their strength h(x) (Hölder exponent)

WAVELET ANALYSIS OF THE DNA SEQUENCE OF THE BACTERIOPHAGE λ



$$T_{g(N)}[f](x,a) = (-a)^N \frac{\mathrm{d}^N}{\mathrm{d}x^N} T_{g(0)}[f](x,a).$$

A UNIQUE WAY TO DISPLAY RESULTS

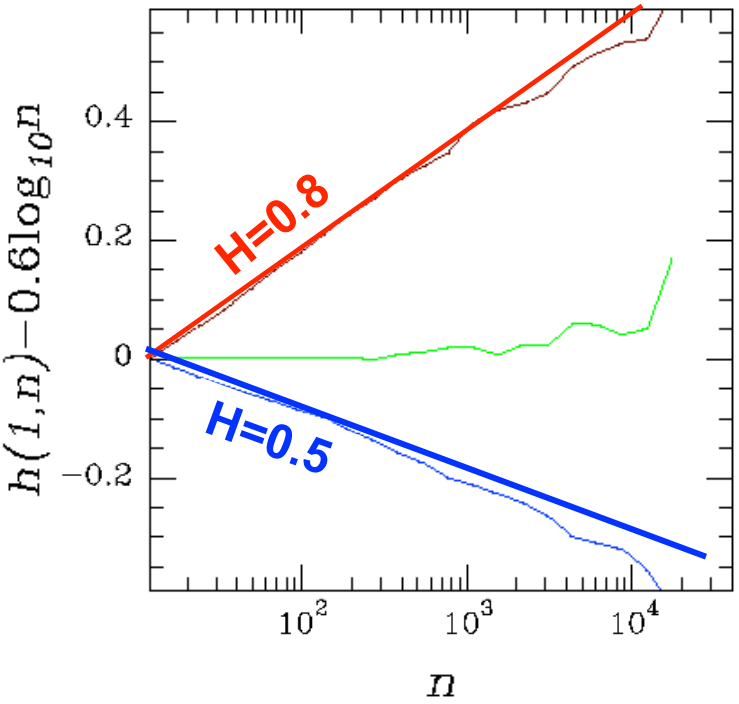


1. Straight line ⇔ scale invariance properties

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

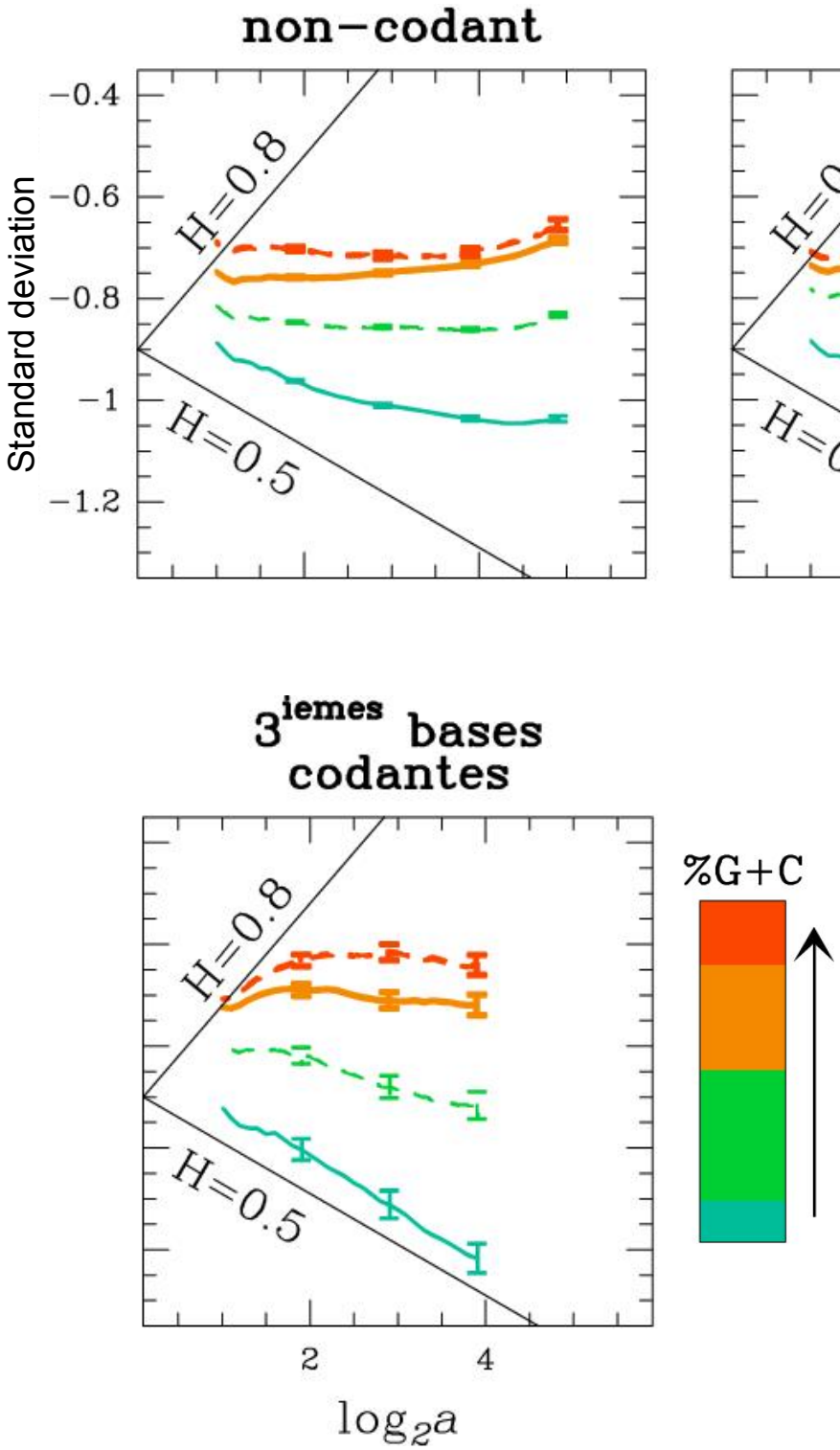


1. Straight line ⇔ scale invariance properties

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

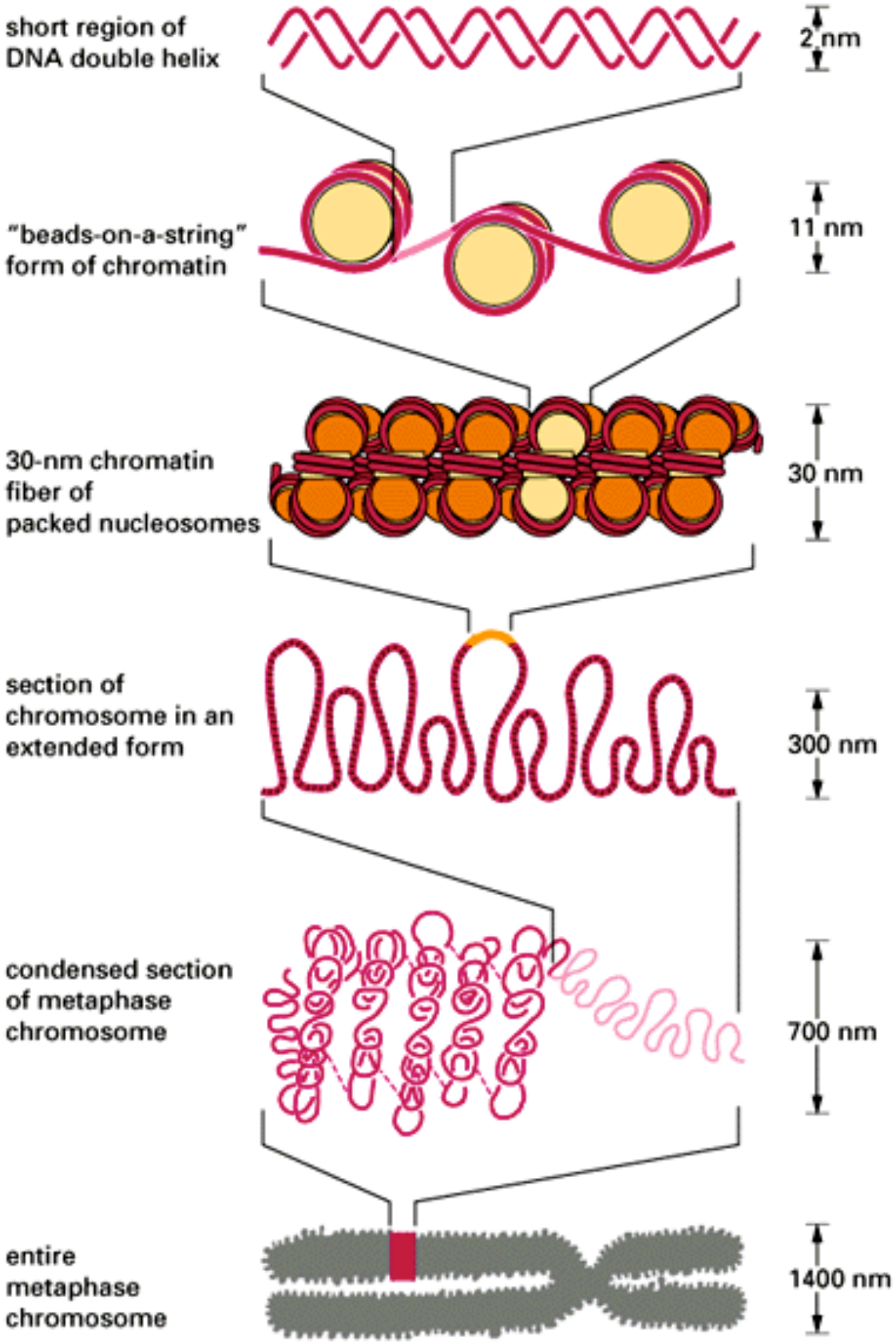
  Chromatine fiber

  Higher order folding up to the metaphase chromosome

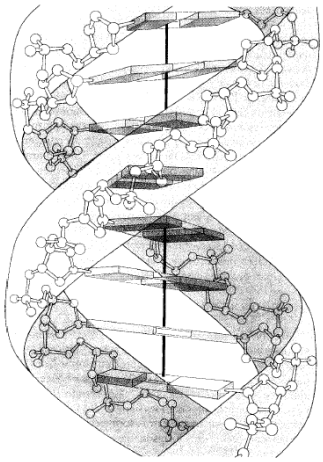
  Fractal model of chromosomes (Takahashi 89)

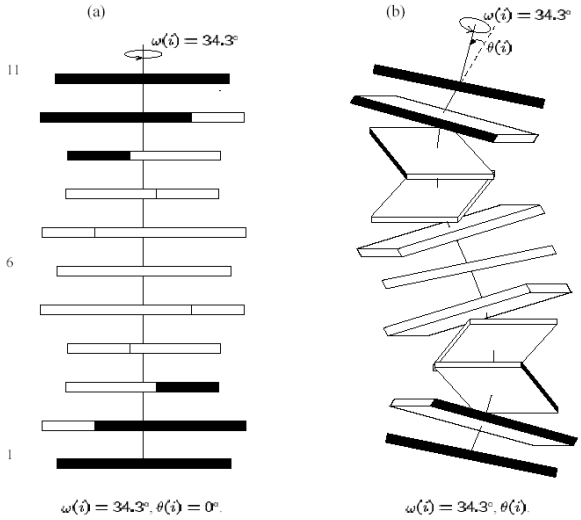
  Crumpled globule model (Grosberg *et al.* 93)

HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA



DNA WALKS THAT REFLECT THE STRUCTURE OF THE DNA POLYMER





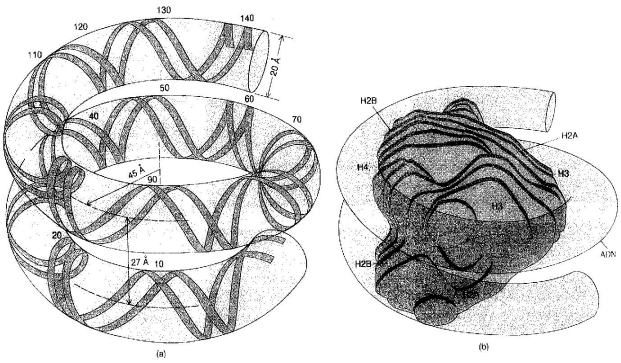
2 trinucleotide codings based on experiments :

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

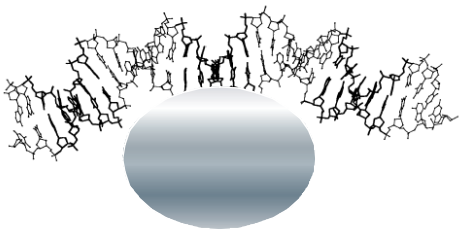
Nucleosome positioning

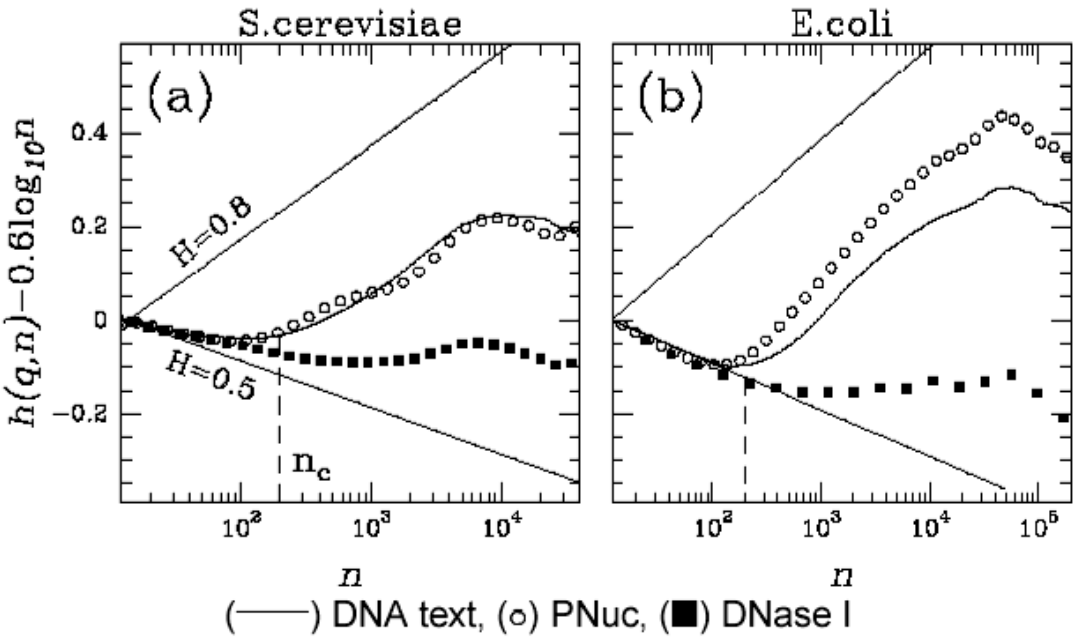
local curvature



Dnase I sensitivity

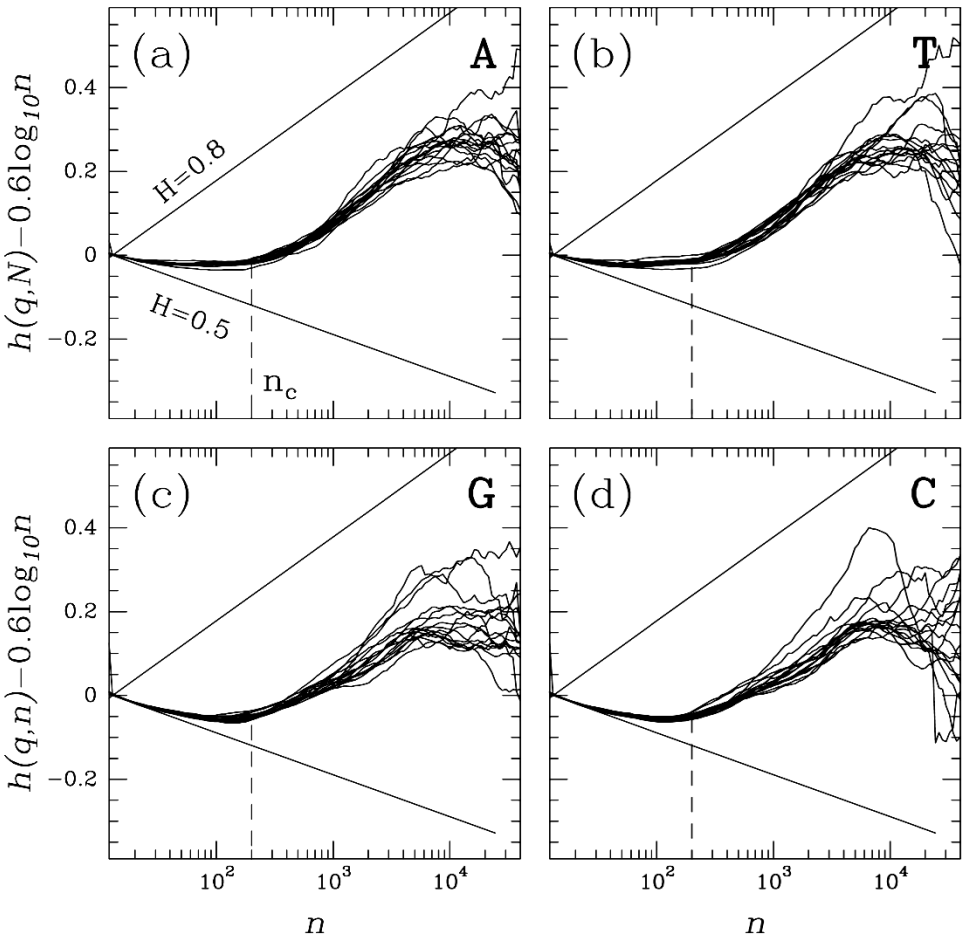
Local flexibility





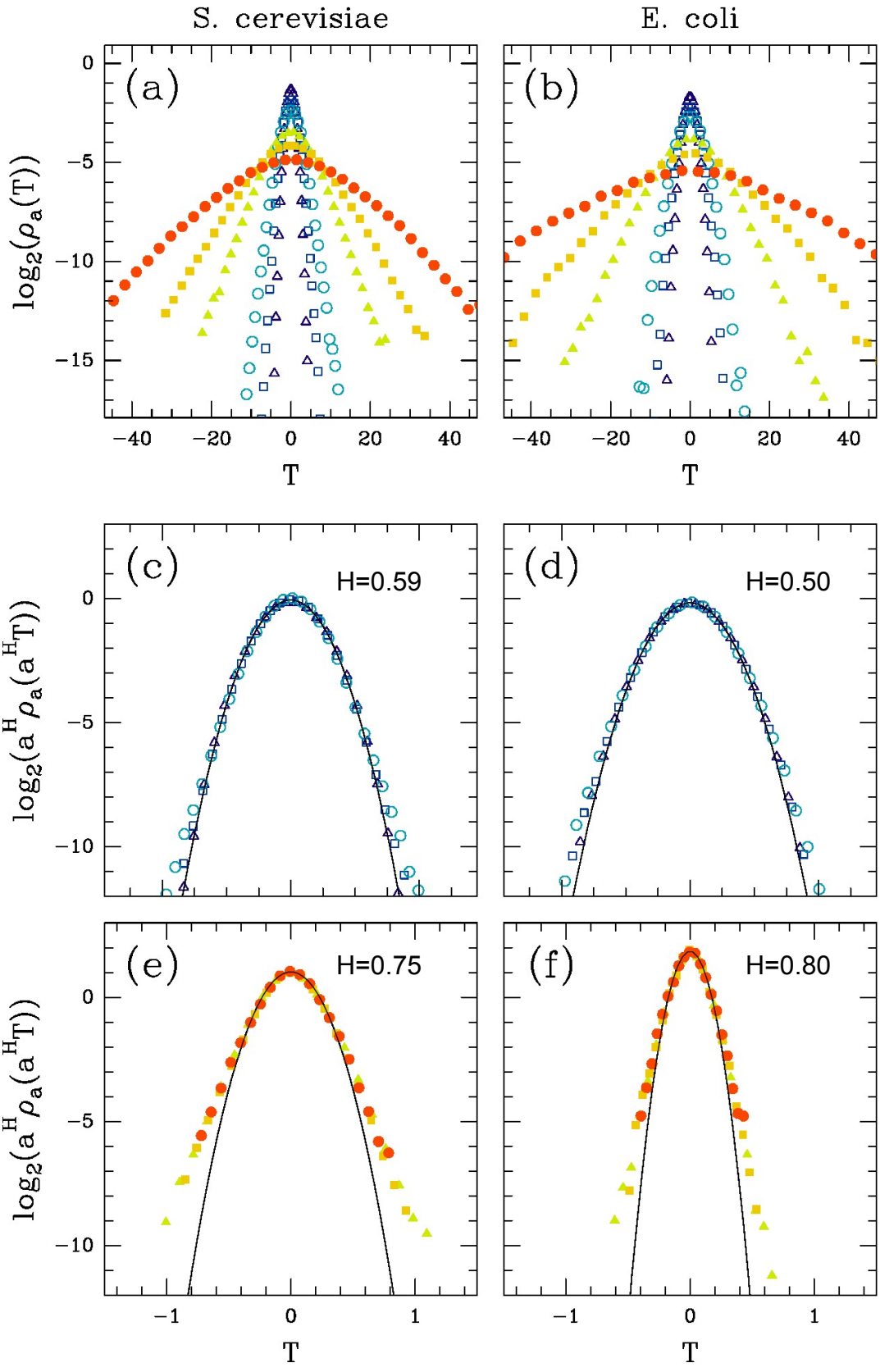
Hypothesis: LRC in the small scales regime is the signature of of the nucleosomal structure

