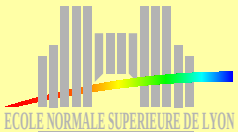
ENS Lyon

Can we predict DNA biological activity from the study of its local fluctuations?

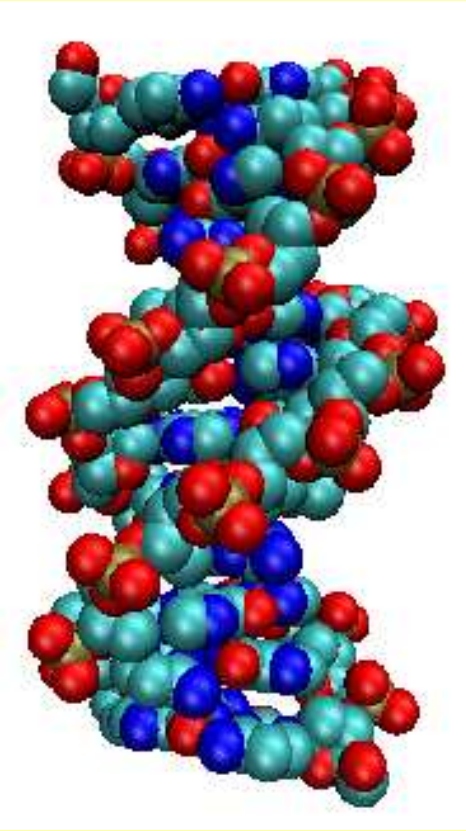
Michel Peyrard
Ecole Normale Supérieure de Lyon, France
Michel.Peyrard@ens-lyon.fr

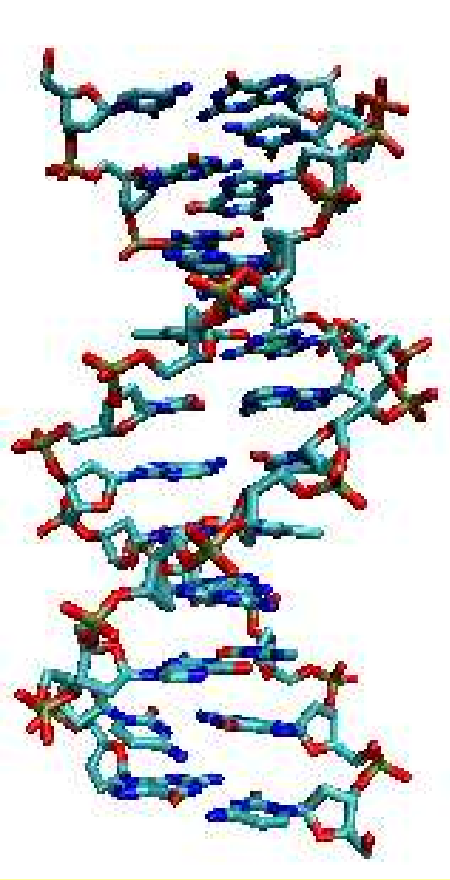
and:
Theory: Titus S. van Erp (Lyon – Leuven) Santiago Cuesta Lopez (Lyon)
Johannes-Geert Hagmann (Lyon – Karlsruhe)
Experiments: Dimitar Angelov (Lyon)

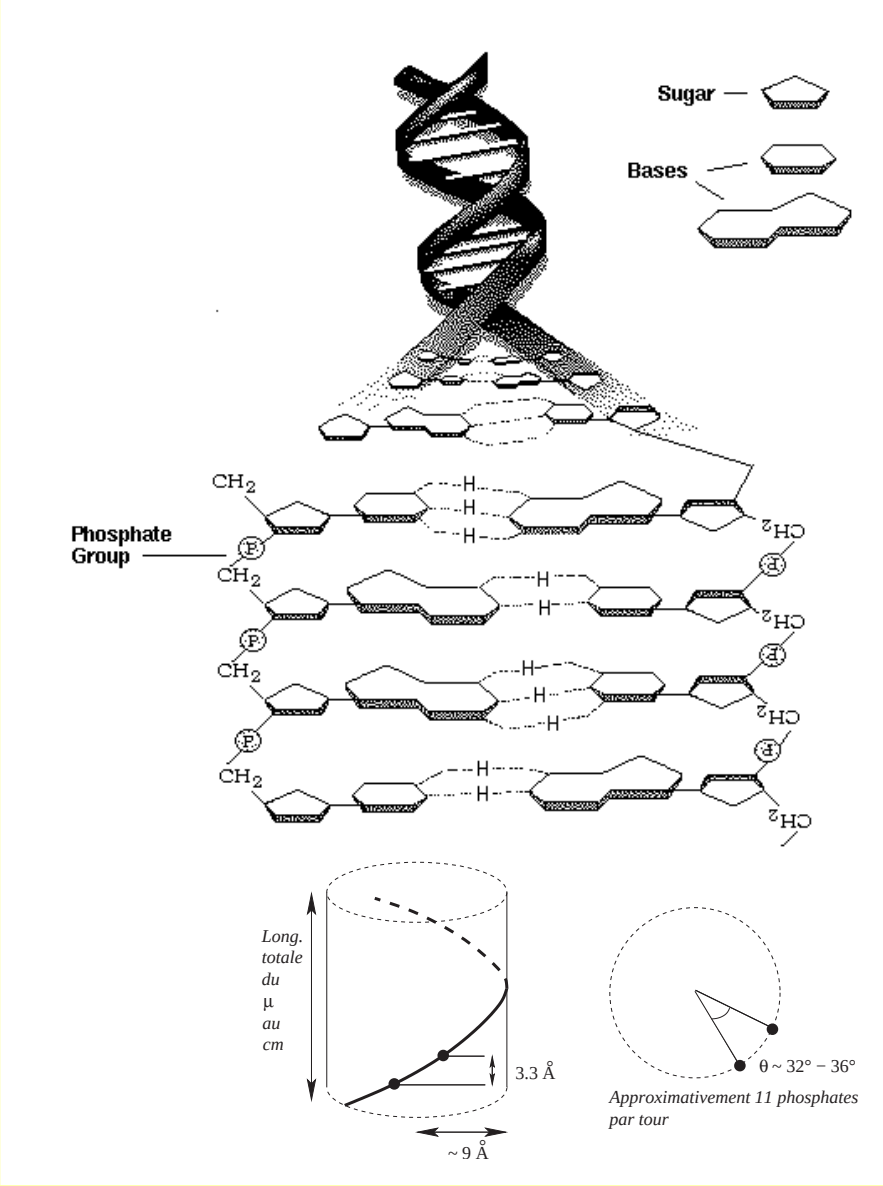
1. A quick reminder of DNA properties
2. Modelling DNA at a mesoscale
3. A biological application: detecting transcription start sites from model studies?
4. Improving models combining experiments and theory (in progress)



Static view: the average structure.







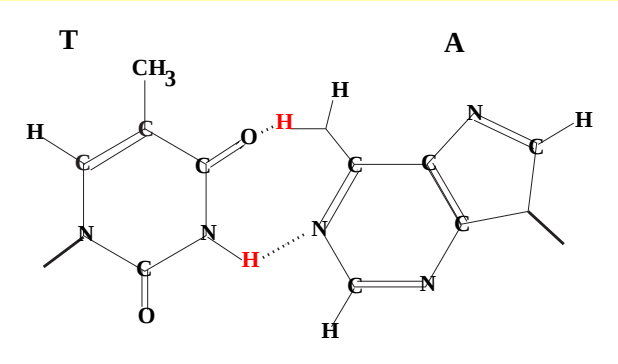
DNA is a highly dynamical entity.

“Breathing of DNA” has been known from biologists for decades.

• Observation: kinetic of imino-proton exchanges ⇒ is the rate limiting step the chemistry of the exchange or fluctuational opening? Use catalysts to discriminate.

Guéron et al. used NMR.
(Gueron et al. Nature **328**, 89 (1987))

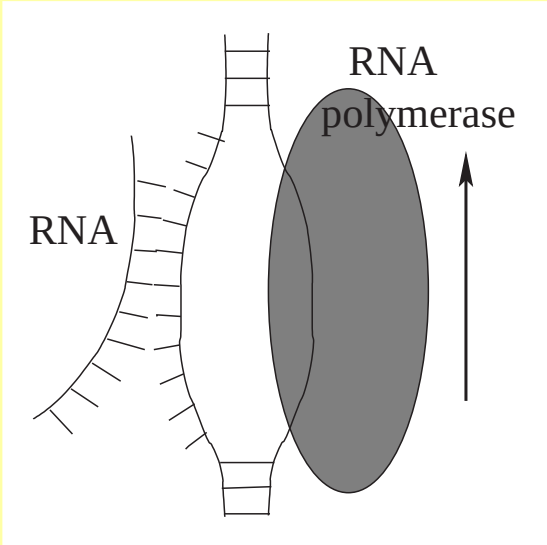
- Lifetime of a base-pair (closed time): 10 ms.
- **Individual** base pair opening (neighbours have different lifetimes)



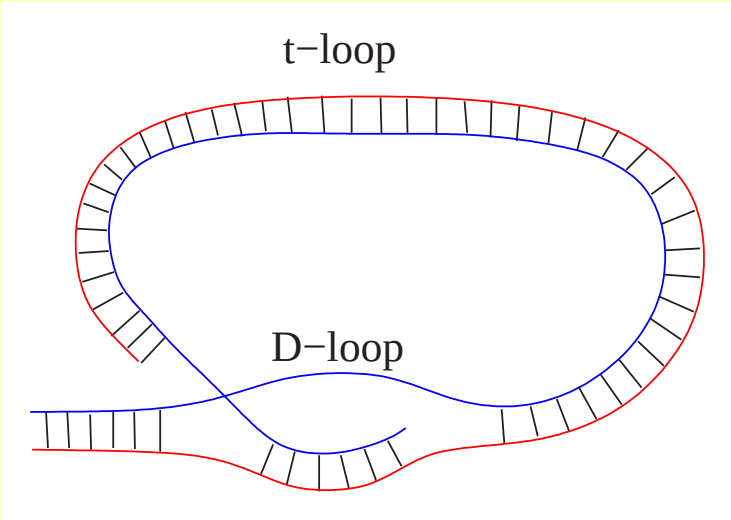
• Local studies of the dynamics: spectroscopy of a bound chromophore ⇒ large anharmonic motions (Cupane et al., Biophy. J. **73**, 959 (1997))

• Also observed in homopolymers (not disorder-induced localisation)

DNA “bubbles” are important for biological function

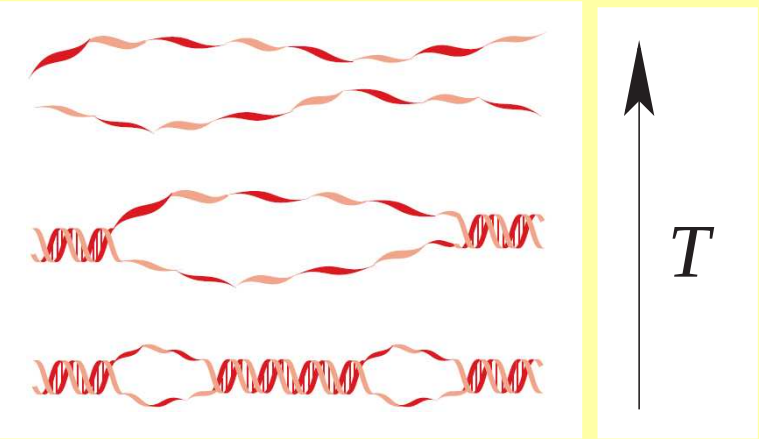


Transcription
(reading a gene)



t–loop of telomers
(protecting chromosome ends)

They can be thermally created (observed in “DNA melting”)

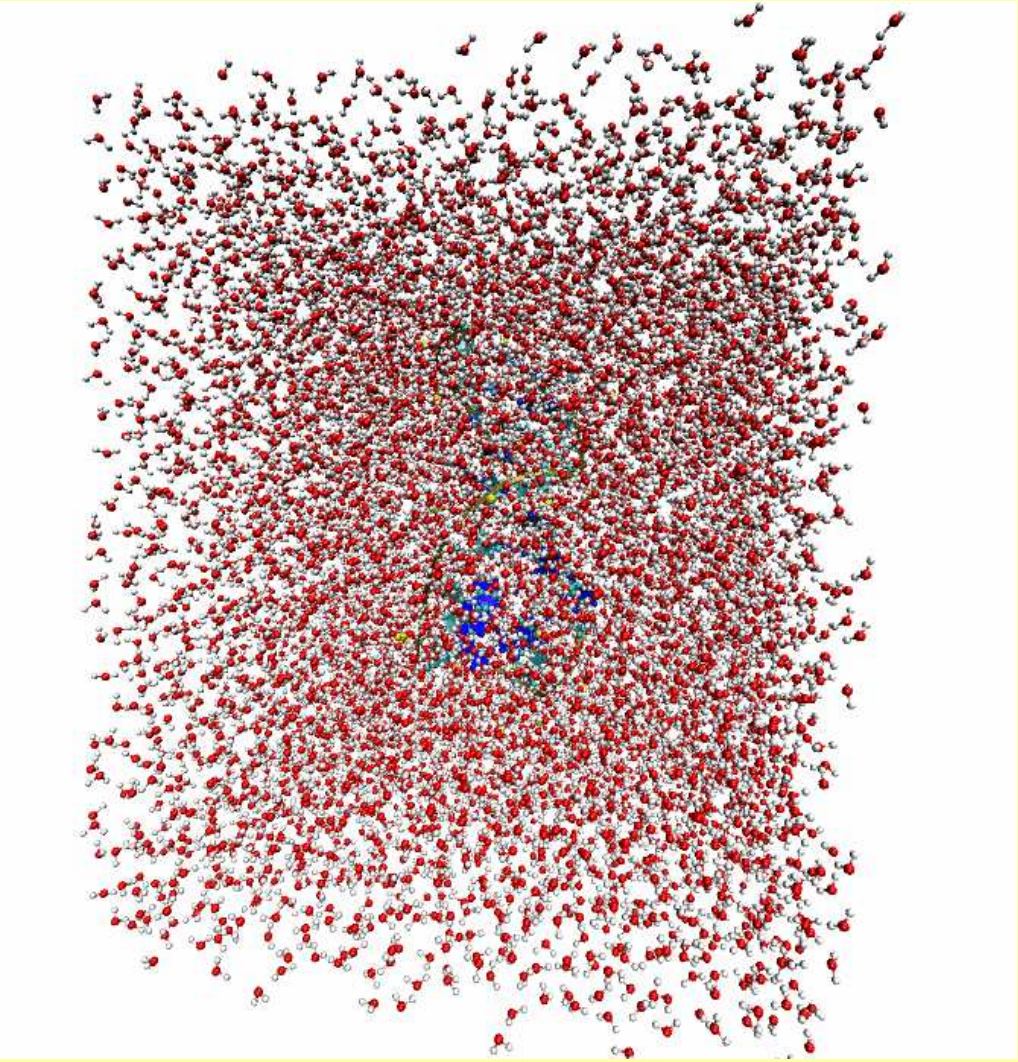


Questions:

- Can we predict bubbles theoretically and compute their sequence dependence?
- Could we use these studies to *predict* the biological activity of some sequences?

Theoretical analysis ⇒ we need a model for the DNA molecule.

All atom studies:



Hopeless for studies needing statistical averages!

Mesoscopic models: at the scale of a base pair

Ising models:

A base pair can be either closed (0) or open (1): adapted to long chains (10000 bases pairs or more).

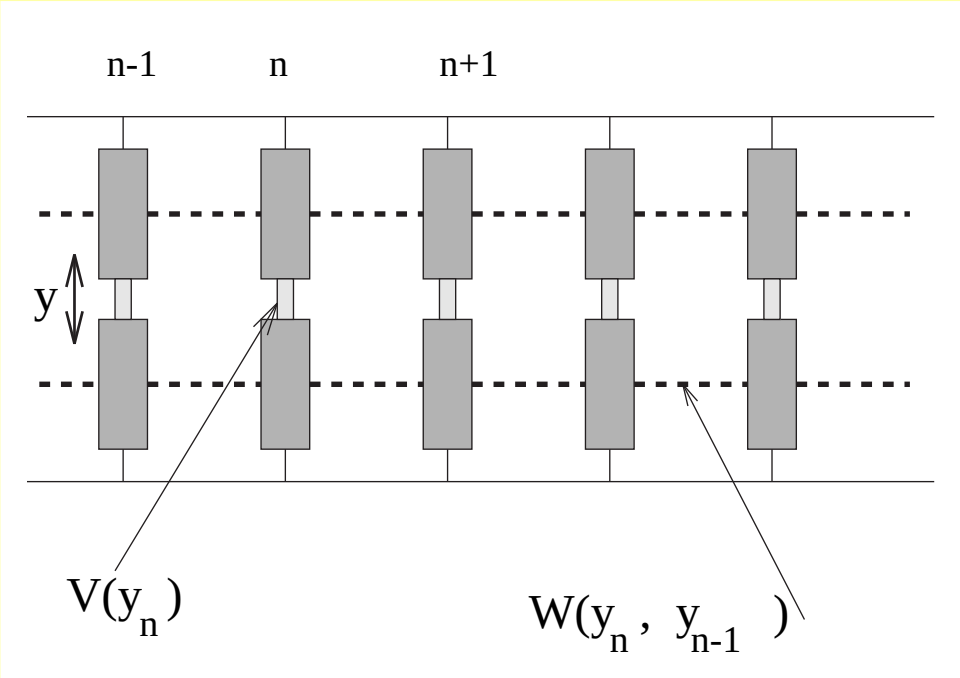
Used by biologists to design sequences for the Polymerase Chain Reaction (PCR – amplification of DNA sequences)
(M. Zuker, *Mfold web server for nucleic acid folding and hybridisation prediction* Nucl. Acids Res. **31** 3406-3415 (2003))