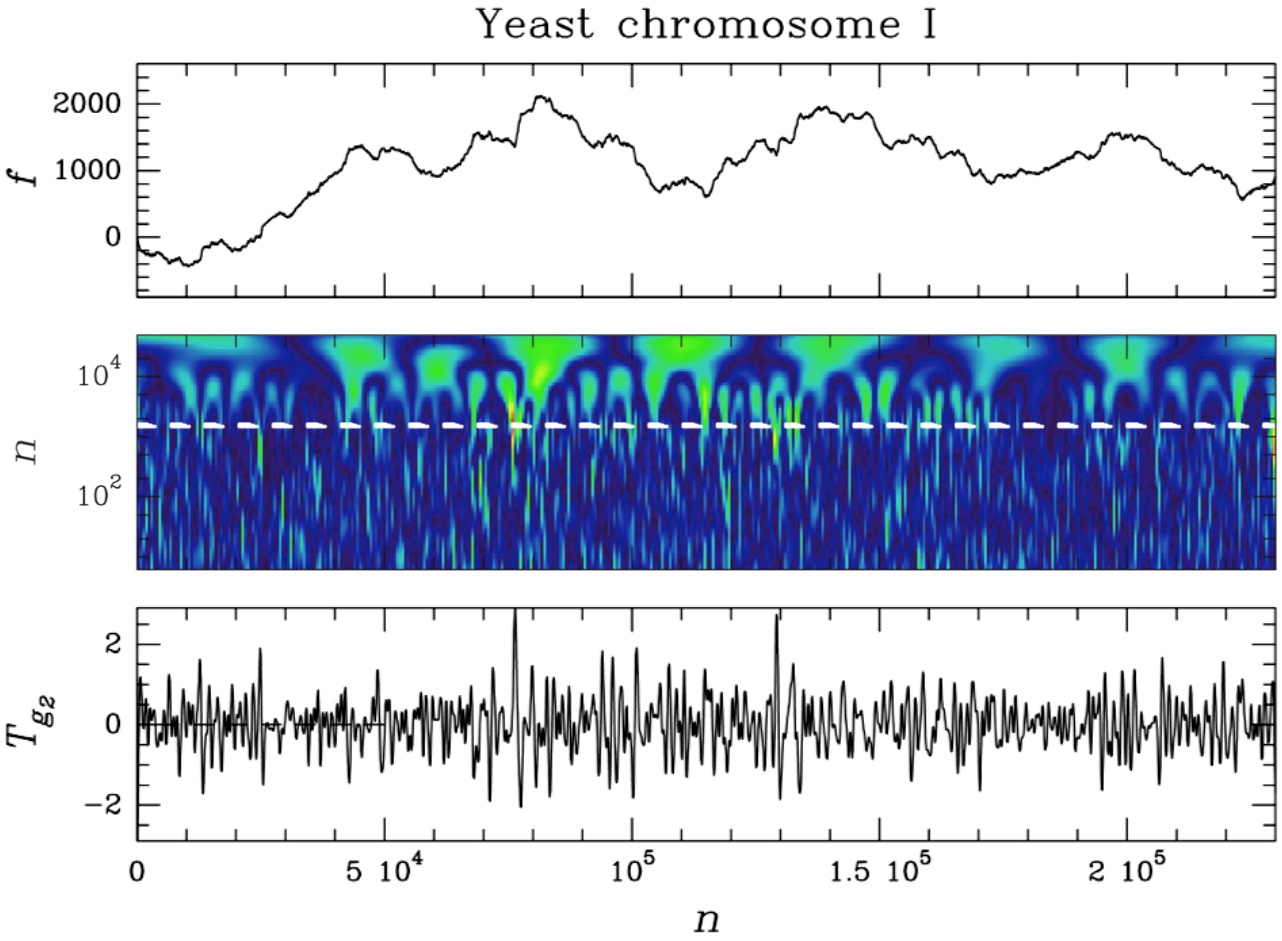
STATISTICAL ANALYSIS OF THE EUKARYOTIC GENOME OF *Saccharomyces cerevisiae*

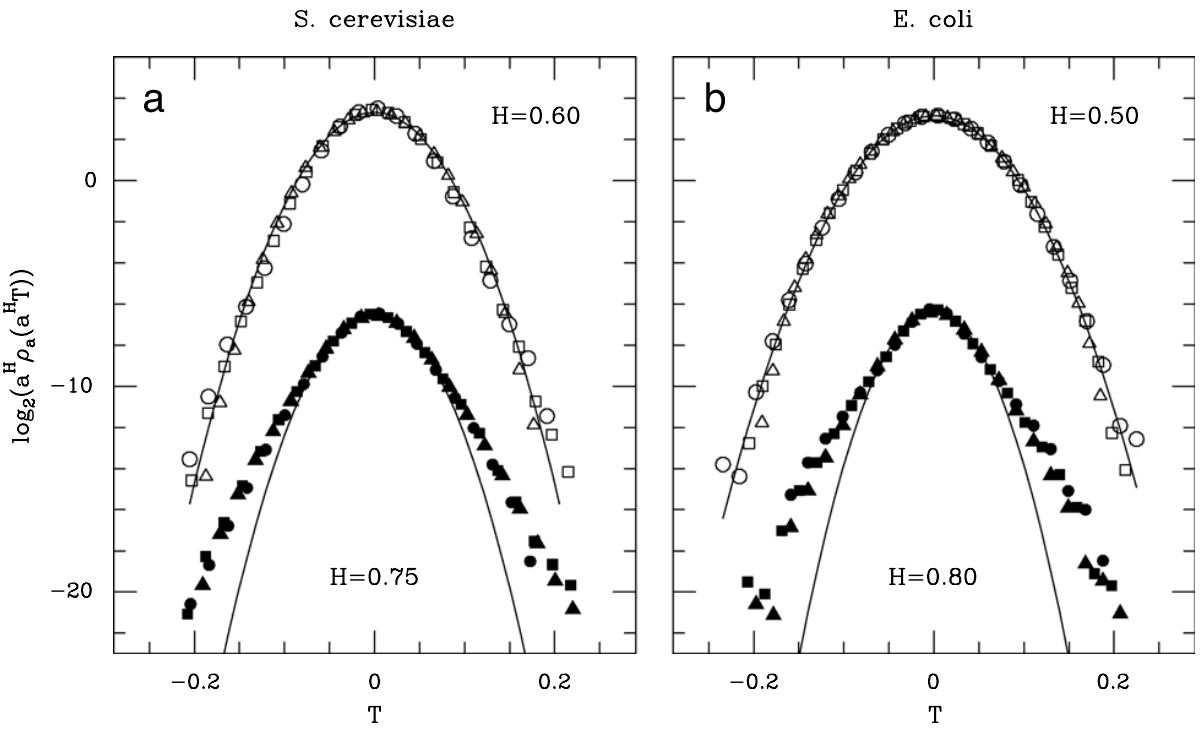


Universality between the 16 chromosomes of yeast
Universality between the 4 mononucleotidic codings
$n_c \sim$ 200bp is a characteristic length scale

Pdf of wavelet coefficient values of "A" DNA walks



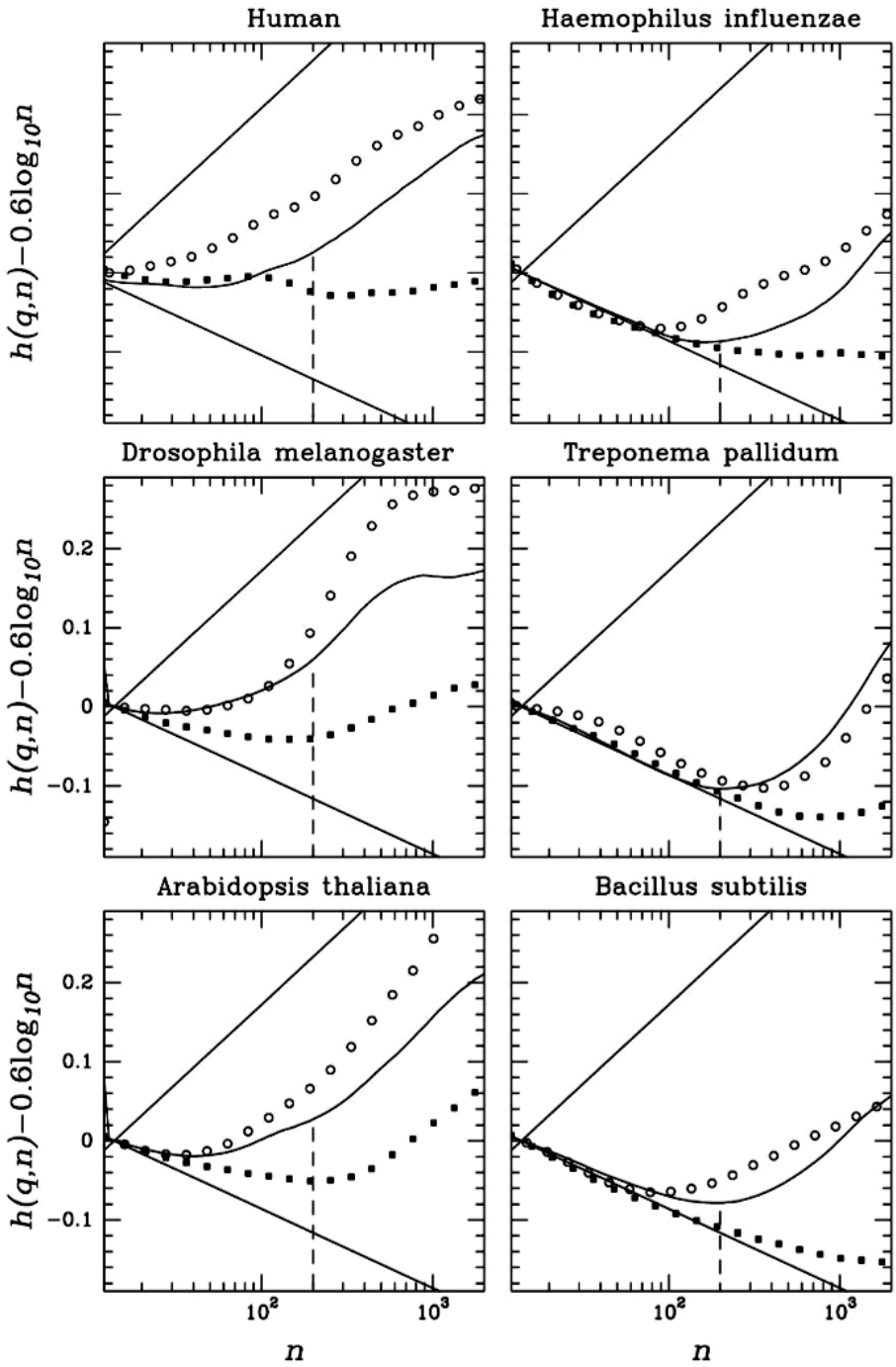




Gaussian statisics at small scales ($n \leq 200$bp)

Non Gaussian (fat tails) statistics at large scale ($n \geq 200$bp)

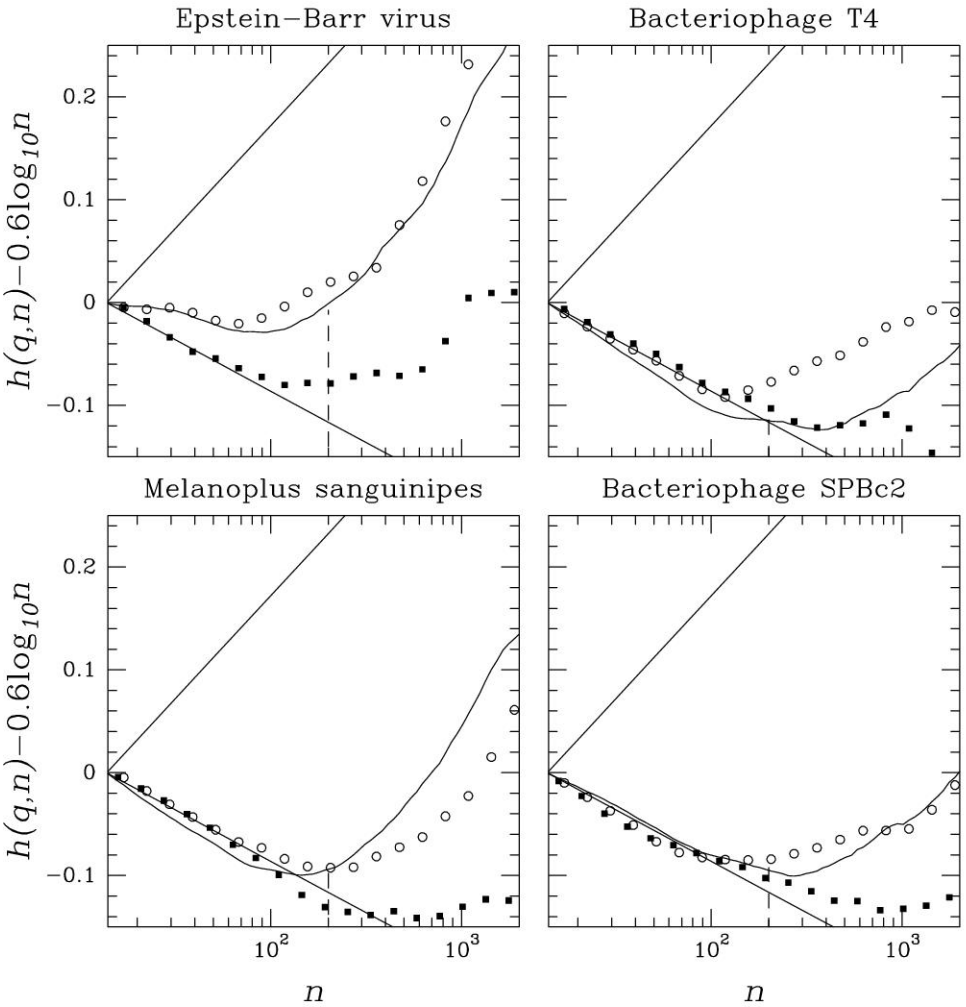
Eucaryotes Bacteria



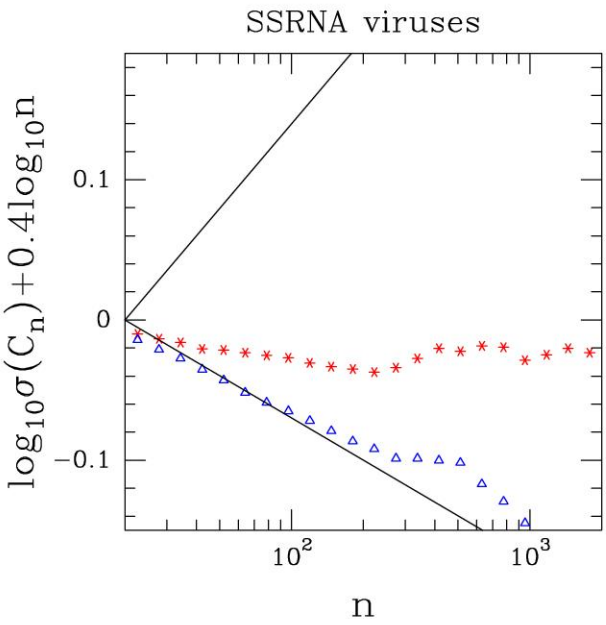
(——) DNA text, (○) PNuc, (■) DNase I

Nucleosomes No nucleosomes

SMALL SCALES LRC ARE RELATED TO NUCLEOSOME LIKE STRUCTURES

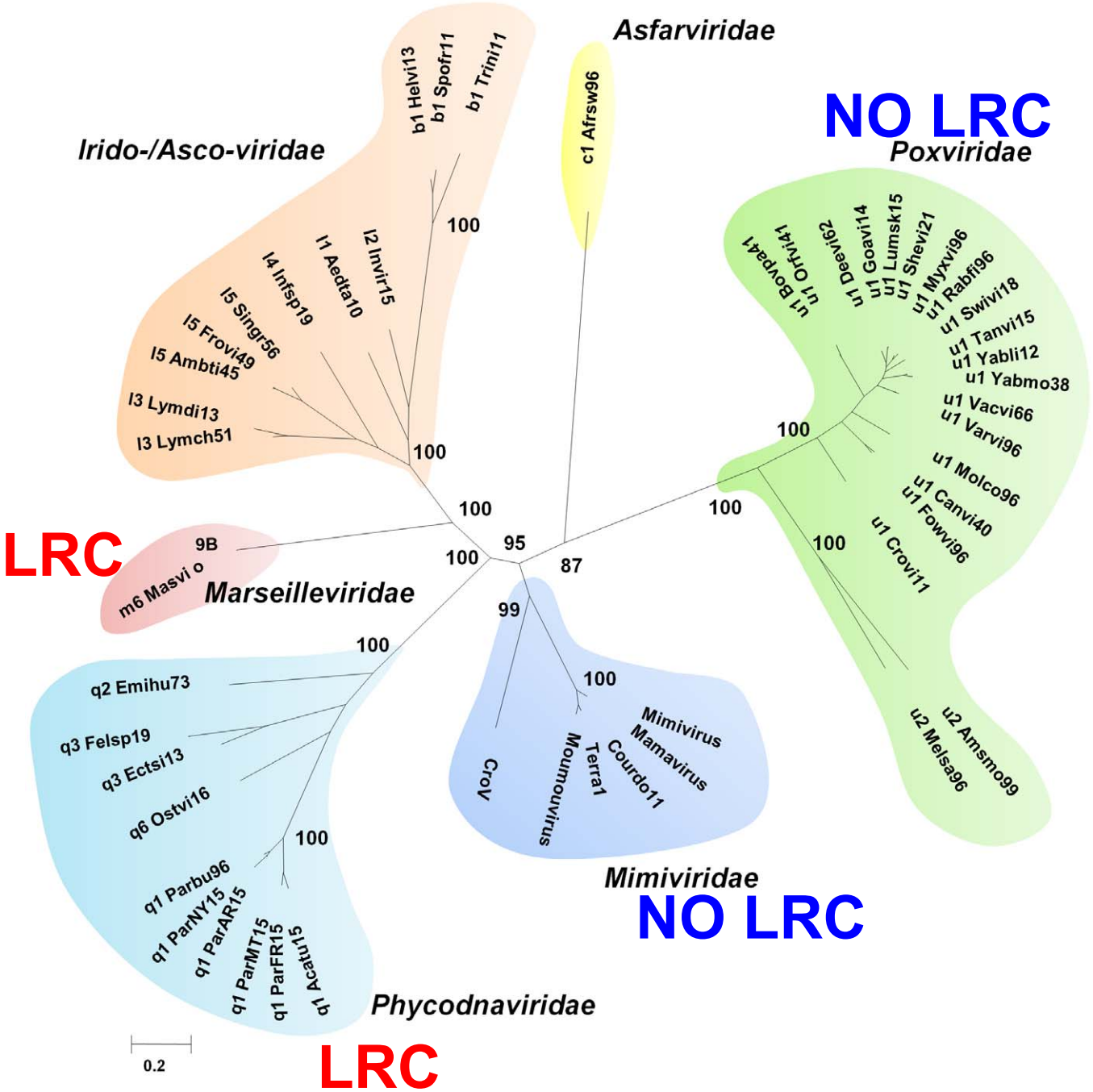


Pox virus don't display LRC in the small scale regime



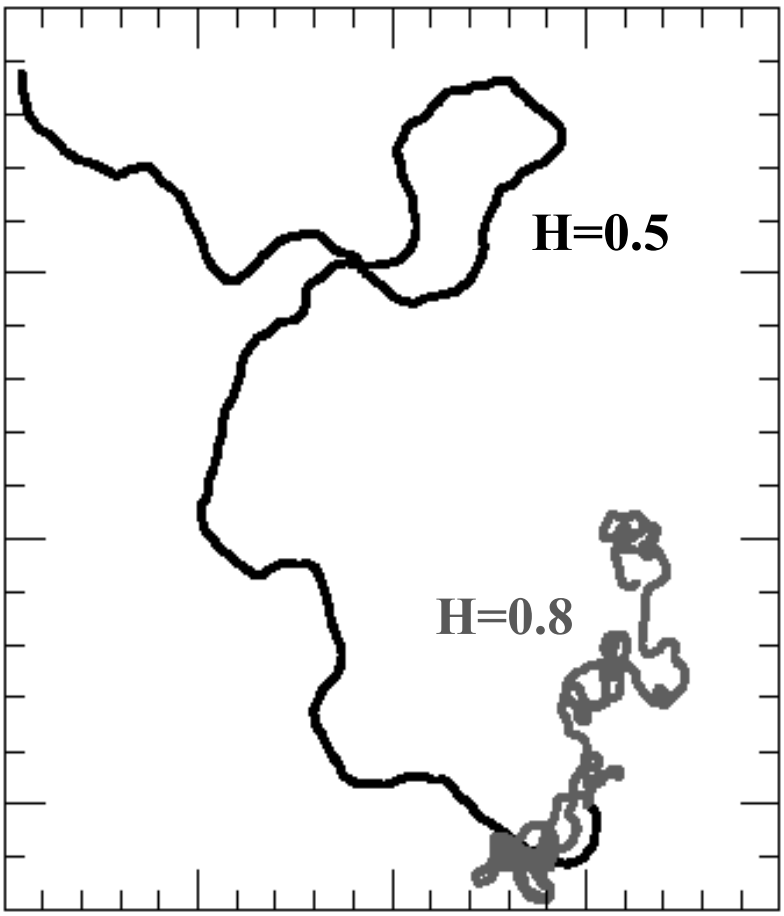
Among the SSRNA viruses only retroviruses display LRC in the small scale regime

Nucleocytoplasmic large DNA viruses



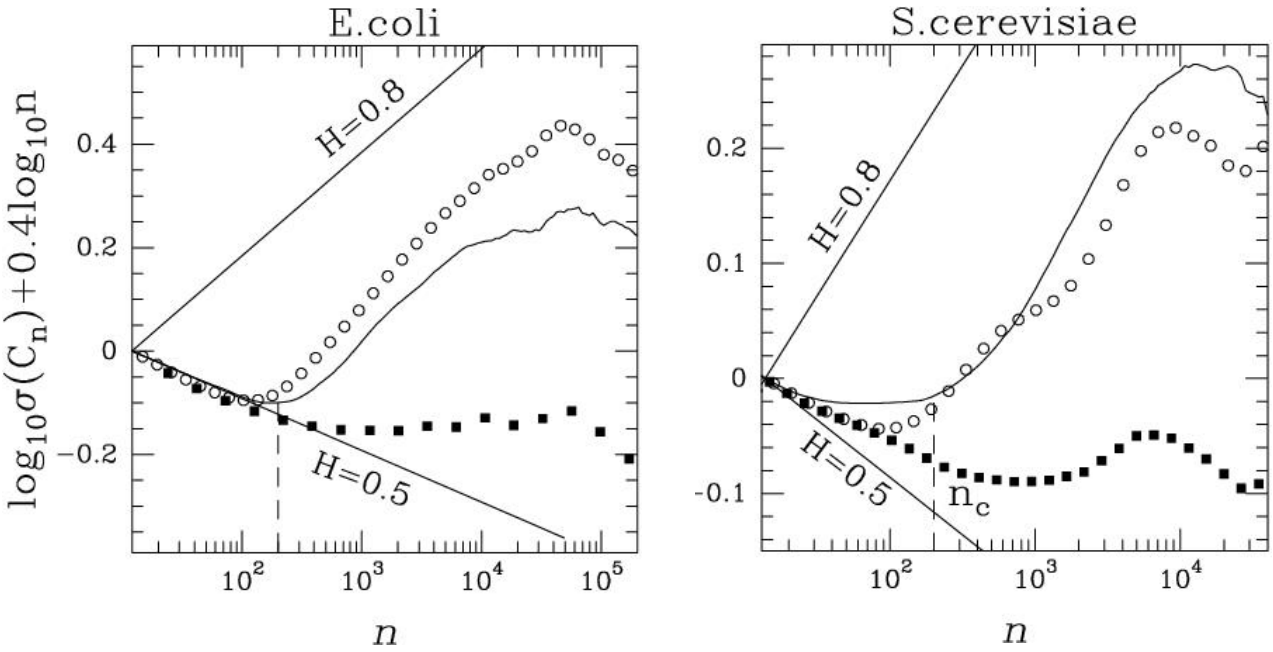
Influence of the DNA sequence on the formation and dynamics of nucleosomes

Uncorrelated DNA

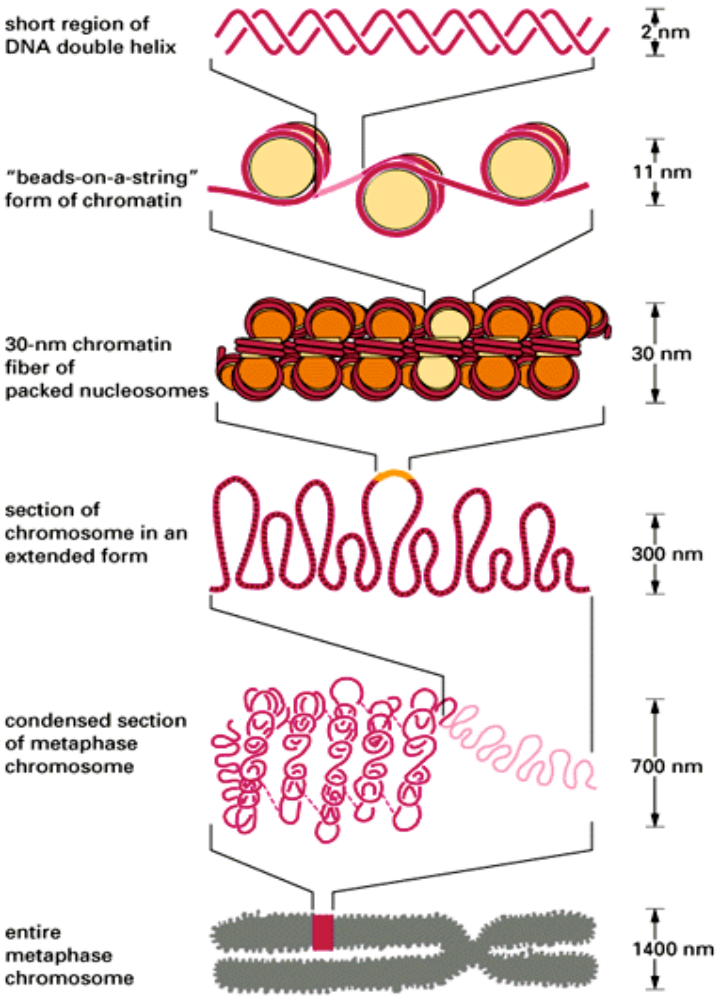


Long-range correlated DNA

L_{DNA} = 3000 bp

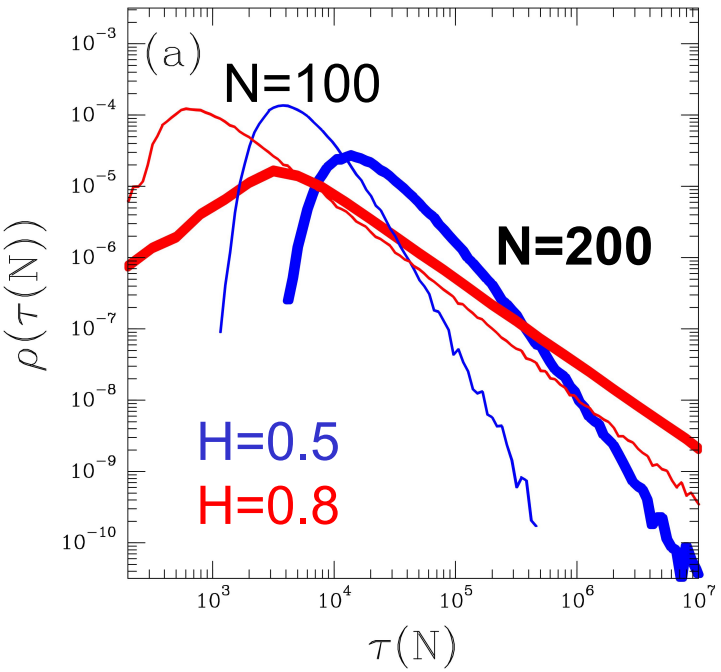


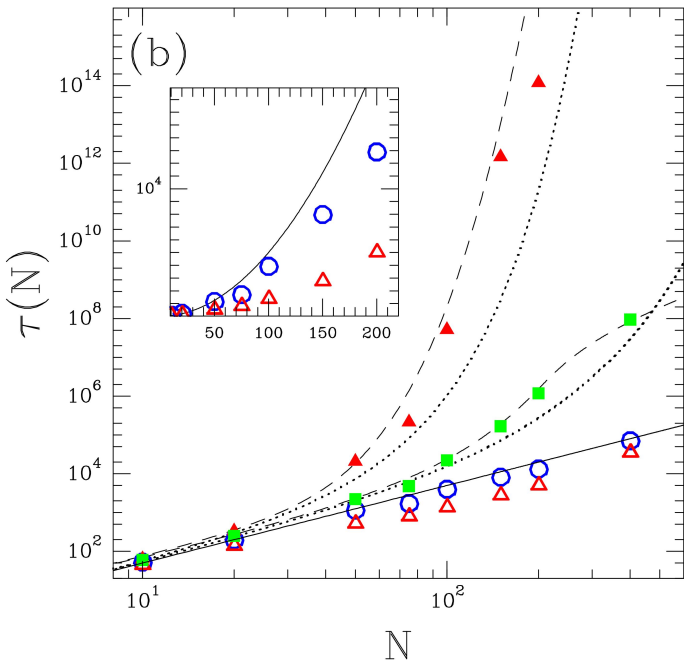
HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA



LRC favour nucleosome mobility

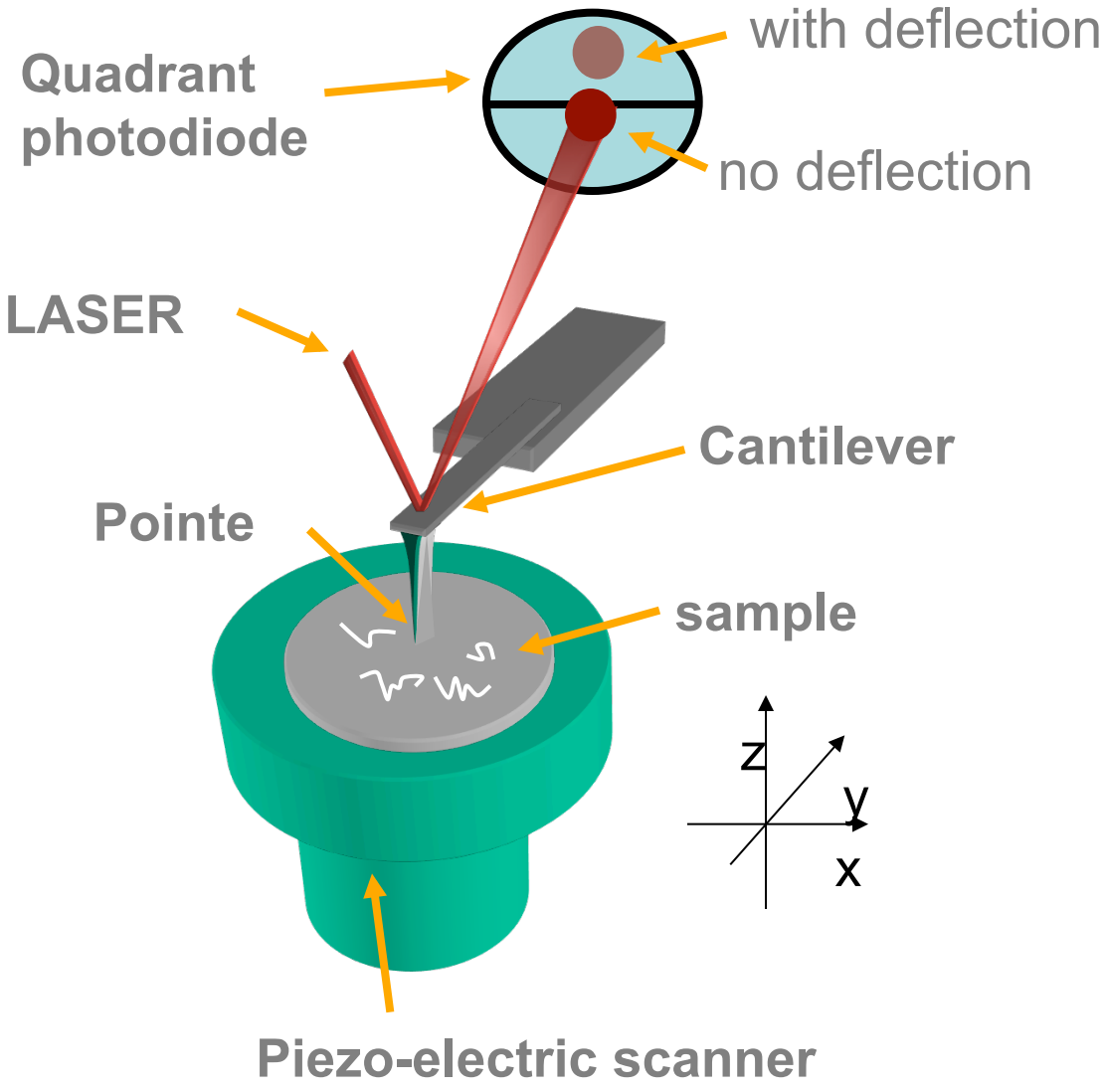
Mean First Passage Time τ at distance N during the diffusion of a nucleosome (l=200) along a semi-flexible chain of length $L >> l$
(**2D elastic model**)