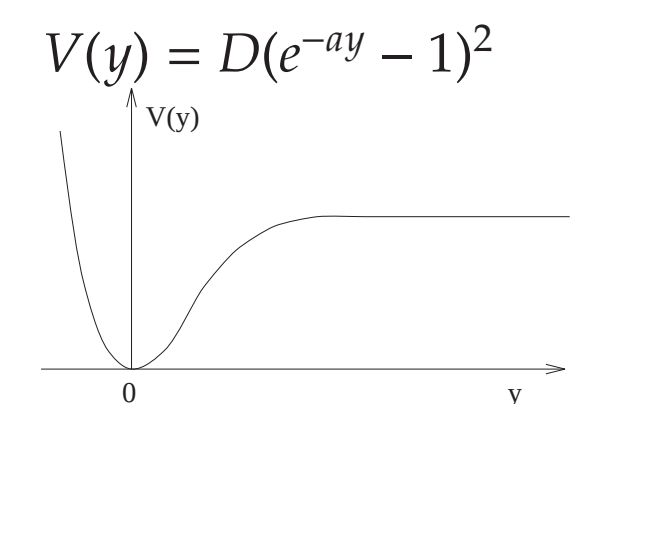
- Many empirical parameters
  (J. SantaLucia, H.T. Allawi, and P. Ananda Senevirante, *Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability* Biochemistry **35** 3555-3562 (1996))

- Not suitable to study dynamics, or local fluctuations.

A dynamical model (M. Peyrard and A.R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989))

Beyond Ising: describe dynamics → real variable y for base-pair stretching

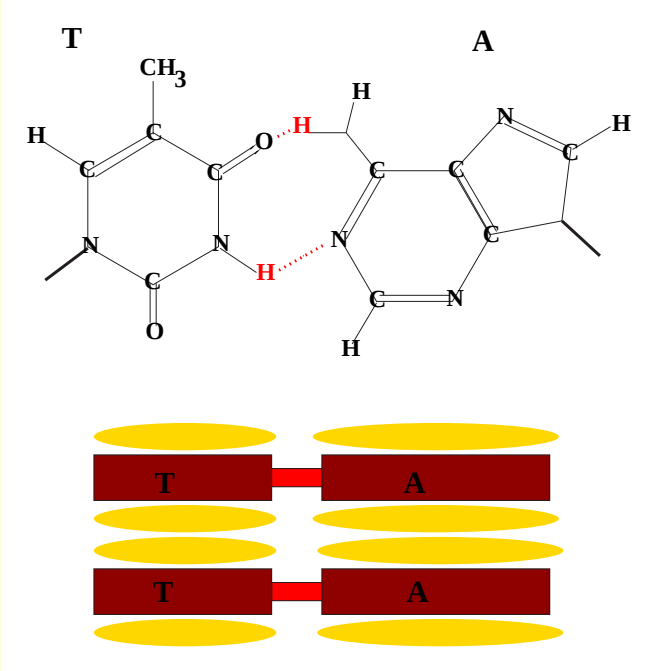




DNA structure only encoded in the potentials

$$H = \sum_n \left[\frac{1}{2} m \left(\frac{dy_n}{dt} \right)^2 + W(y_n, y_{n-1}) + V(y_n) \right]$$

Coupling potential W (stacking) : crucial for the statistical mechanics
(T. Dauxois, M. Peyrard and A.R. Bishop, Phys. Rev. E **47**, R44-R47 (1993))



$$W(y_n, y_{n-1}) = \frac{1}{2}K\left[1+\rho e^{-\alpha(y_n+y_{n-1})}\right](y_n - y_{n-1})^2$$

if y_n **and** $y_{n-1} \ll 1/\alpha \Rightarrow K' \approx K(1+\rho)$
if y_n **or** $y_{n-1} \gg 1/\alpha \Rightarrow K' \approx K$

Model parameters:

Sequence introduced **at the level of the Morse potential** (GC → 3 hydrogen bonds, AT → 2 hydrogen bonds)

One “standard” parameter set determined by fitting of denaturation curves of short DNA sequences (A. Campa and A. Giansanti, Phys. Rev. E **58**, 3585-3588 (1998))

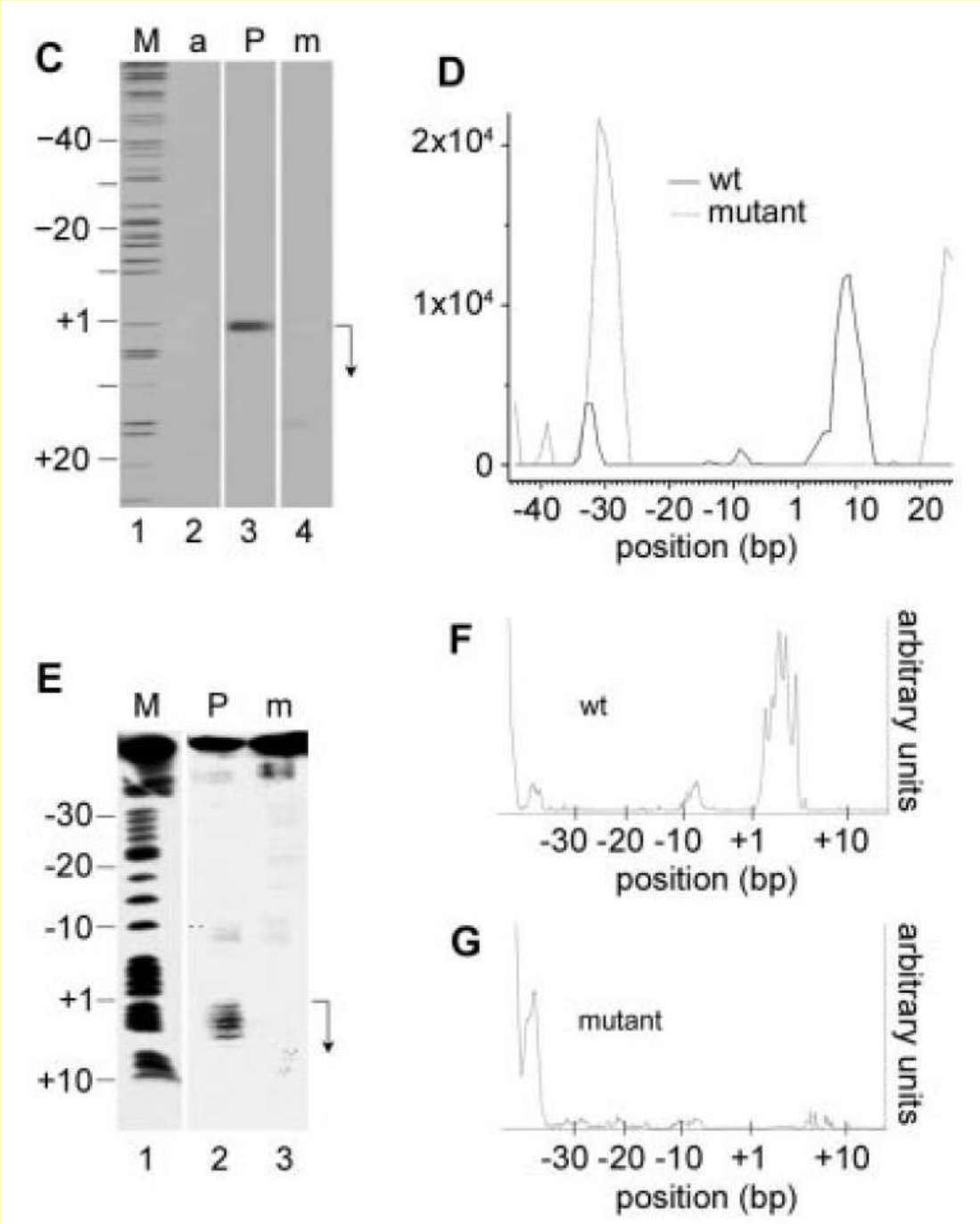
Is it relevant for biology?