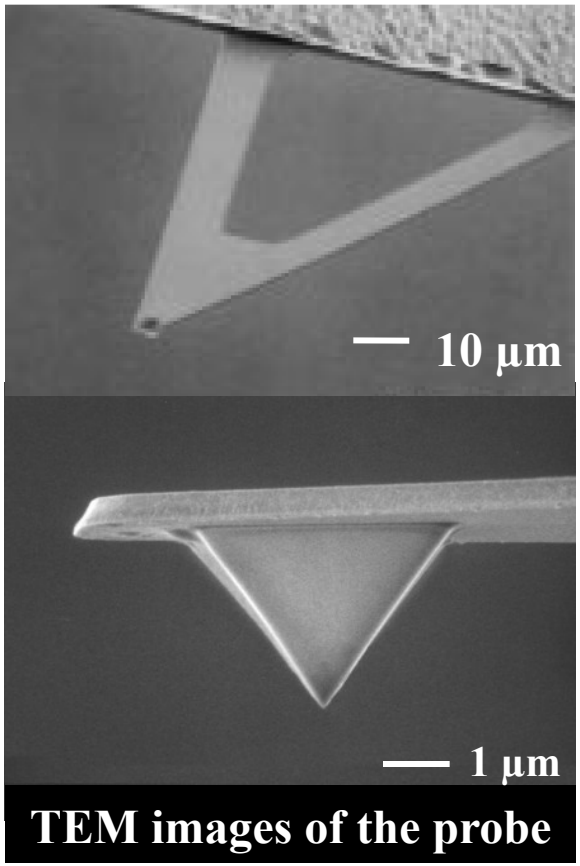




Probing long-range correlations in eukaryotic DNA with Atomic Force Microscopy

Atomic Force Microscopy

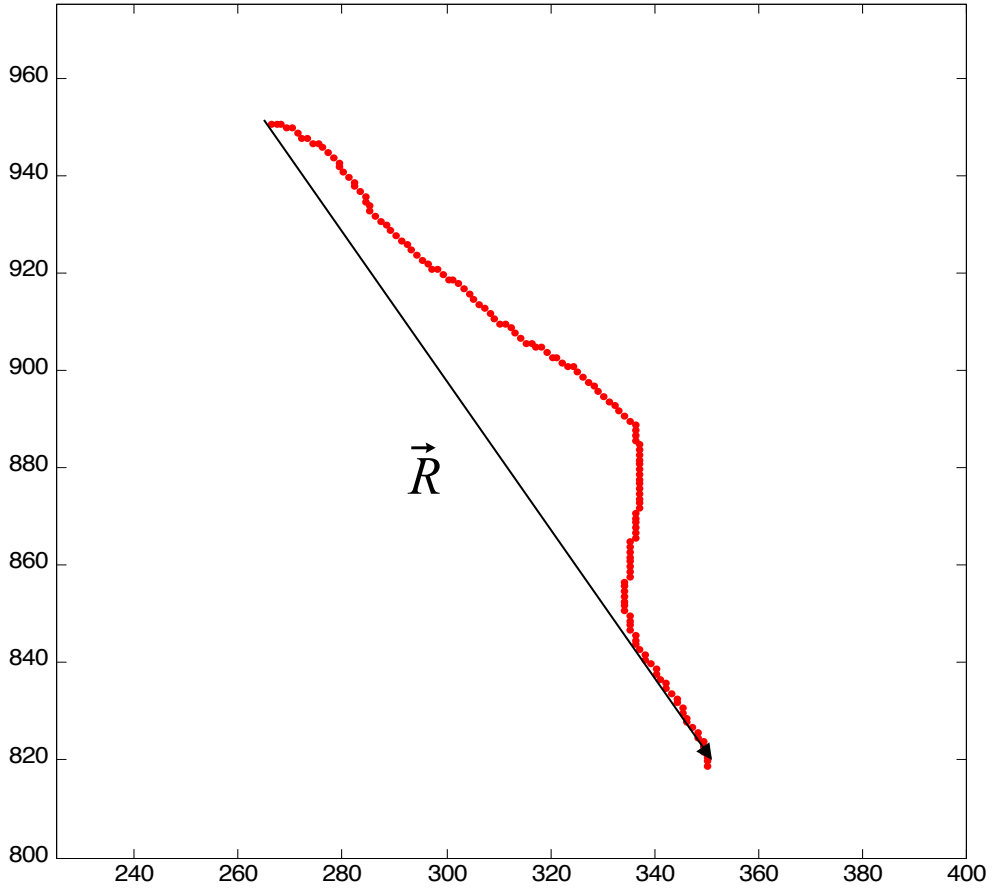




TEM images of the probe

Image analysis

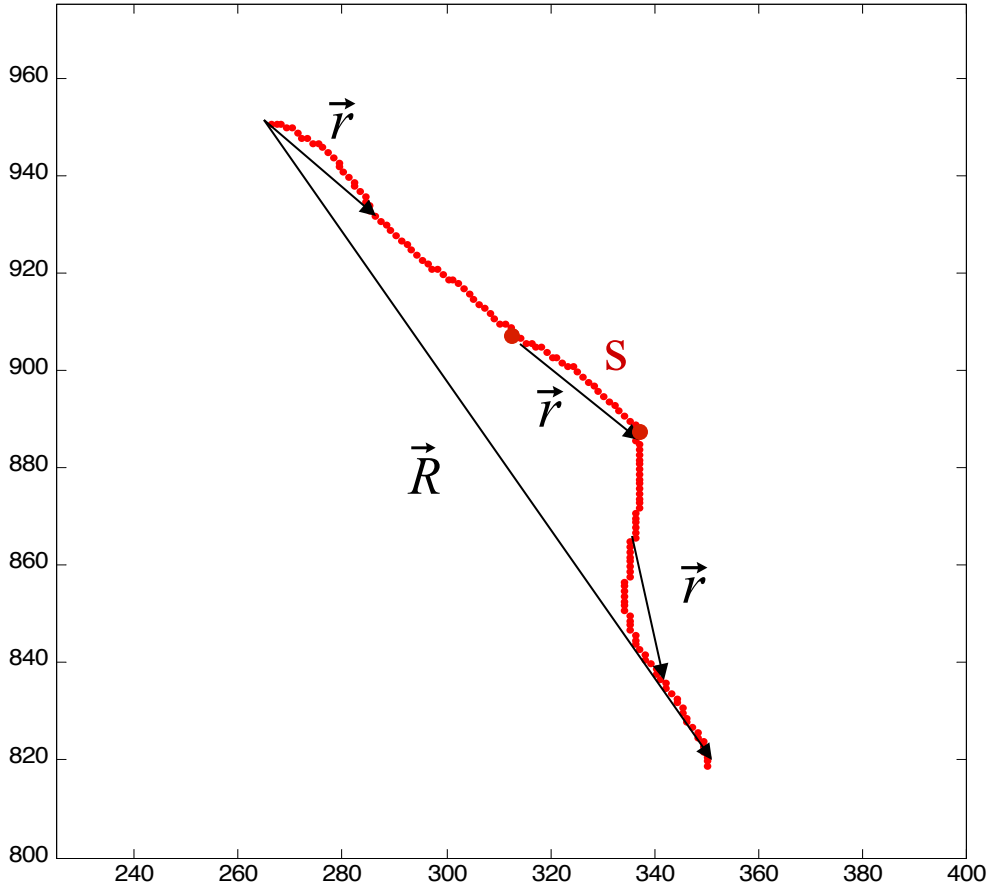
local persistence length calculation



$$l_p = \lim_{L\to\infty} \frac{\langle \vec{R}^2 \rangle}{2L}$$

Image analysis

local persistence length calculation

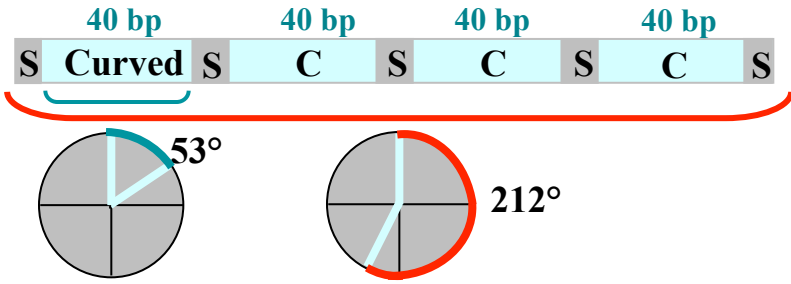


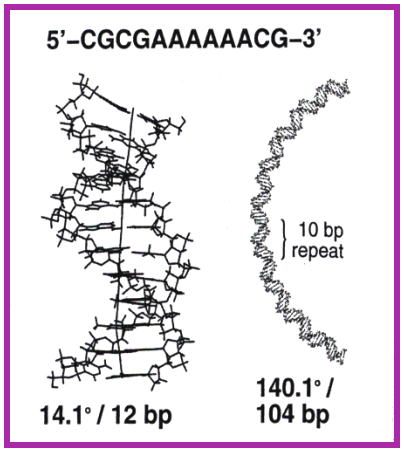
$$l_p = \lim_{L\to\infty} \frac{\langle \vec{R}^2 \rangle}{2L}$$

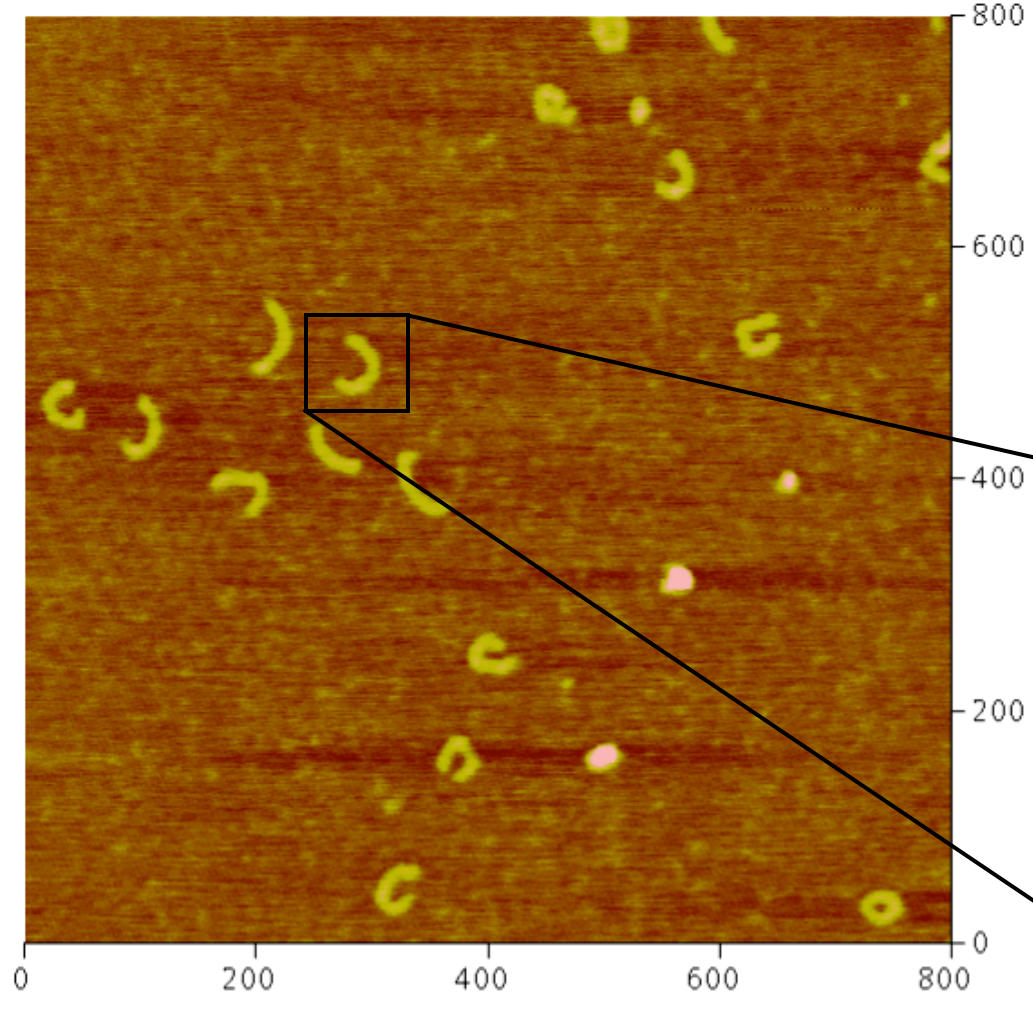
$$l_p(s) = \frac{\langle \vec{r}^2 \rangle}{2s}$$

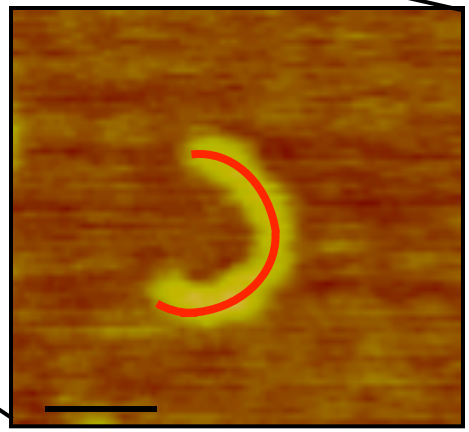
Highly curved DNA fragment (200 bp)

Construction of an artificial DNA with very high curvature:



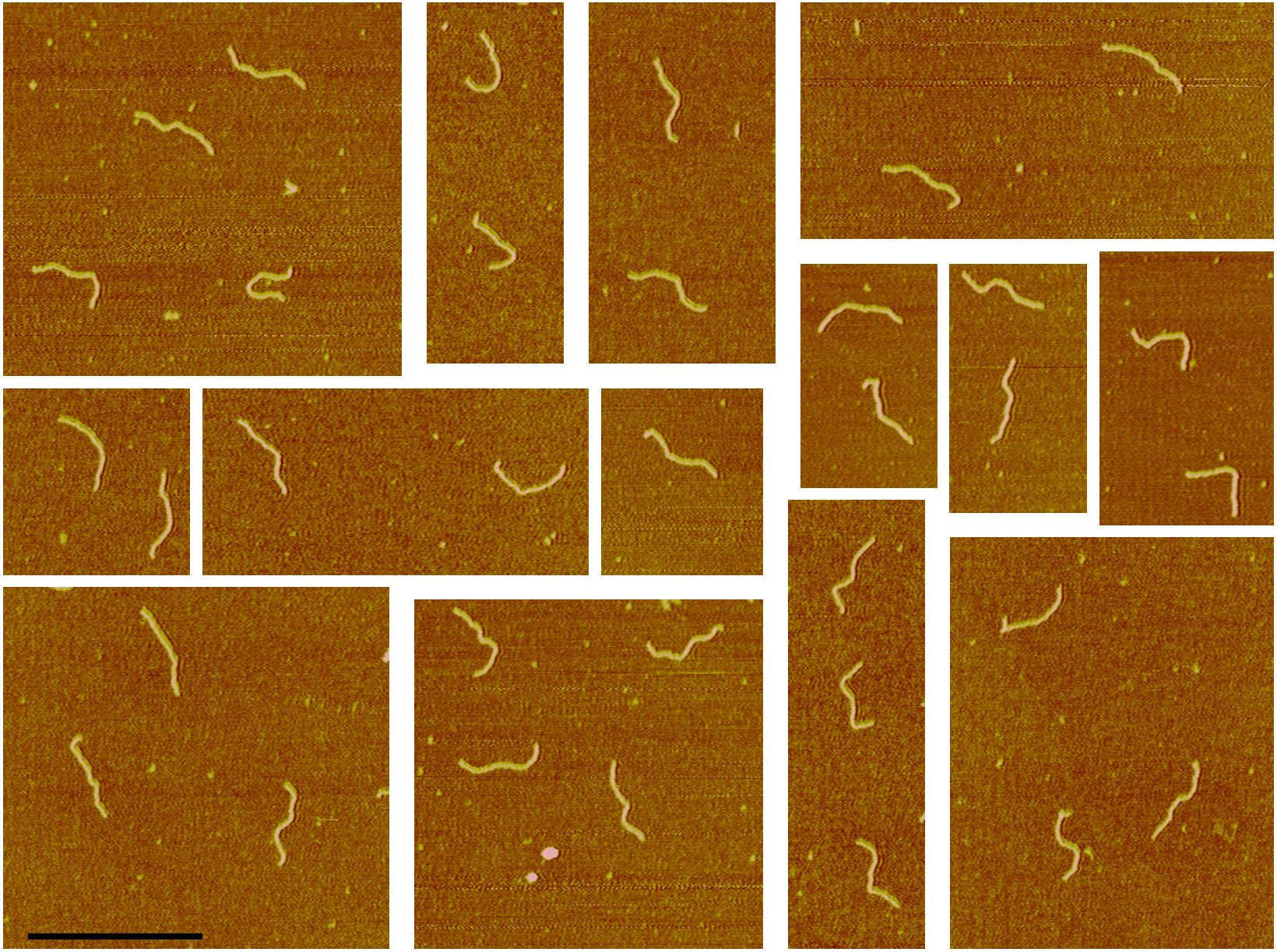




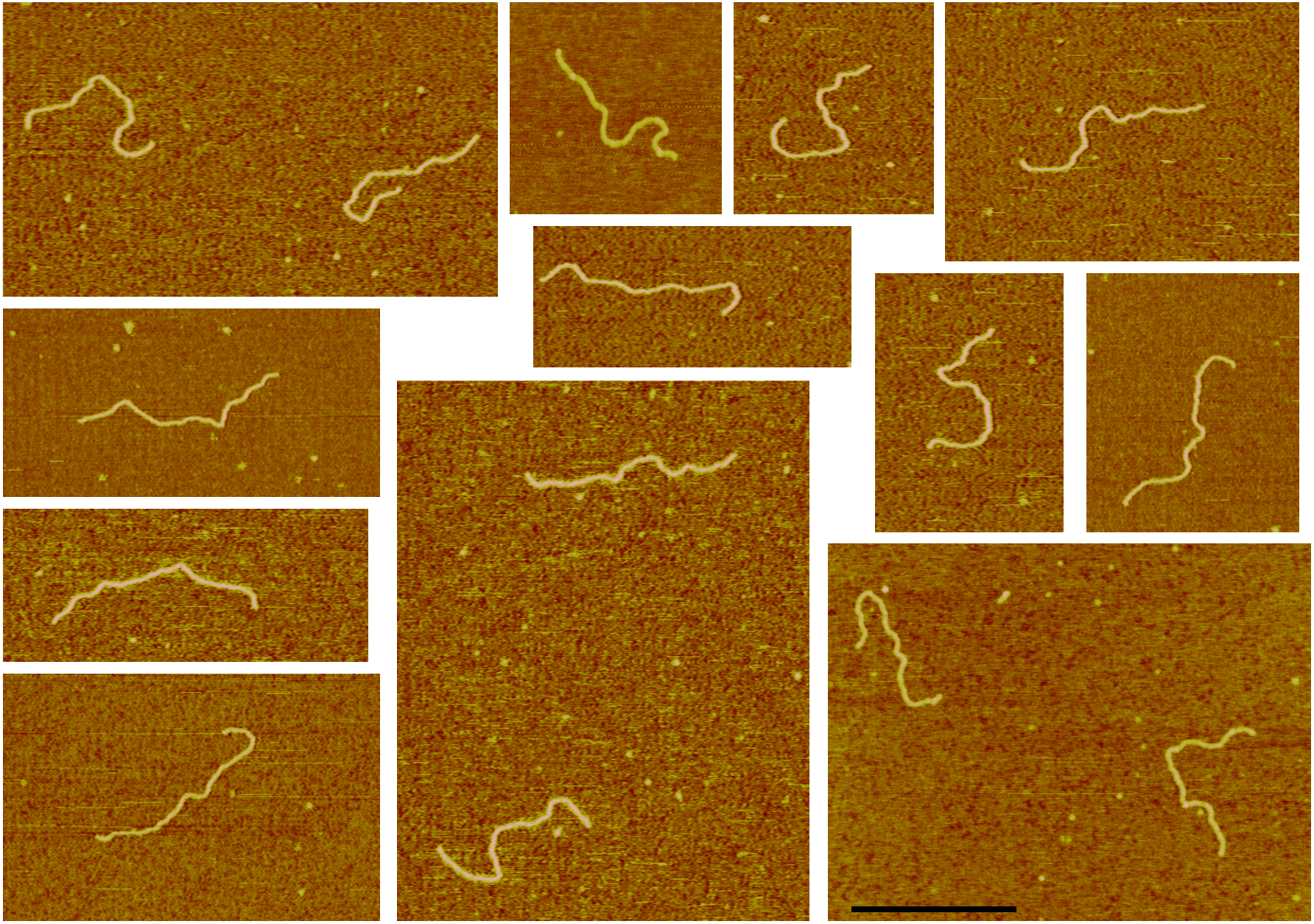


50 nm

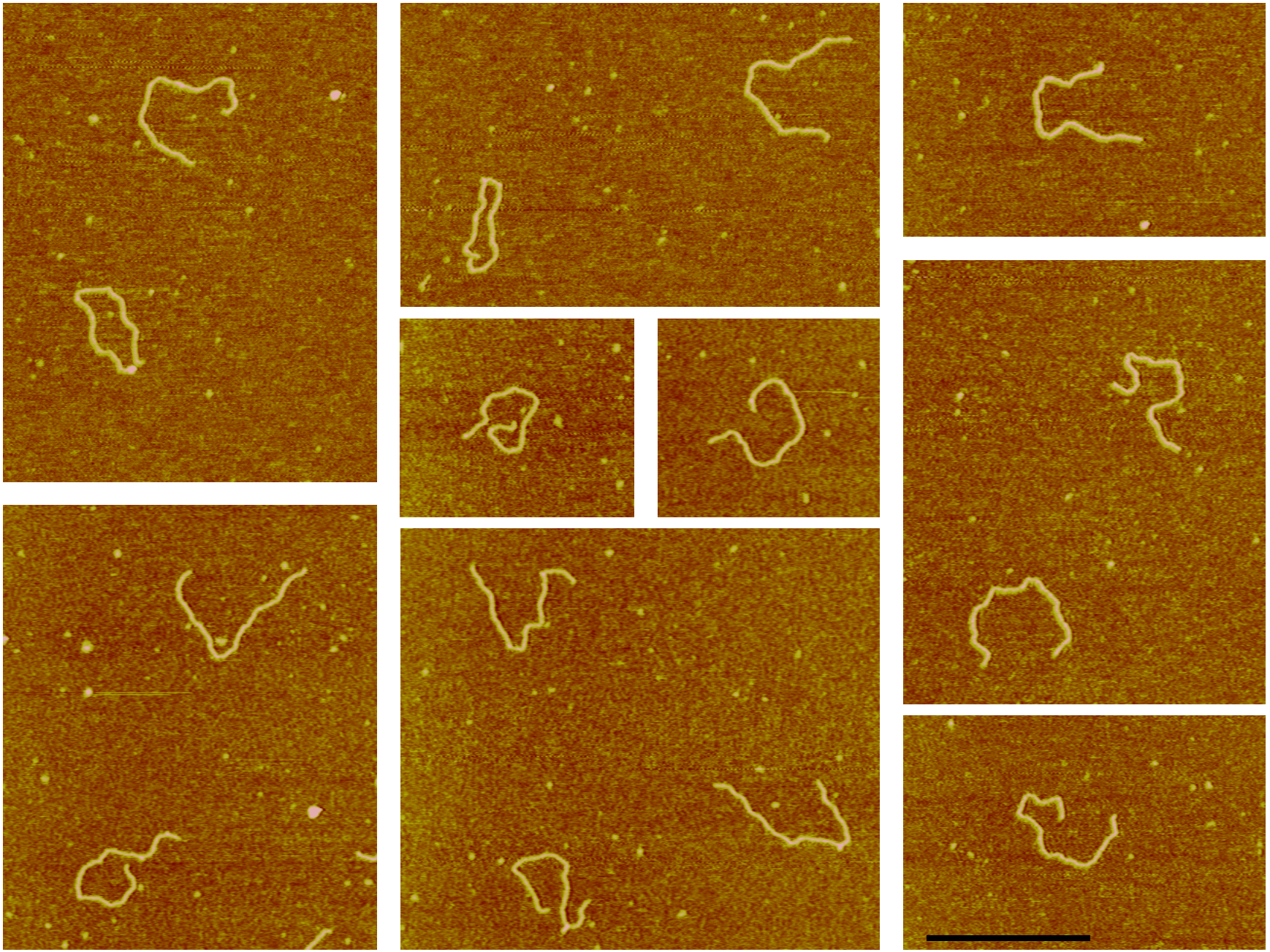
AFM images of 800 bp straight DNA (no structural disorder, $\sigma \sim 0$) in 2D



AFM images of 2200 bp virus DNA (uncorrelated structural disorder H =0.5) in 2D



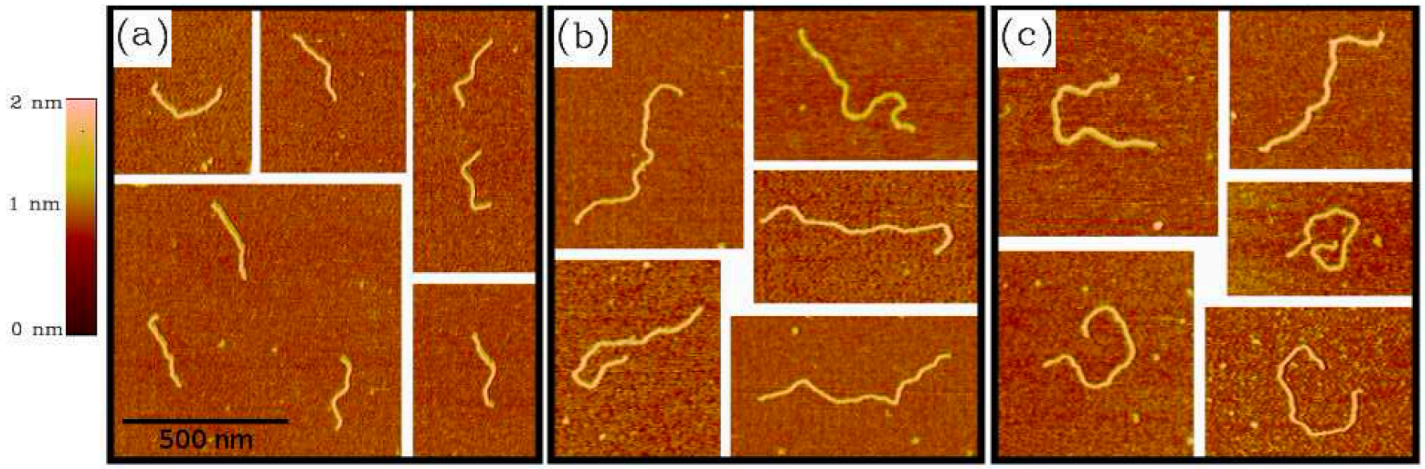
AFM images of 2200 bp human DNA from chromosome 21 (LRC, H~0.8) in 2D

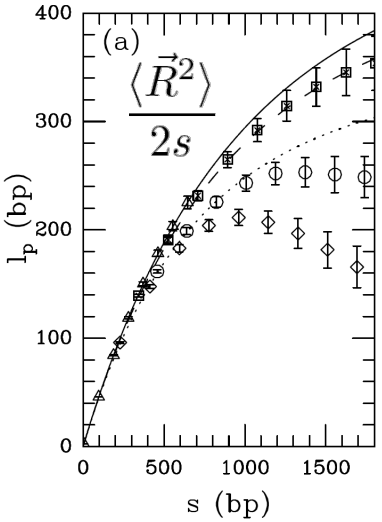


Generalisation of the Worm Like Chain Model to LRC polymers

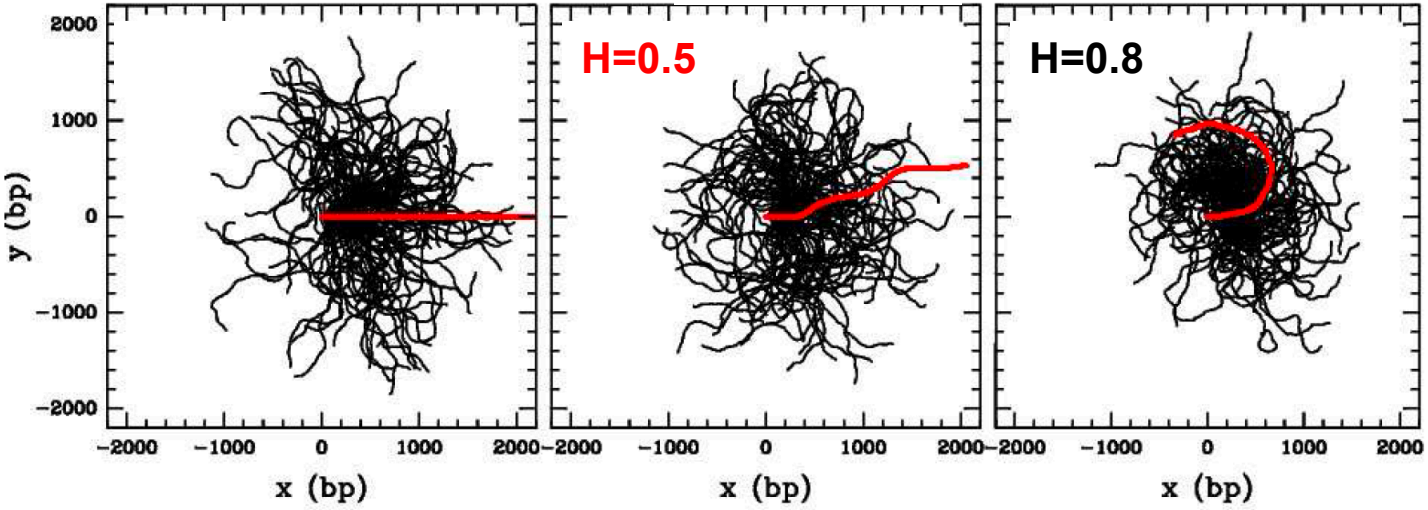
Moukthar et al., Phys. Rev. Lett. (2007); J. Phys. Chem B (2010)

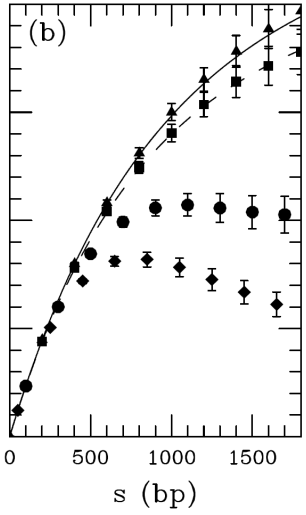
Straight DNA - 800bp **HCV DNA - 2200bp** **Human DNA - 2200bp**





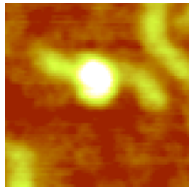
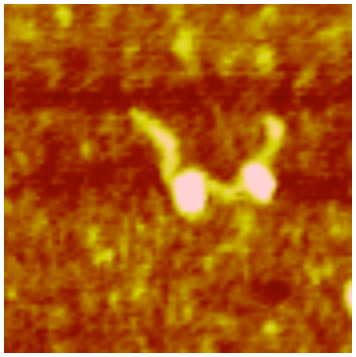
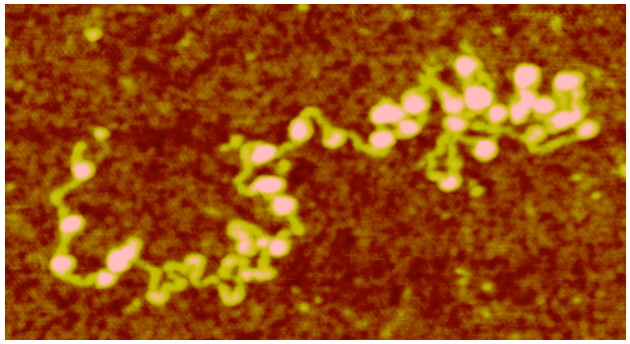
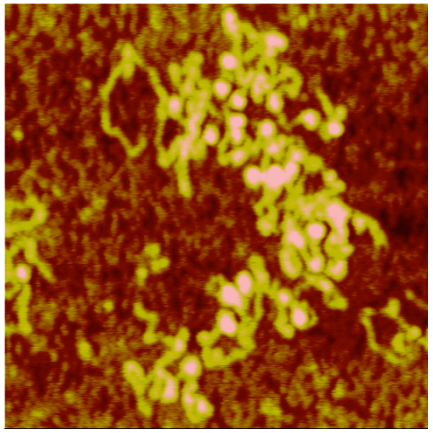
WLC $\ell_p = 2A$ WLC $\frac{1}{\ell_p^{eff}} = \frac{1}{2A} + \frac{2}{\sigma_o^2}$ **GWLC** $l_p = f(A, \sigma_o, H)$



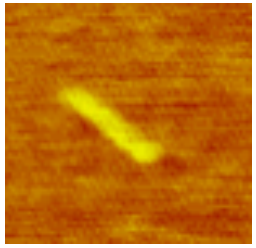
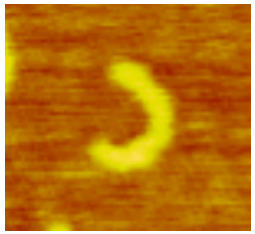
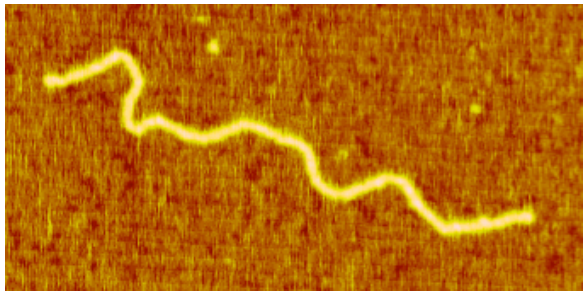
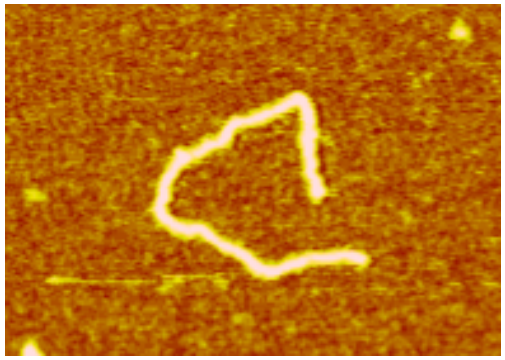


Moukthar et al., Eur. Phys. Lett. (2009); Arneodo et al., Phys. Rep. (2011)

Structure and dynamics of nucleosomes : histone variants, and chromatin remodeling

DNA sequence effects on the structural and mechanical properties of the double helix

DNA sequence effect on the nucleosomal organization of the eukaryotic chromatin fiber

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
Cédric Vaillant

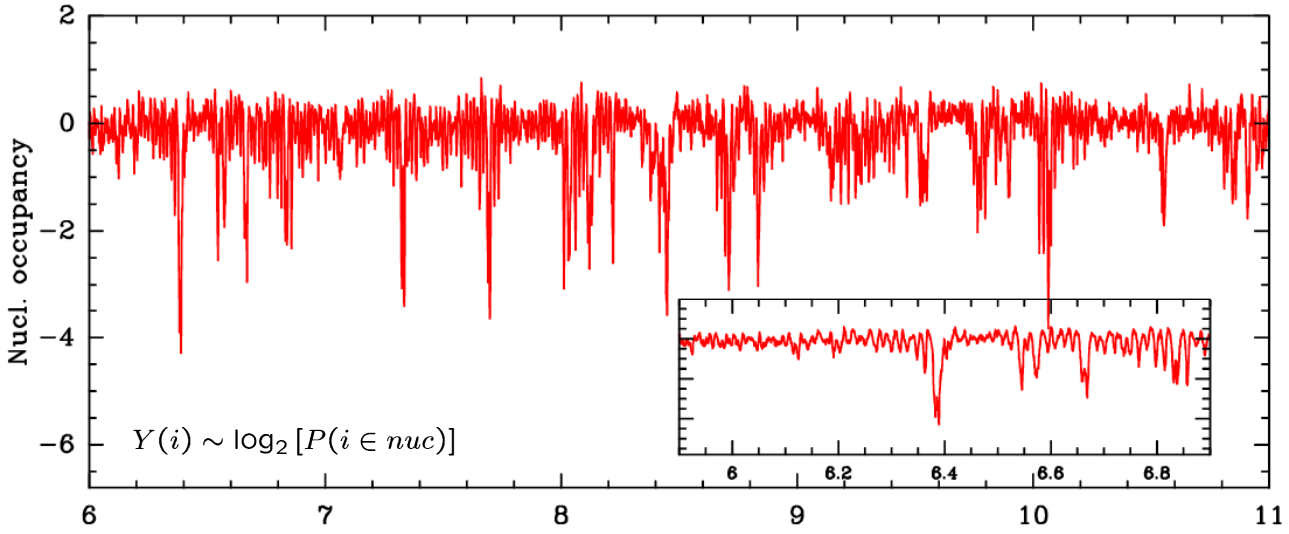
Françoise Argoul
Monique Marilley
Pascale Milani
Philippe Bouvet
Zofia Haftek-Terreau

ENS de Lyon, France

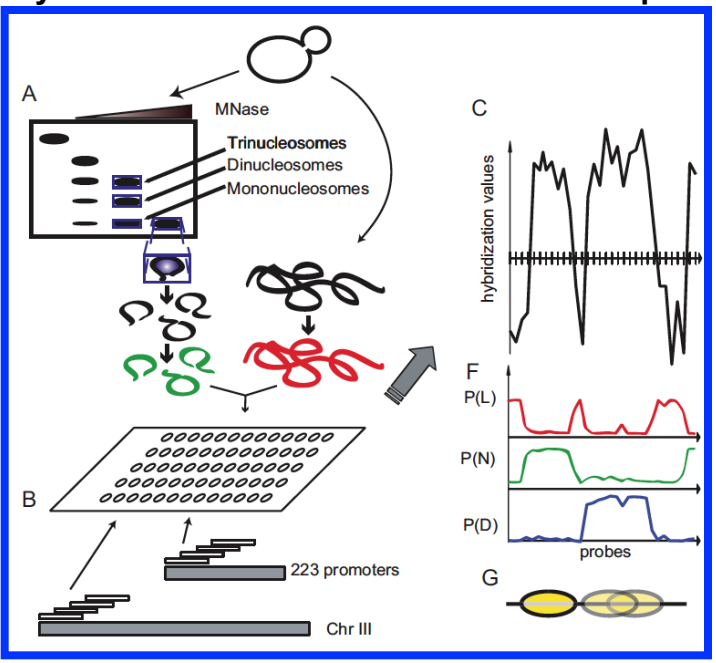
Yves d'Aubenton-Carafa
Claude Thermes

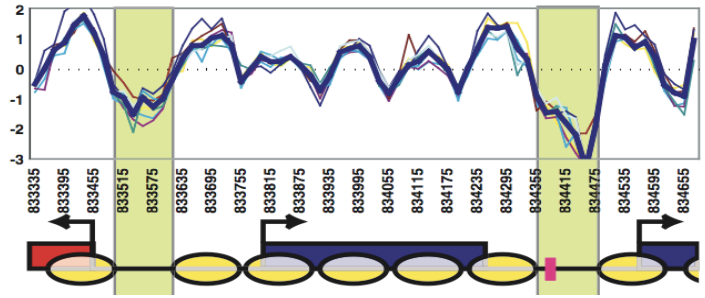
CGM, Gif-sur-Yvette, France

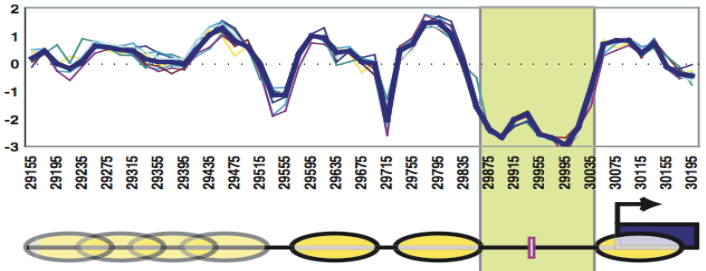
Genome-Scale Identification of Nucleosome Positions in *S. cerevisiae* (Yuan et al., Science 309)