Dynamical analysis of the base sequence: C.H. Choi, G. Kalosakas, K. O Rasmussen, M. Hiromura, A. Bishop and A. Usheva, Nucleic Acid Res. **32**, 1584-1590 (2004)

A -46 -21
GTGGC CATTTAGGG TATATATGGCC GAGTGAGCGA
GCAGGATCTC CATTTTGACC GCGAAATTTG AACG
+1 +23

B -46 -21
GTGGC CATTTAGGG TATATATGGCC GAGTGAGCGA
GCAGGATCTC CGCTTTGACC GCGAAATTTG AACG
+1 +23

Summary: The "secret" of the transcription start site is not in the protein that reads the code but in a characteristic of DNA itself.



The method of Choi et al.'s study:

Generalise the model to heterogeneous DNA chains: introduce the sequence.

Interactions within a base pair differ for AT (2 hydrogen bonds) and GC (3 hydrogen bonds)

$$V_n(y_n) = D_n[e^{-\alpha_n y_n} - 1]^2$$

Stacking interaction still assumed homogeneous

$$W(y_n, y_{n-1}) = \frac{1}{2}K\left[1 + \rho e^{-\delta(y_n + y_{n-1})}\right](y_n - y_{n-1})^2$$

Parameters deduced from experimental denaturation curves
(A. Campa and A. Giansanti, Phys. Rev. E **58**, 3585 (1998)):

$D_{AT} = 0.05\,\text{eV}$ $\quad D_{GC} = 0.0755\,\text{eV}$
$\alpha_{AT} = 4.20\,\text{Å}^{-1}$ $\quad \alpha_{GC} = 6.90\,\text{Å}^{-1}$
$K = 0.025\,\text{eV/Å}^2$ $\quad \rho = 2.0$ $\quad \delta = 0.35\,\text{Å}^{-1}$

Use molecular dynamics simulations and ***count as "open" states that belong to a bubble of size 10***

But . . . Molecular dynamics provide very bad statistics to study large bubbles.

⇒ **Compute bubble opening probabilities from statistical physics.**
(T. S. van Erp, S. Cuesta-Lopez, J.-G. Hagmann, M. Peyrard, Phys. Rev. Lett. (2005))

$\theta_i(y_i) = \theta(y_i - y_0), \qquad \bar{\theta}_i(y_i) = \theta(y_0 - y_i)$ → 1 if base-pair i is open, 0 otherwise

$$\theta_i^{[m]} \equiv \bar{\theta}_{i-\frac{m}{2}} \bar{\theta}_{i+\frac{m}{2}+1} \prod_{j=i-\frac{m}{2}+1}^{i+\frac{m}{2}} \theta_j \text{ for } m \text{ even}$$

(and a related definition for m odd)
→ 1 for each base pair in the middle of a bubble of **exactly** size m

$$\left\langle \theta_i^{[m]} \right\rangle_\mu \equiv \frac{\left\langle \theta_i^{[m]} \mu \right\rangle}{\langle \mu \rangle} \equiv \frac{Z_{\theta_i^{[m]}}}{Z - Z_\Pi} \quad \text{with} \quad \mu = 1 - \prod_{i=1}^{N} \theta_i$$

Probability to have bubble of **exactly** size m at site i

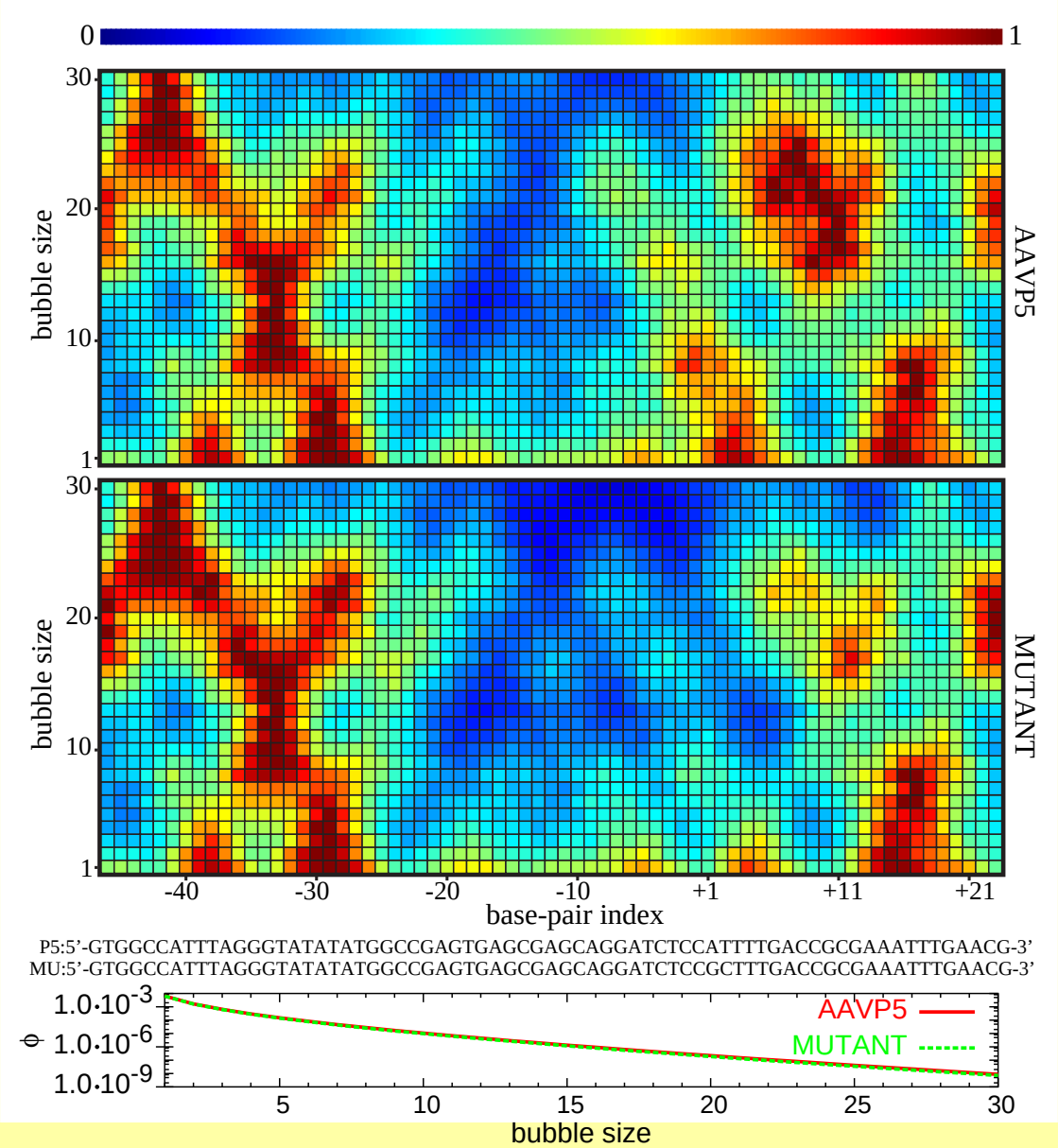
1D allows exact numerical calculation (factorisable form) → efficient iterative algorithm.

Model does predict openings where experiments find them, but

- Transcription start site is not the most preferential bubble site
- The mutation only has a local effect.

The model does not predict sites with biological activity!

Normalised bubble probability matrix



The question is still there: can theory predict biological activity ?

Two possibilities

1. DNA *does not* direct its own transcription as expected (and suggested by experiments)

2. *It does* but the model is not good enough

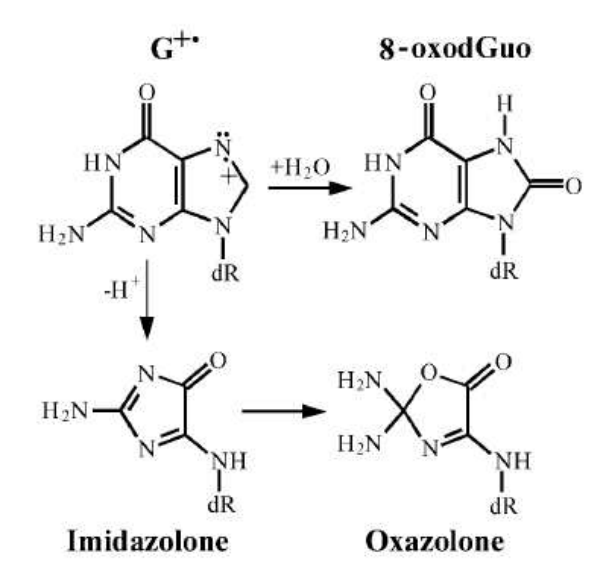
$\Rightarrow$ An improved model is required.

It must be based on **local** measures of DNA fluctuations.

UV Laser oxidative Guanine modifications:
a probe of local stacking fluctuations.

Dimitar Angelov, ENS Lyon (with A. Spassky)

High-intensity UV laser pulses → guanine oxidative lesions strongly modulated by the local structure.



Oxozalone sensible to cleavage by piperidine glycosylase

8-oxodG sensible to cleavage by Fpg protein, and appears only in helicoidal stacking

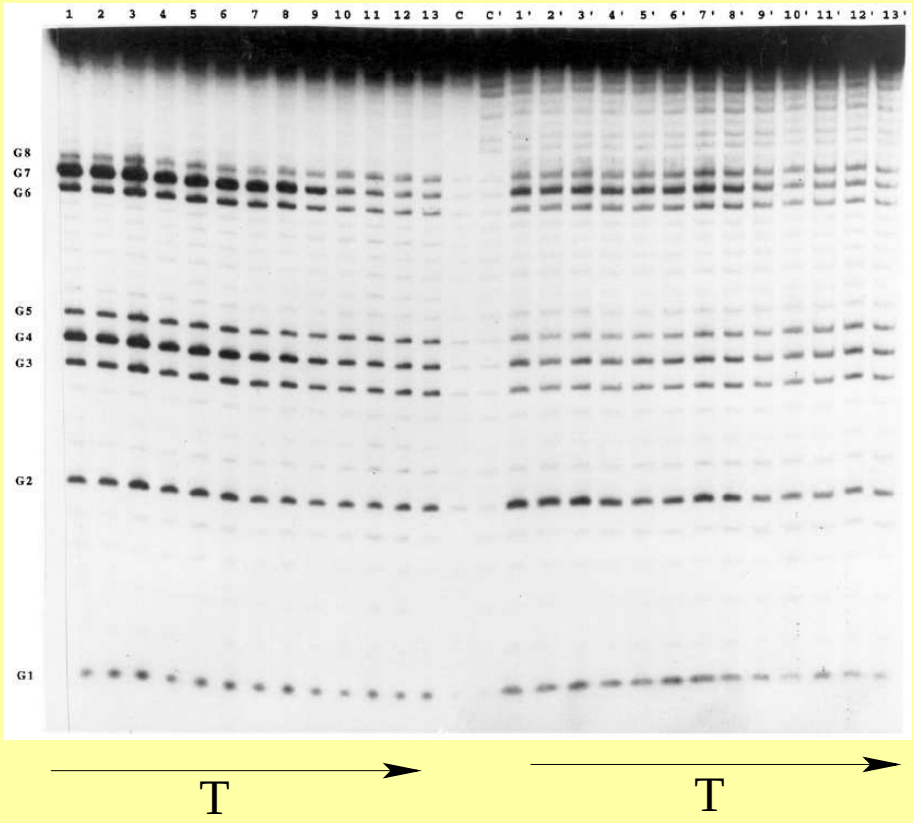
→ R_{pip}/R_{FpG} indicative of local opening

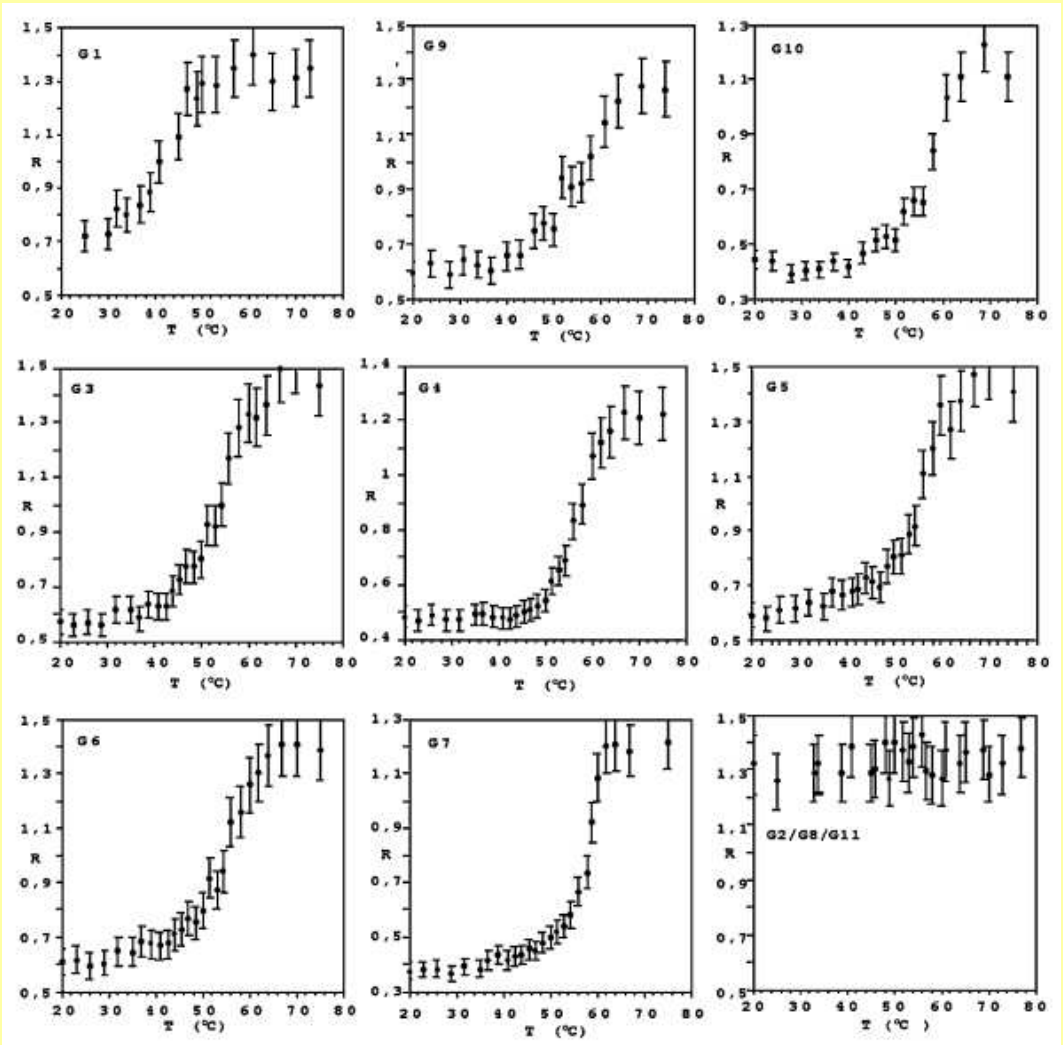
Opens the possibility to evaluate fluctuational opening versus sequence.

From A. Spassky and D. Angelov, J. Mol. Biol. **323** 9 (2002)

```
                1     2    345         678
5'-ATAAAAGTTACGCTAGGGATAACAGGGTAATATAACG-3'
3'-TATTTTCAATGCGATCCCTATTGTCCCATTATATTGC-5'
                  11 10          9
```