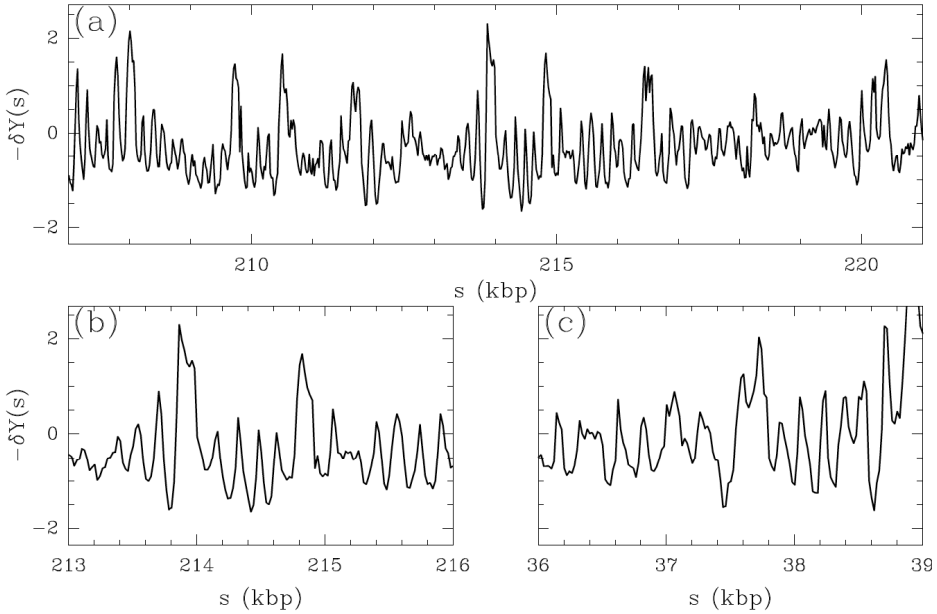
Hybridisation of nucleosomal DNA on chip



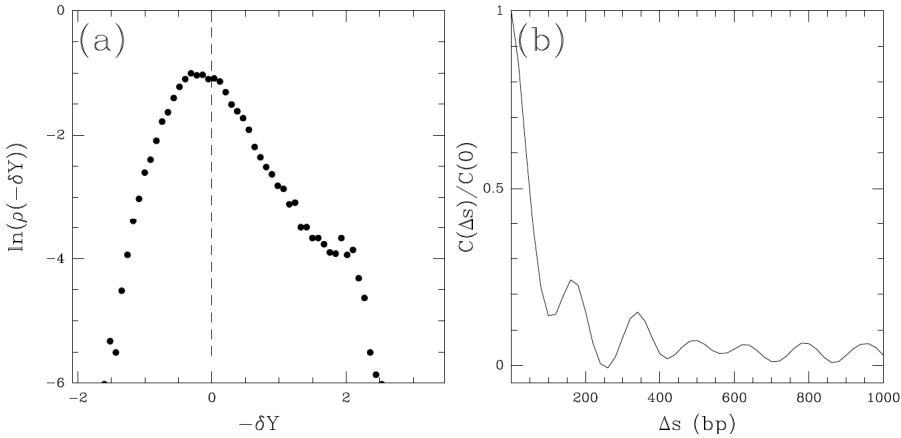


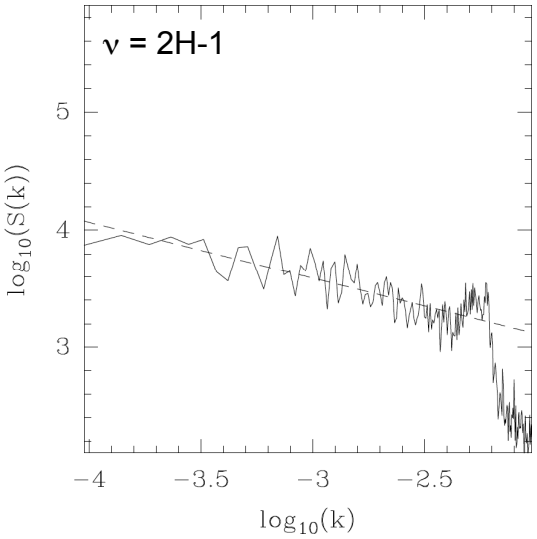


Statistical characterization : I Experiment



Vaillant et al., Phys. Rev. Lett. (2007)





H = 0.74

Large-scale disorder of nucleosome positioning

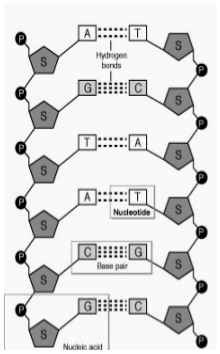
Long-range correlations between 200 and 1000bp with H~0.8

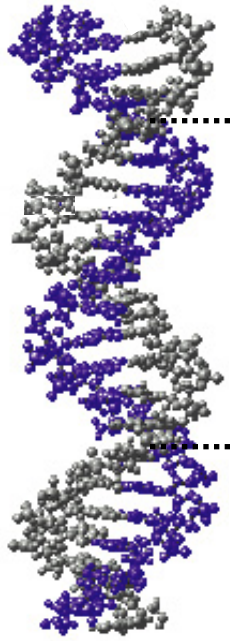
Small-scale periodic component at 160 bp

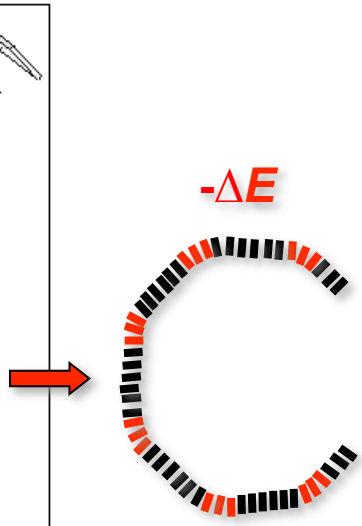
Weak positioning: fat tail (exponential) statistics

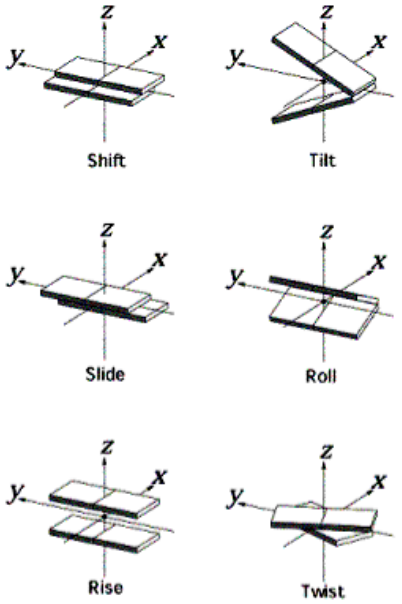
PERIODIC DISTRIBUTION OF DNA BENDING SITES FAVORS NUCLEOSOME FORMATION

Crothers, Travers, Trifonov, Widom





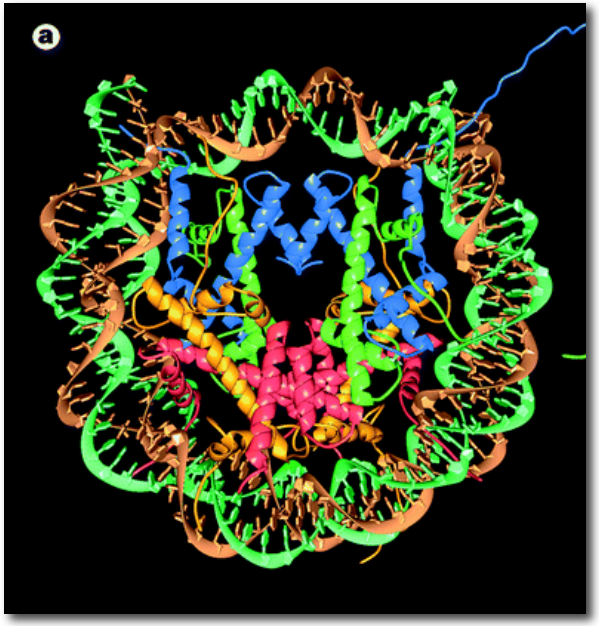




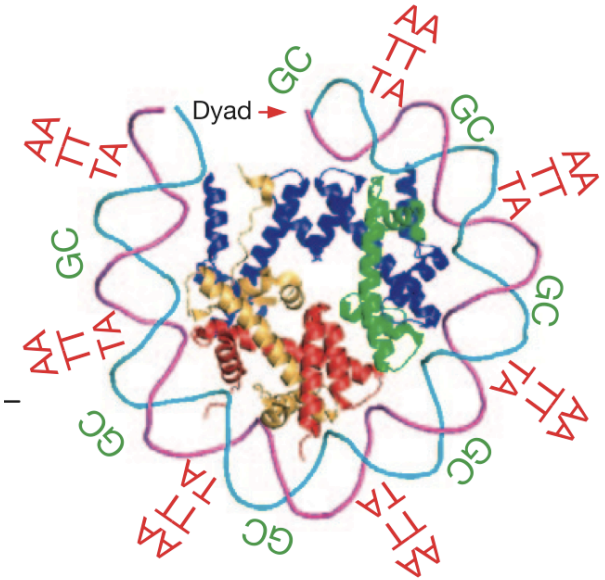
1 helix turn ~10 bp

(c)

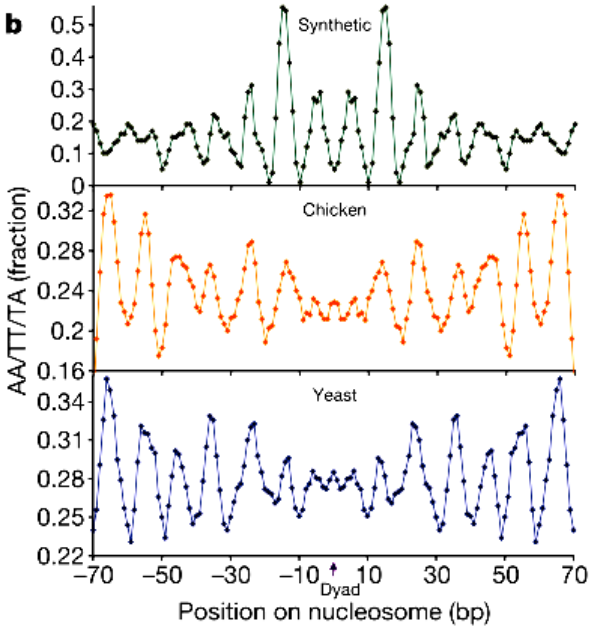
PERIODIC DISTRIBUTION OF DNA BENDING SITES



Luger et al., Nature (1997)



Segal et al., Nature (2006)



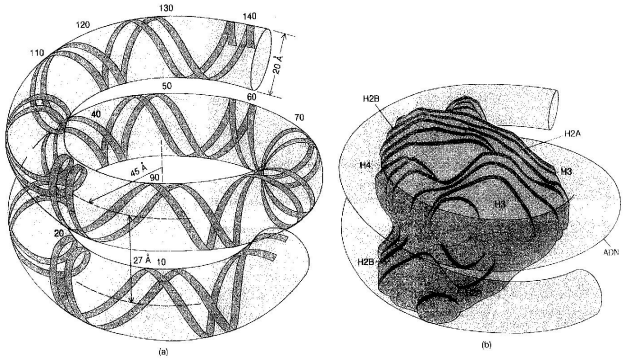
⇨ But it concerns only a small proportion of nucleosomes

(15% according to *Peckham et al., Genome Res. (2007)*)

Trinucleotidic structural tables based on experiments

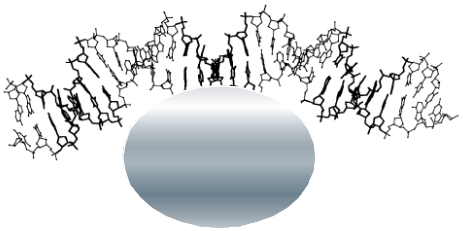
Nucleosome positioning

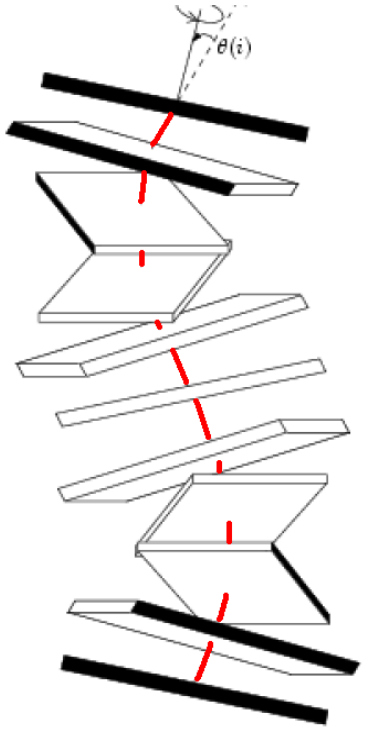
local curvature



Dnase I sensitivity

Local flexibility





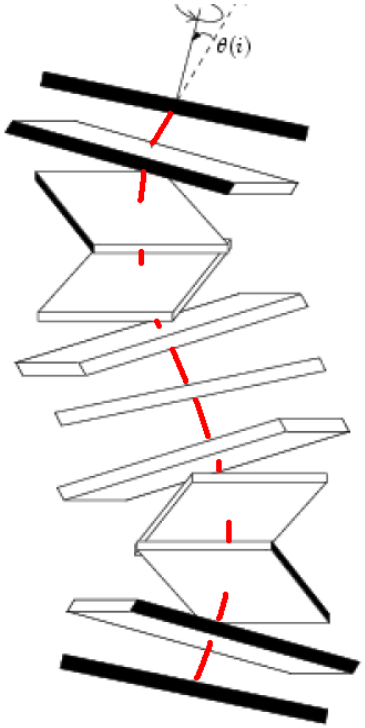
Trinucleotide	PNuc	DNase I
AAA/TTT	0.0	0.1
AAC/GTT	3.7	1.6
AAG/CTT	5.2	4.2
AAT/ATT	0.7	0.0
ACA/TGT	5.2	5.8
ACC/GGT	5.4	5.2
ACG/CGT	5.4	5.2
ACT/AGT	5.8	2.0
AGA/TCT	3.3	6.5
AGC/GCT	7.5	6.3
AGG/CCT	5.4	4.7
ATA/TAT	2.8	9.7
ATC/GAT	5.3	3.6
ATG/CAT	6.7	8.7
CAA/TTG	3.3	6.2
CAC/GTG	6.5	6.8

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

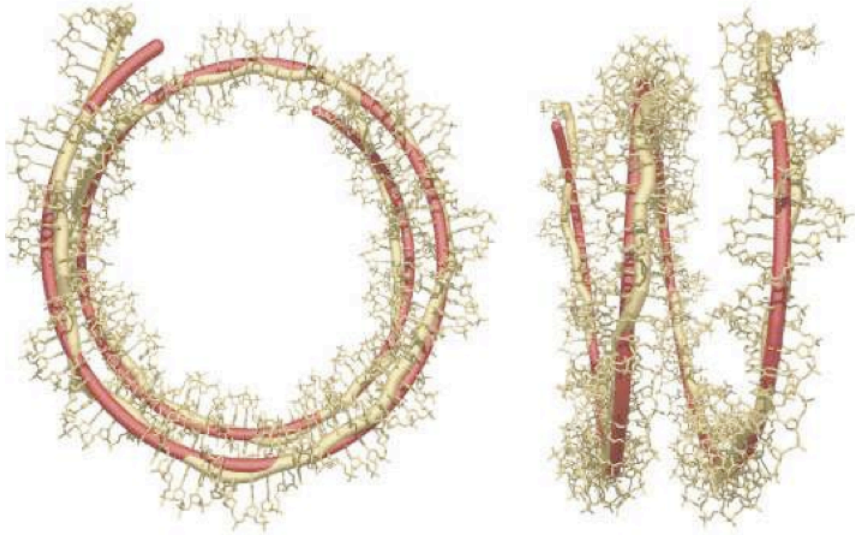
trajectory of the double helix axis

COMPUTING THE FREE-ENERGY NECESSARY TO BEND THE DNA DOUBLE HELIX TO FORM NUCLEOSOMES

Vaillant et al., Phys. Rev. Lett. (2007)



$\Delta E(s)$

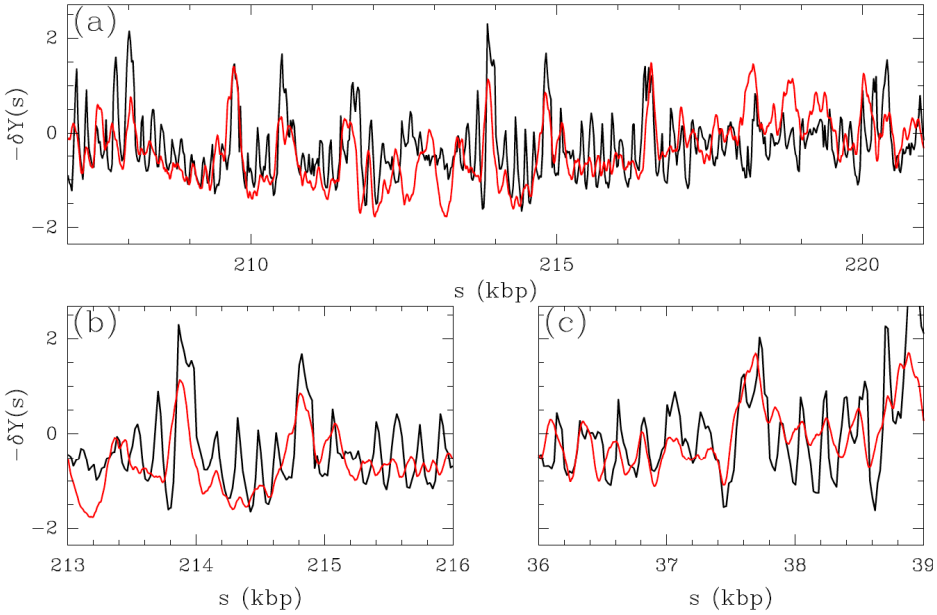


Local polymer structure locally defined by 3 angles: $(\Omega_1(s), \Omega_2(s), \Omega_3(s))$
and 3 flexibilities: $(A_1(s), A_2(s), C(s))$

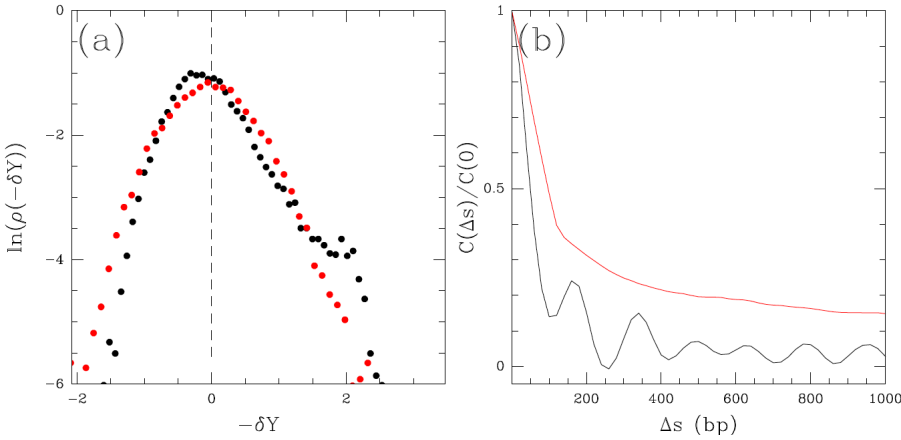
$$\frac{\delta E}{k_B T} = \frac{A_1}{2}(\Omega_1 - \Omega_{o1})^2 + \frac{A_2}{2}(\Omega_2 - \Omega_{o2})^2 + \frac{C}{2}(\Omega_3 - \Omega_{o3})^2 ds$$