FpG protein hot piperidine

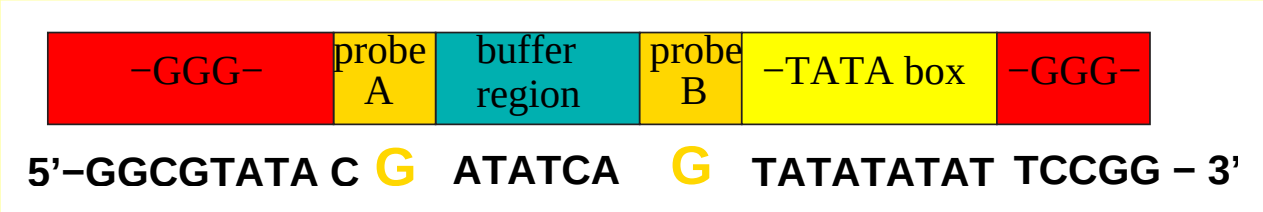




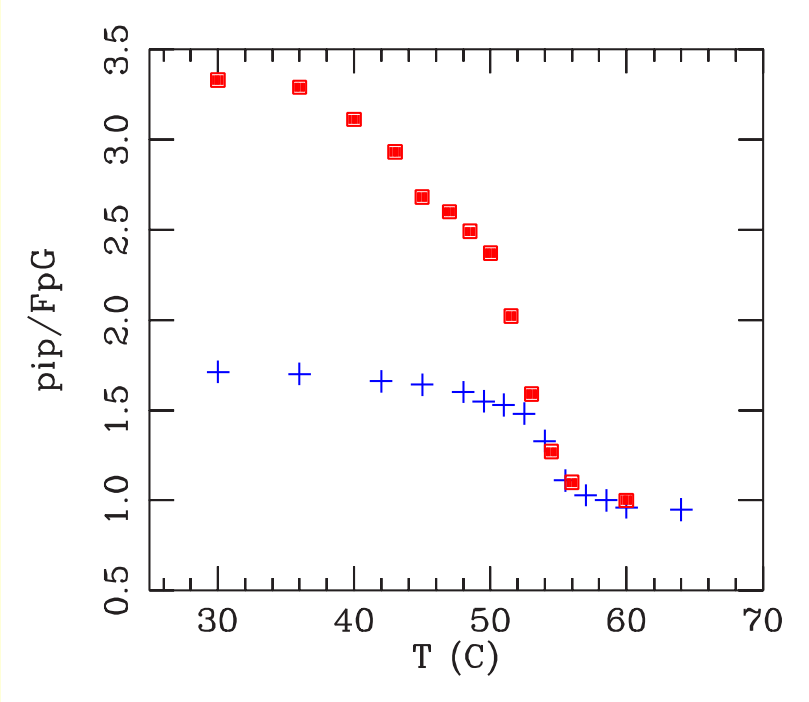
Provides data for all *GC* pairs, contrary to fluorophore–quencher methods.

A first observation:

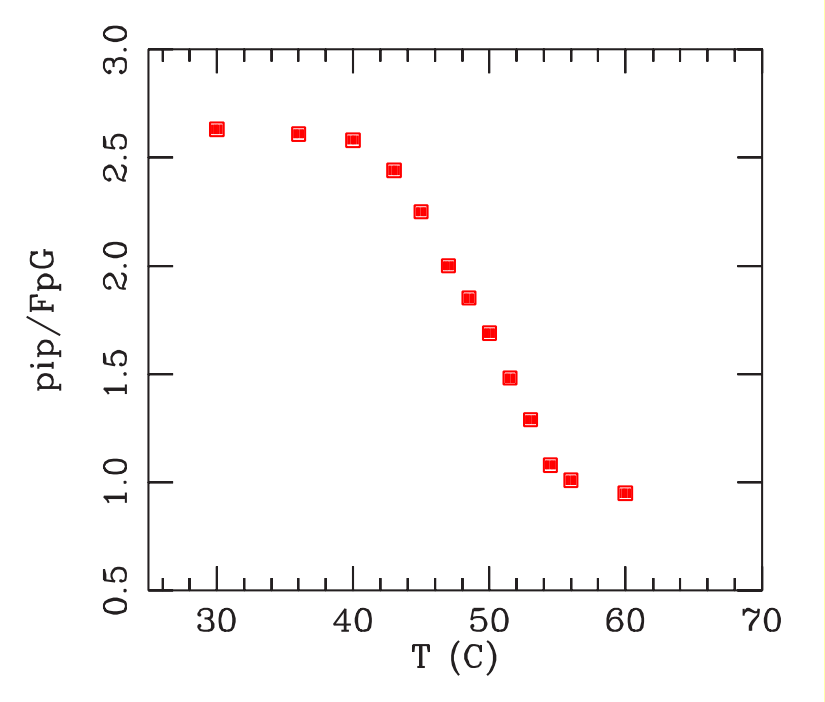
DNA fluctuations can influence sites a few base-pair away.



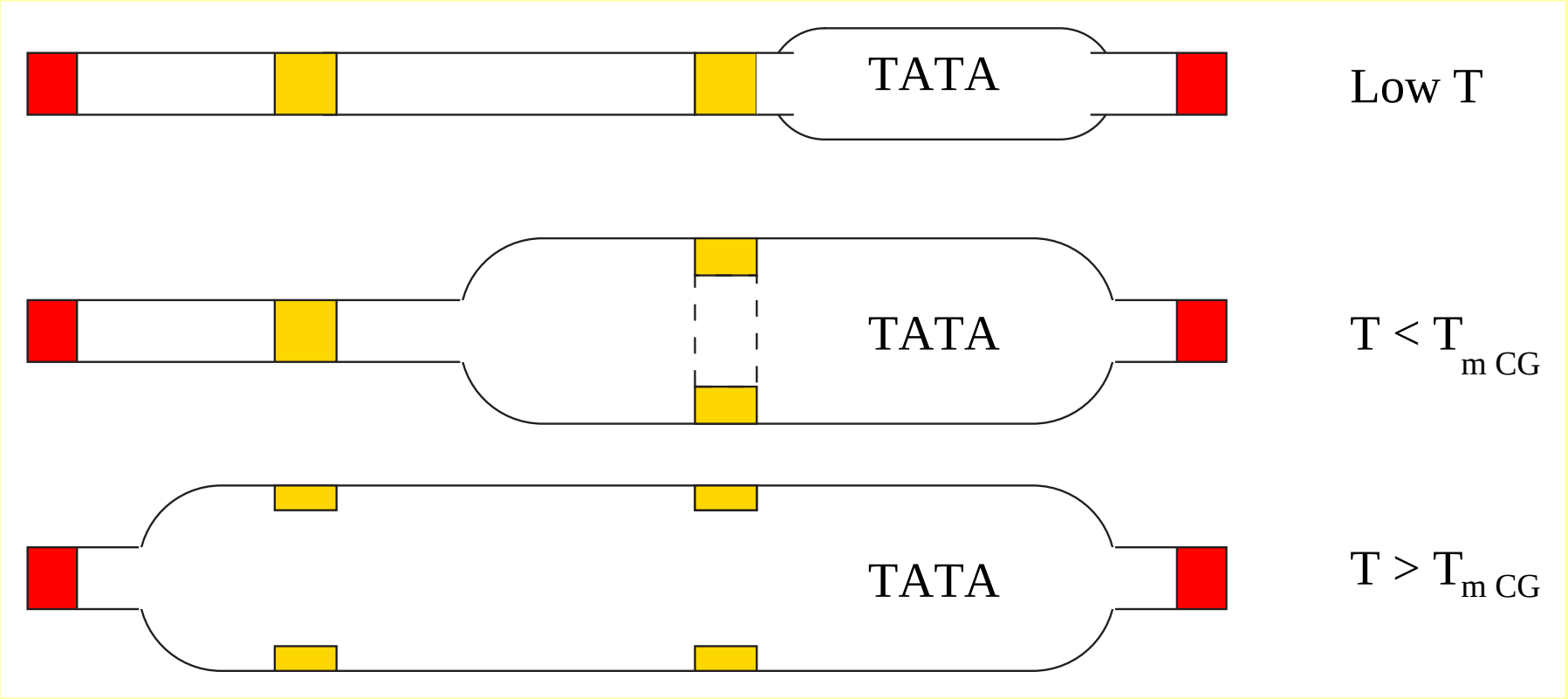
with TATA box



no TATA box, no probe B



A possible mechanism

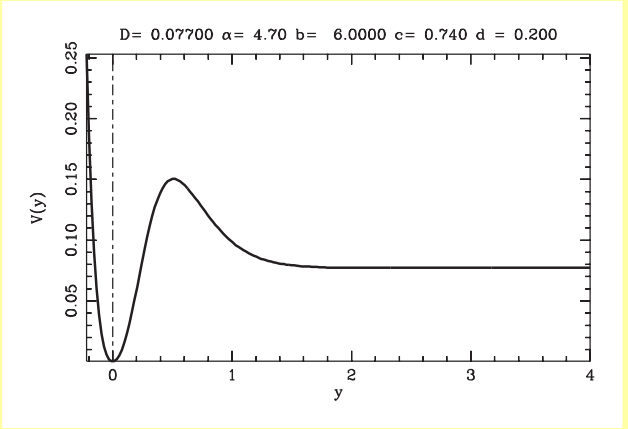


Improving the model.