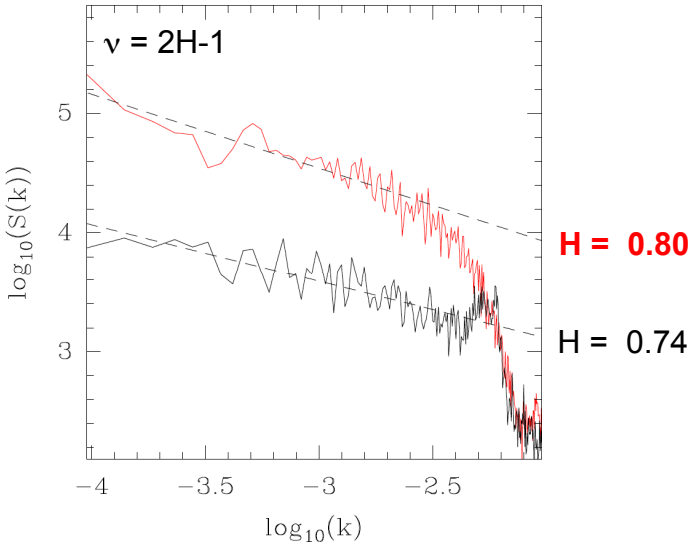
Equilibrium configuration: $(\Omega_{o1}(s), \Omega_{o2}(s), \Omega_{o3}(s)) \neq (0, 0, 0)$
$\rightarrow \Omega_{o3} \sim 34°/\text{pb}.$
$\rightarrow \Omega_{oi}$ vs séquence: structural tables (crystallography, AFM, molecular dynamics)

Statistical characterization : II One nucleosome model



Vaillant et al., Phys. Rev. Lett. (2007)
Miele et al., Nucleic Acids Res. (2008)

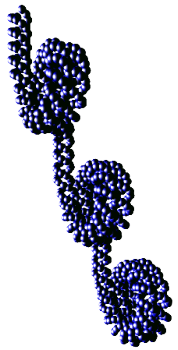




H = 0.80

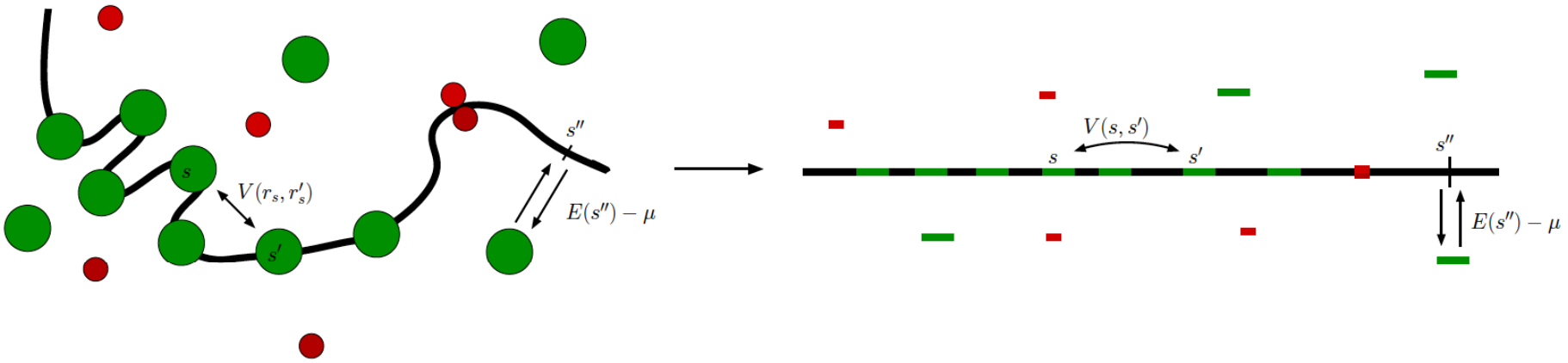
H = 0.74

One nucleosome energy landscape based on a simple sequence dependent helical model



Radius	R = 4.19 nm
Pitch	P = 2.49 nm
Total length	ℓ = 125bp

Modeling sequence effect on nucleosome organisation



chomatin = non uniform fluid of 1D hard-rods

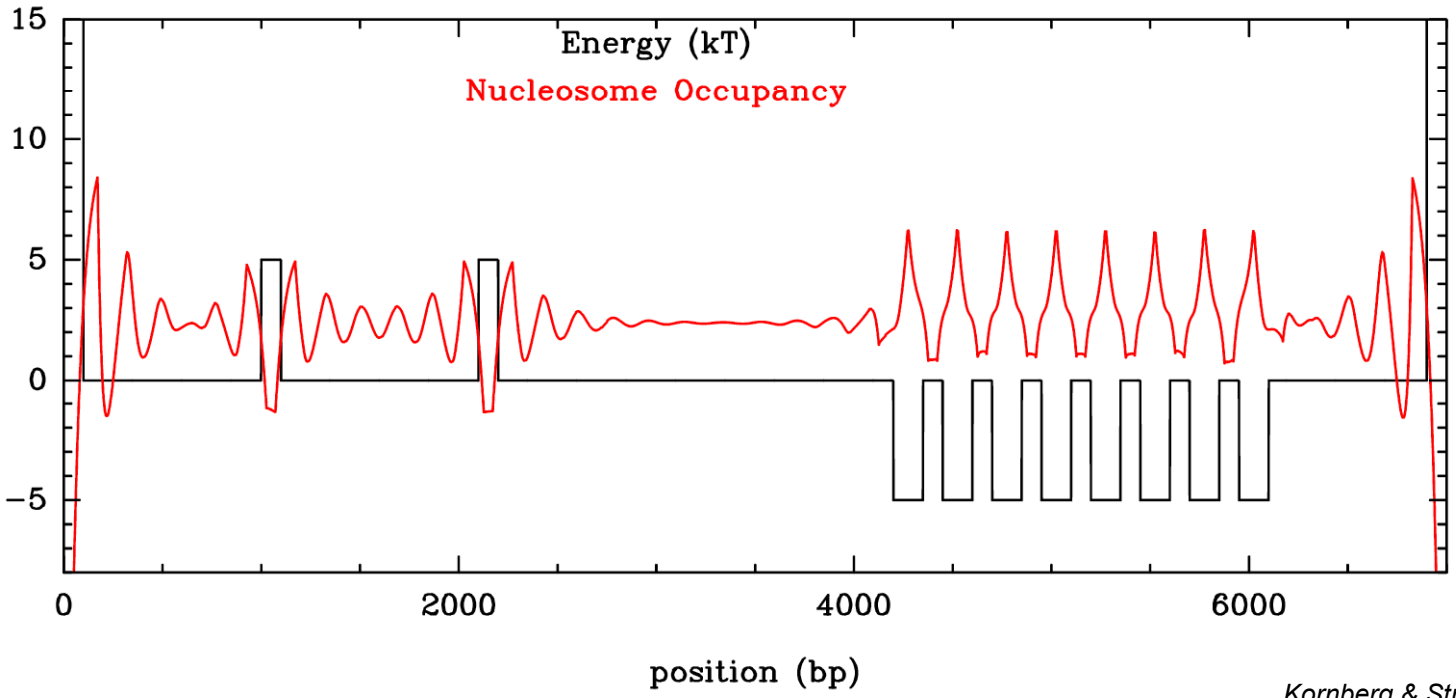
- Dynamics: adsorption, desorption, diffusion.
- Potential energy: non-uniform = sequence dependent $E(s)$.
- Uniform bulk chemical potential μ.
- Interactions: hard-core repulsion.
- Thermodynamical Equilibriium: thermal bath, kT.

Parking phenomenon by excluding energy barriers

Percus relation:

Percus, J. Stat. Phys. (1976)
Vanderlick et al., Phys. Rev. A (1986)

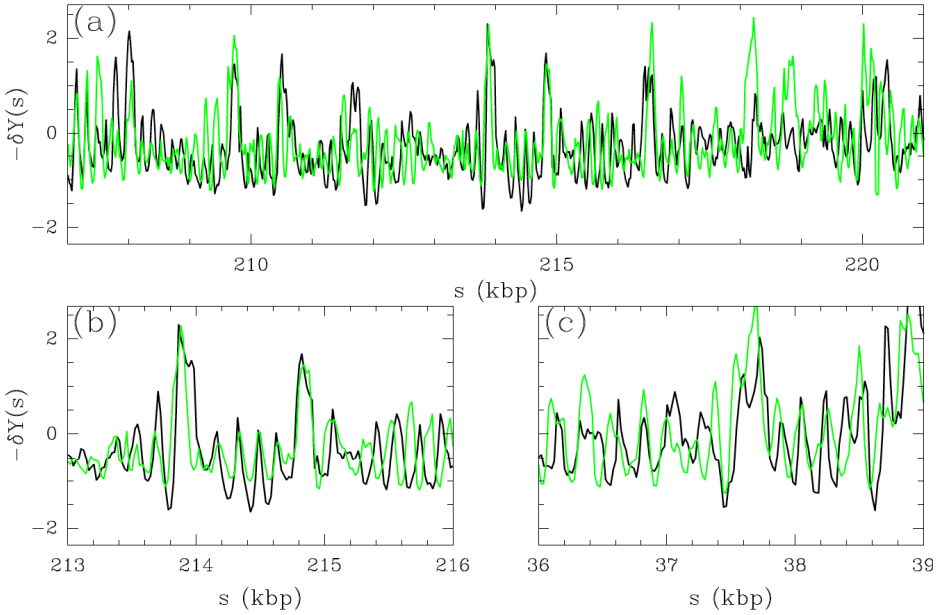
$$\beta\mu = \beta E(s,l) + ln\,\rho(s) - ln\left(1 - \int_s^{s+l} \rho(s')ds'\right) + \int_{s-l}^{s} \frac{\rho(s')}{1 - \int_{s'}^{s'+l} \rho(s'')ds''} ds'.$$



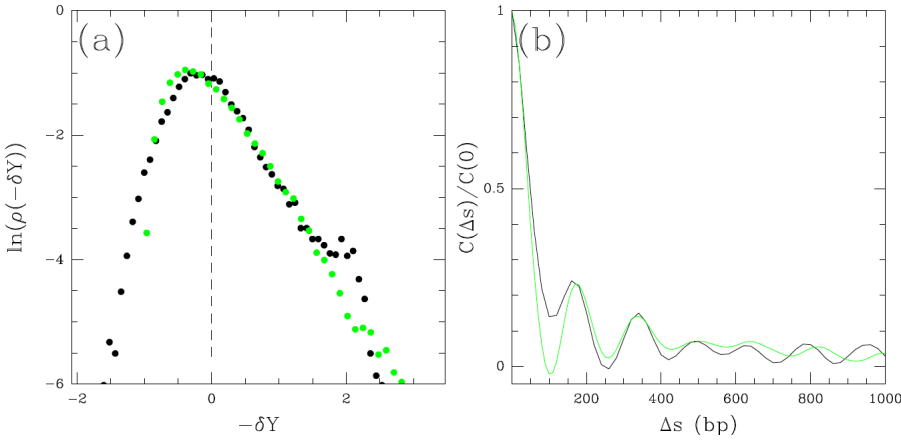
Kornberg & Stryer, Nucleic Acids Res. (1988)

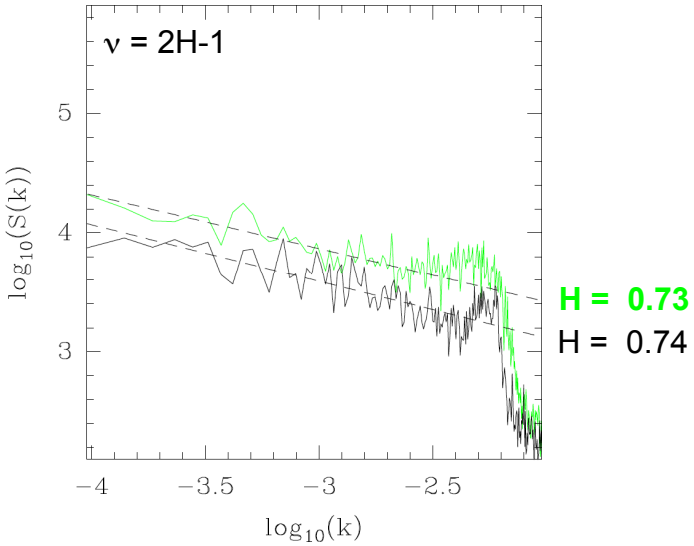
- Long-range statistical ordering by stable excluding barriers
  Barriers: proteins, unfavorable sequence, fixed nucleosome
- Nucleosome energy (barriers) *vs* sequence ?

Statistical characterization: III Many nucleosomes simulations



Vaillant et al., Phys. Rev. Lett. (2007)



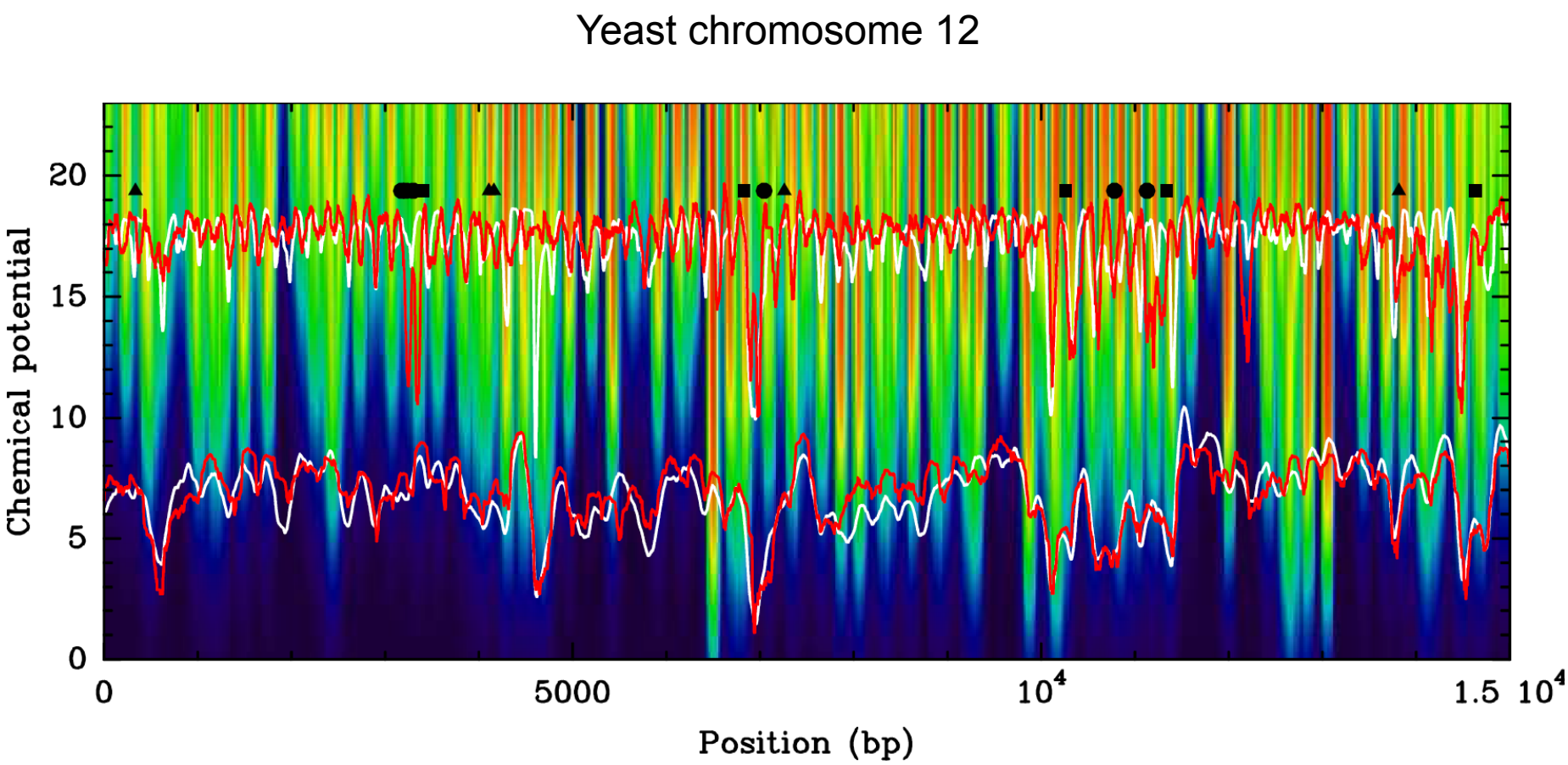


Many nucleosomes Grand Canonical Monte-Carlo simulations of the nucleosomal array

Chemical potential: 1nuc / 200bp

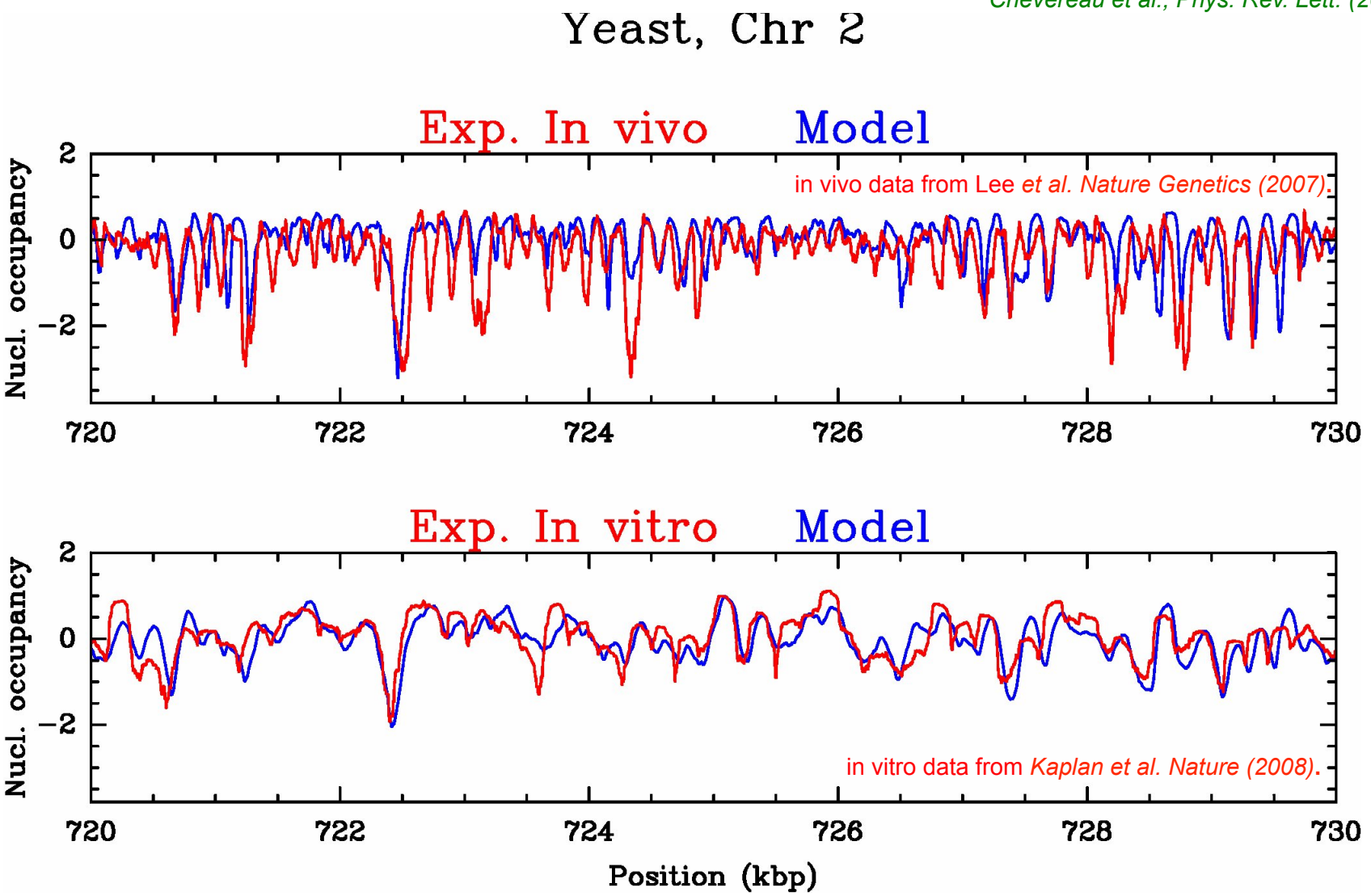
Experimental evidence that long-range correlations influence nucleosomal organisation

Physical modeling of *in vitro* and *in vivo* nucleosome occupancy



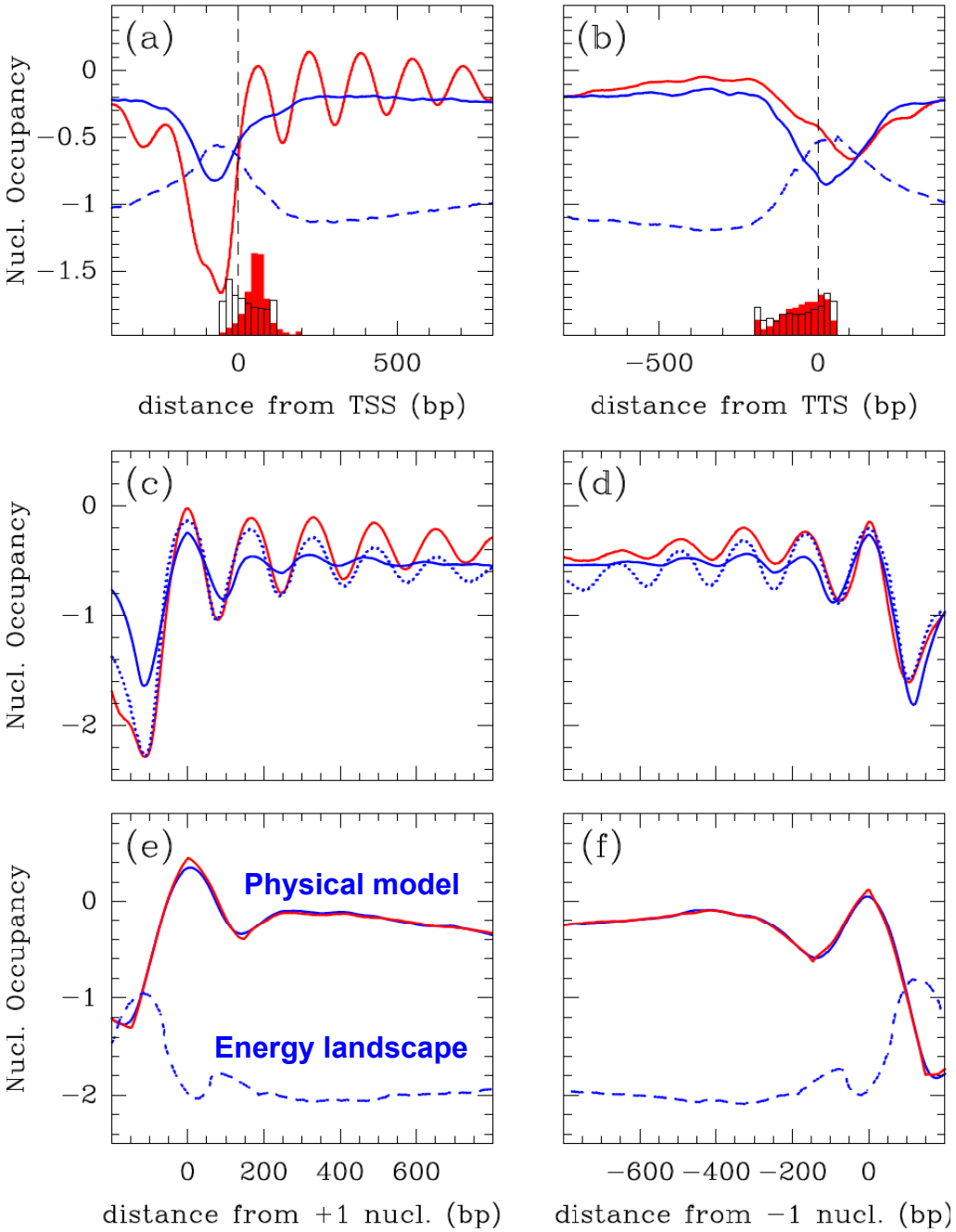
Genome-wide prediction of nucleosome occupancy versus *in vitro* data

Chevereau et al., Phys. Rev. Lett. (2009)



Nucleosome Free Regions at transcription start and end sites

Chevereau et al., Phys. Rev. Lett. (2009)

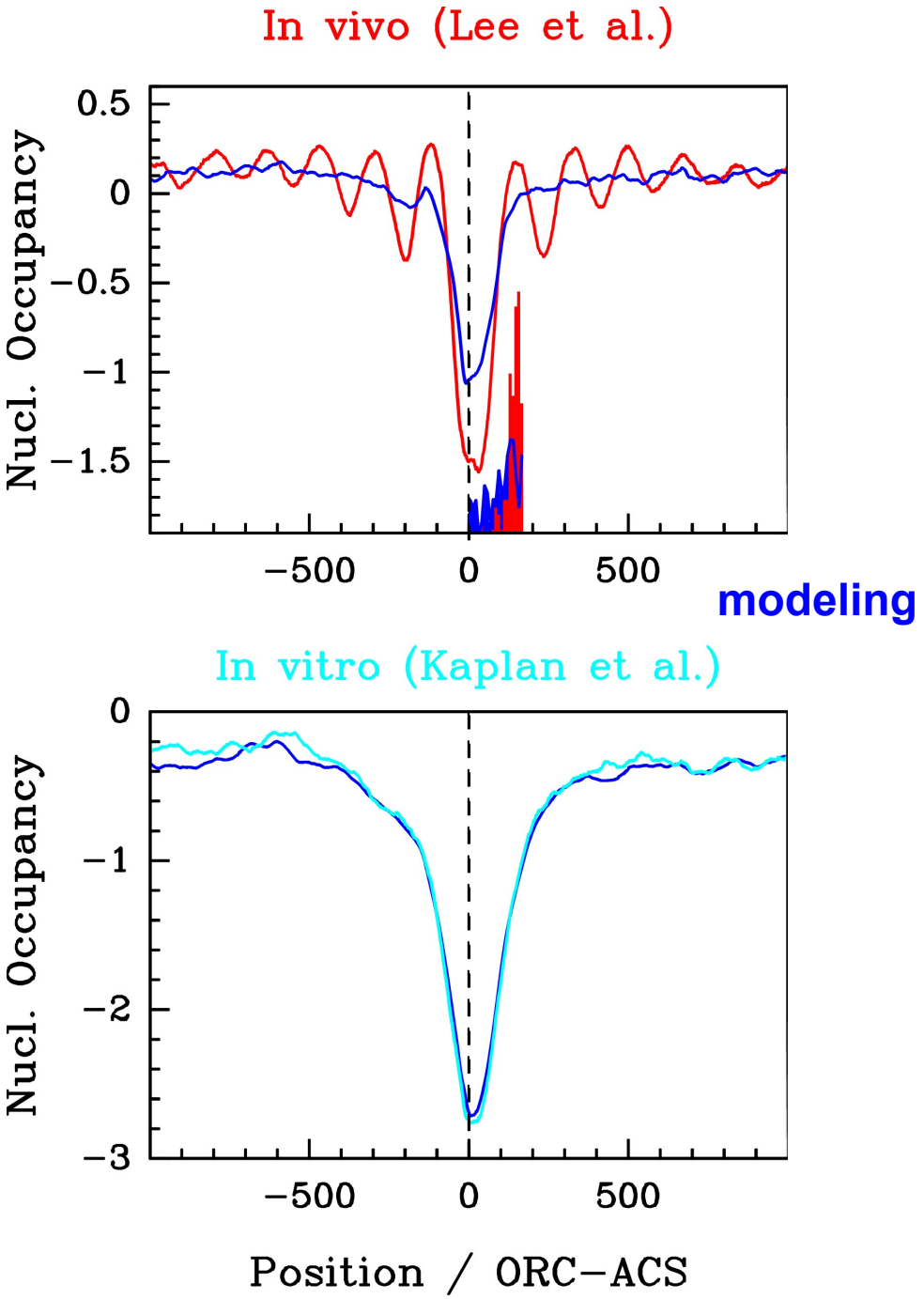


Lee et al., Nat. Genet. (2007)

Kaplan et al., Nature (2009)

ORC binding induces ordering of flanking nucleosomes

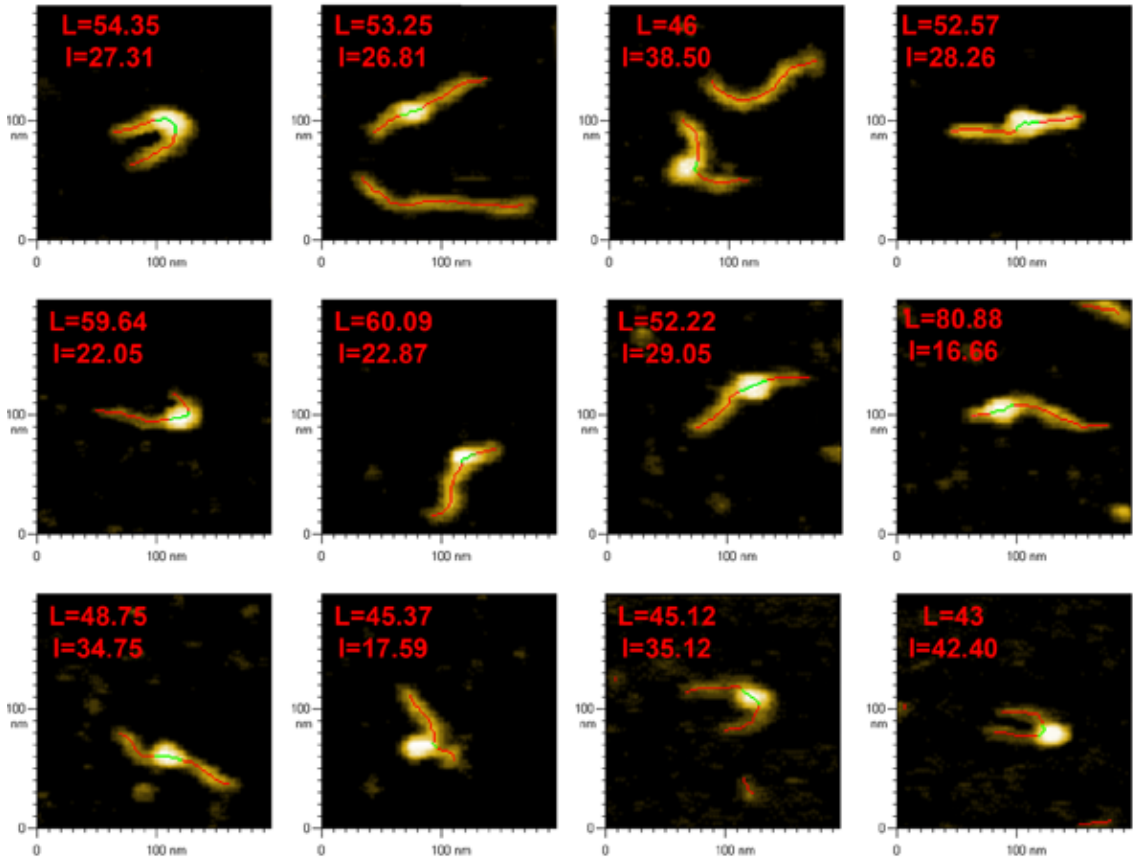
NFR at ORC-ACS is coded in the sequence



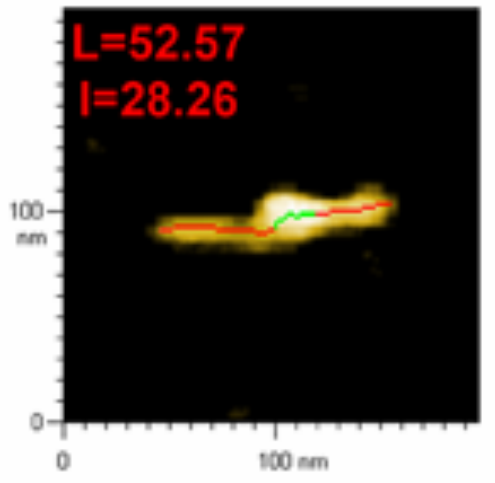
Imaging nucleosome positioning along small DNA fragments using Atomic Force Microscopy in liquid

Experimental evidence of sequence dependent nucleosome positioning

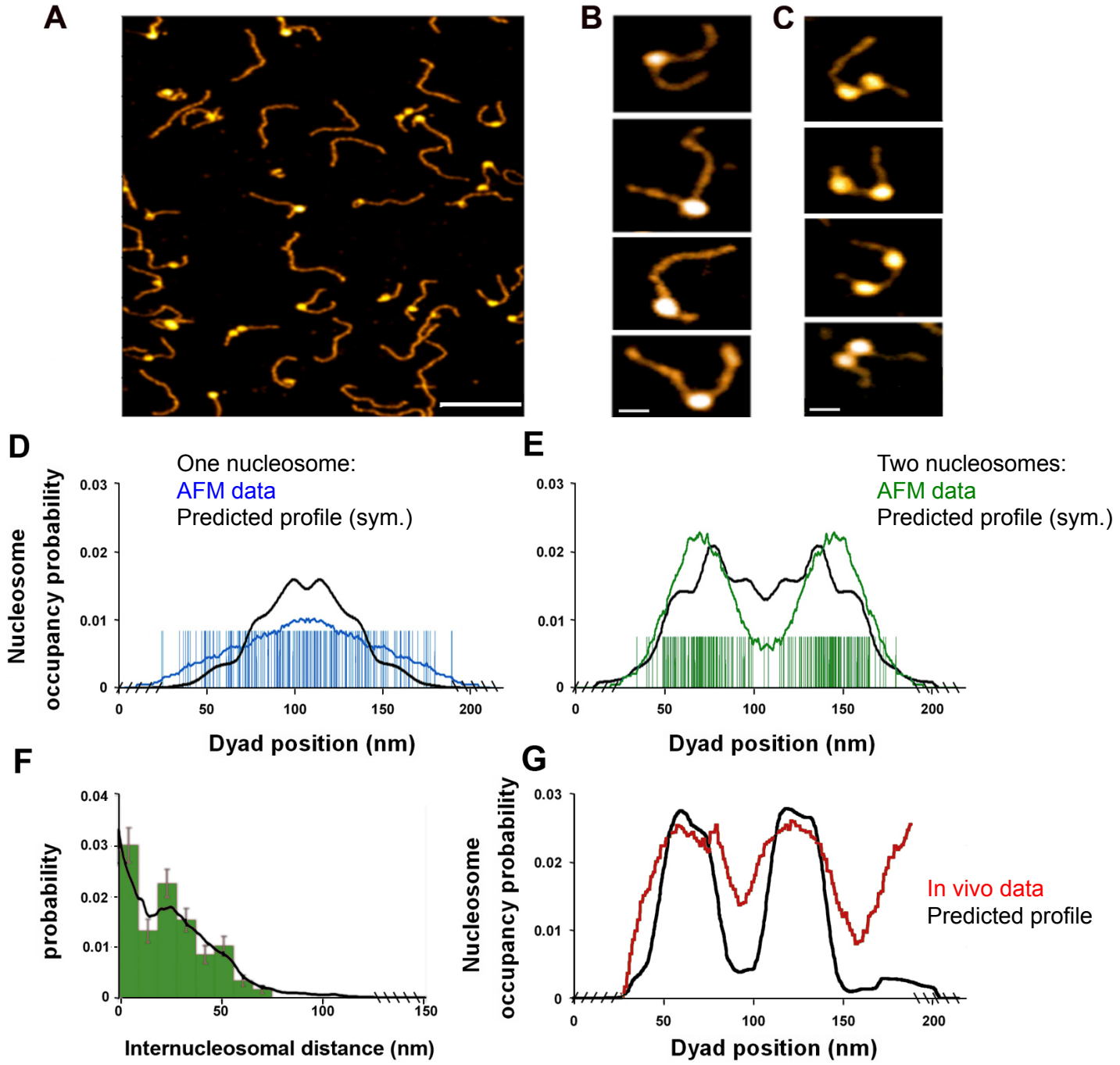
Milani et al., PNAS (2009)



AFM imaging in liquid



Fragment from *S. Cerevisiae* Chr. 7 (595 bp) including gene YGR105W (450 bp).



Physics Reports 498 (2011) 45–188

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

Multi-scale coding of genomic information: From DNA sequence to genome structure and function

Alain Arneodo [a,b,*], Cédric Vaillant [a,b], Benjamin Audit [a,b], Françoise Argoul [a,b], Yves d'Aubenton-Carafa [c], Claude Thermes [c]

[a] Université de Lyon, F-69000 Lyon, France

[b] Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS, Ecole Normale Supérieure de Lyon, F-69007 Lyon, France

[c] Centre de Génétique Moléculaire, CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Accepted 24 September 2010
Available online 8 October 2010
editor: H. Orland

Keywords:
DNA sequence
Chromatin
Nucleosome
Genome organization
Epigenetics
Transcription
Replication
Compositional strand asymmetry
Statistical physics
Hetero-polymer
Generalized worm-like-chain model
Scale-invariance
Multi-fractal
Multi-scale analysis
Wavelet transform
Long-range correlations
Atomic force microscopy

ABSTRACT

Understanding how chromatin is spatially and dynamically organized in the nucleus of eukaryotic cells and how this affects genome functions is one of the main challenges of cell biology. Since the different orders of packaging in the hierarchical organization of DNA condition the accessibility of DNA sequence elements to trans-acting factors that control the transcription and replication processes, there is actually a wealth of structural and dynamical information to learn in the primary DNA sequence. In this review, we show that when using concepts, methodologies, numerical and experimental techniques coming from statistical mechanics and nonlinear physics combined with wavelet-based multi-scale signal processing, we are able to decipher the multi-scale sequence encoding of chromatin condensation–decondensation mechanisms that play a fundamental role in regulating many molecular processes involved in nuclear functions.