More realistic potential within a base-pair

Include a barrier for closing (H bonds with solvent)

$$V_n(y_n) = D_n[e^{-\alpha_n y_n} - 1]^2 + \frac{b_n y_n^3 \theta(y_n)}{\cosh^2[c_n(\alpha_n y_n - d_n \ln 2)]}$$



Stacking interactions are also strongly sequence dependent

$$\begin{array}{ccccccc} \begin{matrix} C-G \\ G-C \end{matrix} & & \begin{matrix} C-G \\ A-T \end{matrix} & & \begin{matrix} G-C \\ T-A \end{matrix} & & \begin{matrix} A-T \\ A-T \end{matrix} \\ & \cdots\cdots & & \cdots\cdots & & & \begin{matrix} A-T \\ A-T \end{matrix} \\ & & \begin{matrix} T-A \\ G-C \end{matrix} & & \begin{matrix} A-T \\ C-G \end{matrix} & & \begin{matrix} T-A \\ T-A \end{matrix} \\ -0.63 & & -0.45 & & -0.28 & & -0.23 \quad -0.16 \quad \text{eV} \end{array}$$

(Quantum chemistry calculations. Thermodynamic measurements → similar conclusions)

In the model stacking interactions **must** depend on the interacting pairs

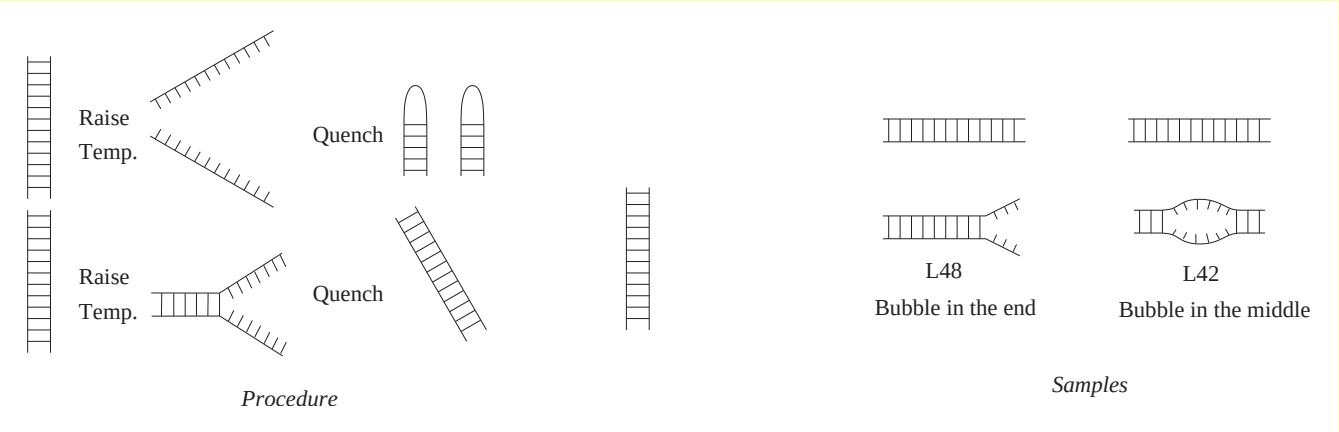
$$W(y_n, y_{n-1}) = \frac{1}{2} K_{n,n-1} \left[1 + \rho_{n,n-1}\, e^{-\delta_{n,n-1}\,(y_n + y_{n-1})} \right] (y_n - y_{n-1})^2$$

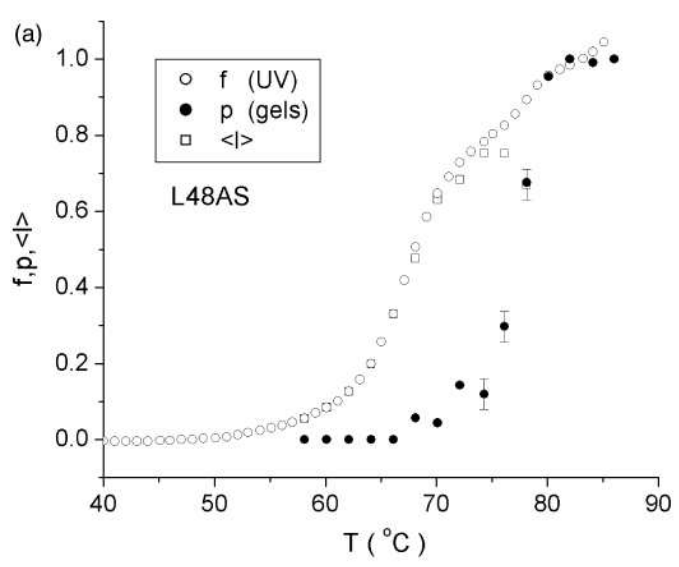
Experiments (observations of local fluctuations, thermodynamic measurements) and Quantum Chemistry → hints for parameter selection

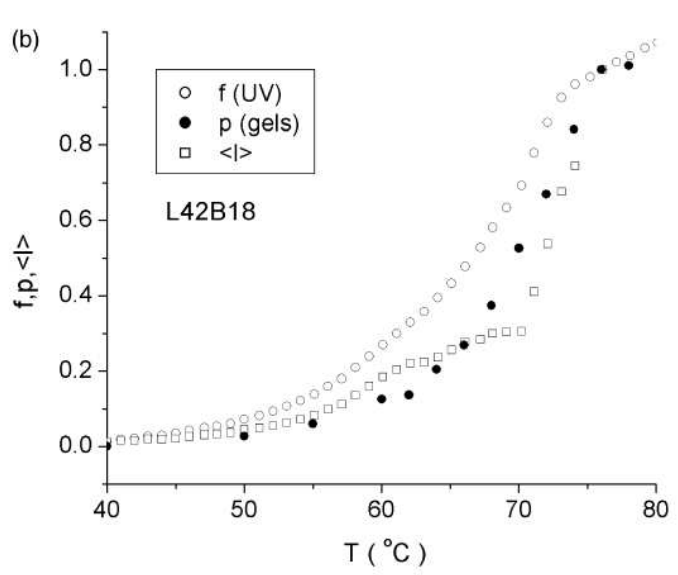
but fit to controlled experiments necessary (similar to Campa–Giansanti)

A clever trick (possible with artificial sequences): use sequence that can form hairpins to discriminate between partly open and fully open molecules.

Y. Zeng, A. Montrichok, G. Zocchi, J. Mol. Biol. **339** 67-75 (2004)







f fraction of open pairs

p fraction of open molecules

c fraction of open pairs within partially open molecules

$f = p + (1 - p)c$

$\sigma = f - p$ fraction of bases in bubble state

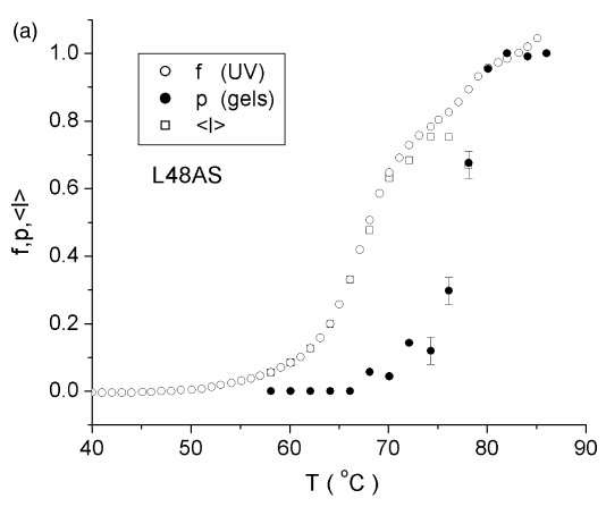
Model with pair-dependent stacking

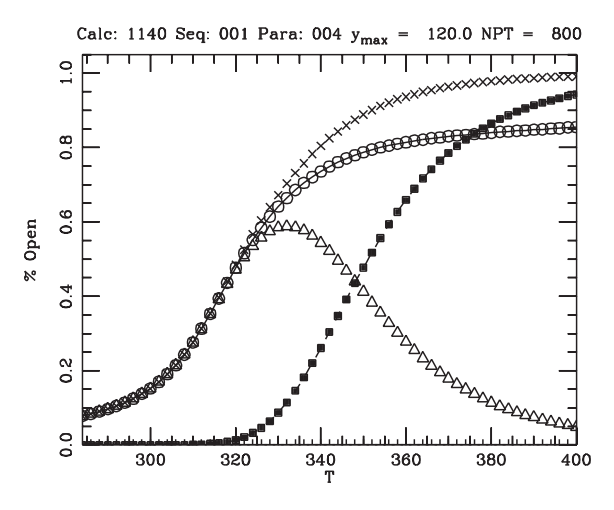
Calculation with full evaluation of the partition function (direct integration)

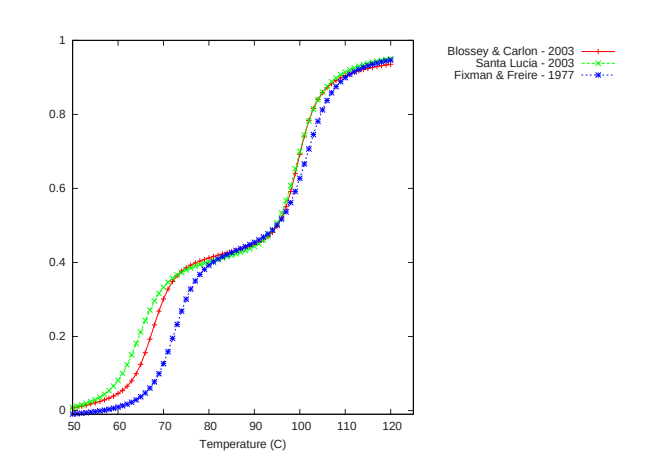
Defining a threshold y_0 for opening, calculate:

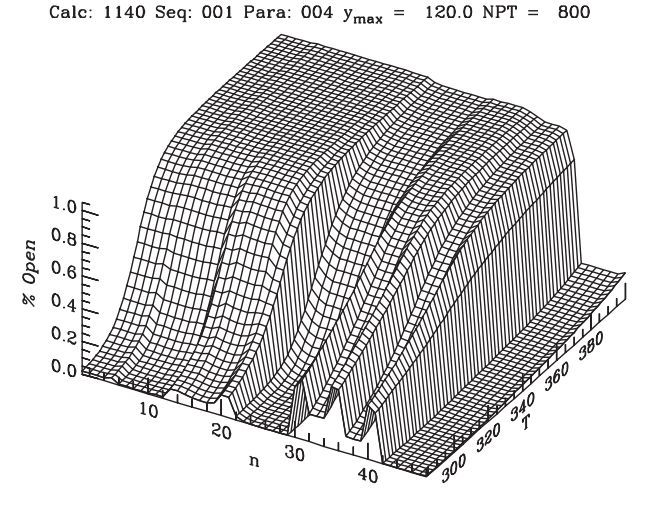
- probability for one base-pair to be open ($\rightarrow c$ fraction of open pairs in partially open molecules)
- probability that they are all open ($\rightarrow p$ fraction of open molecules)

Bubble in the end (L48)

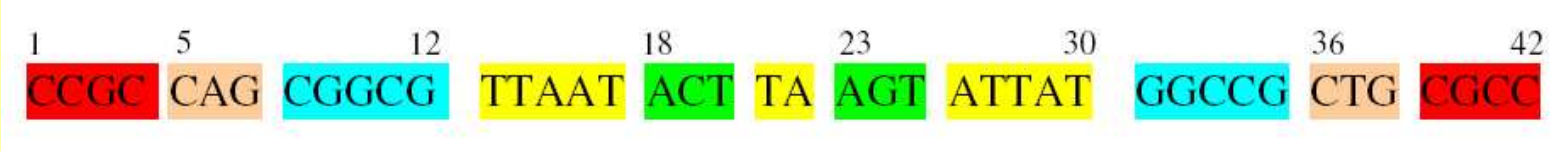


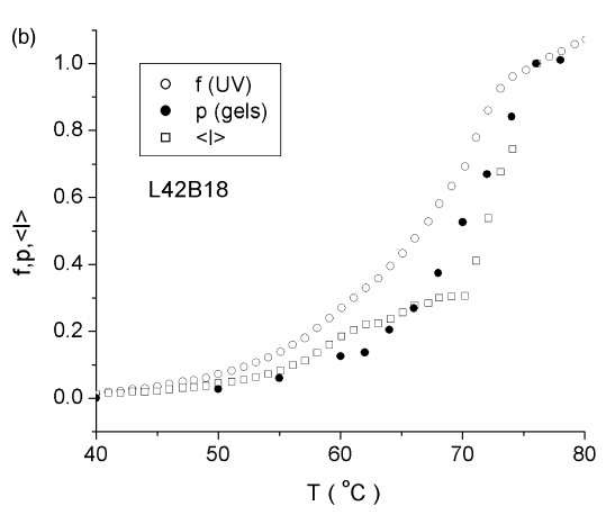


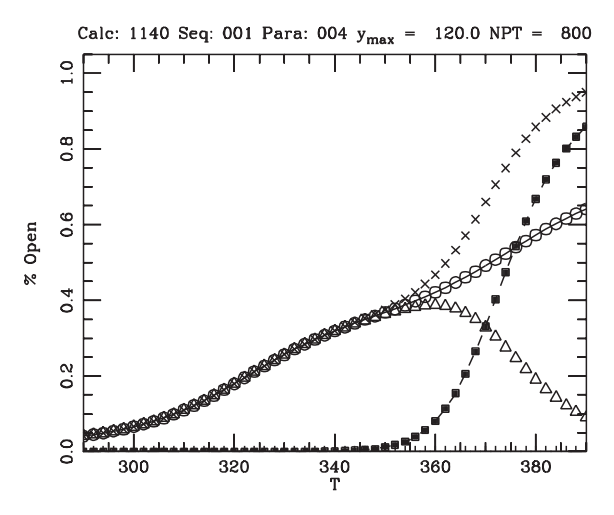


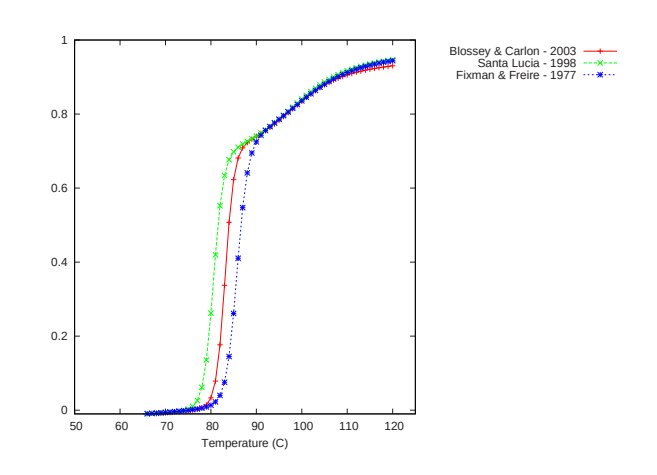


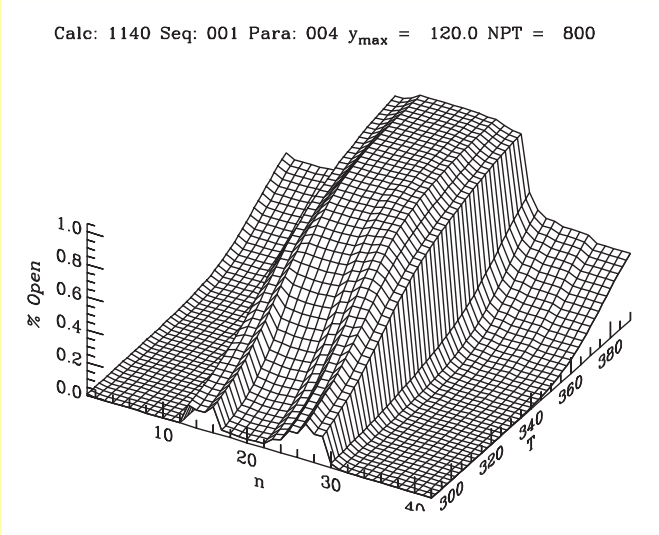
Bubble in the middle (L42)











Results extremely sensitive to details (in agreement with experiments!)

Calculations with Ising models (Poland Scheraga), which are standard for applications in biology (PCR) are *qualitatively wrong*.

We are doing better but ...

Calculation shows that two-body potentials are not enough

Conclusion: can we predict DNA biological activity theoretically?

- Statistical studies required ⇒ mesoscopic models
- Rough prediction of opening probabilities versus sequence agree with experiments but
  - they do not discriminate sites with biological activity
  - sequence effects are very subtle and results are highly sensitive to details (local structure affected by sequence) ⇒ results accurate enough to be biologically significant are hard to get

- Experiments are becoming very precise ⇒ should allow model improvements.
  - Local observation of fluctuations by UV oxidative modifications of Guanine
  - Availability of designed sequence open new possibilities for “old” experiments (melting studies)