Models for DNA denaturation

Henri Orland
SPhT, CEA-Saclay

Outline

- Review of the Denaturation Transition of DNA
- Simple polymer model: the Schrodinger equation approach
- The Peyrard-Bishop model
- The Poland-Scheraga model
  - Recursion equations
  - the Fixman-Freire method

Outline

- Theory for DNA Hybridization
  - effects of mutations and mismatches
- Effect of a torque: Supercoiling

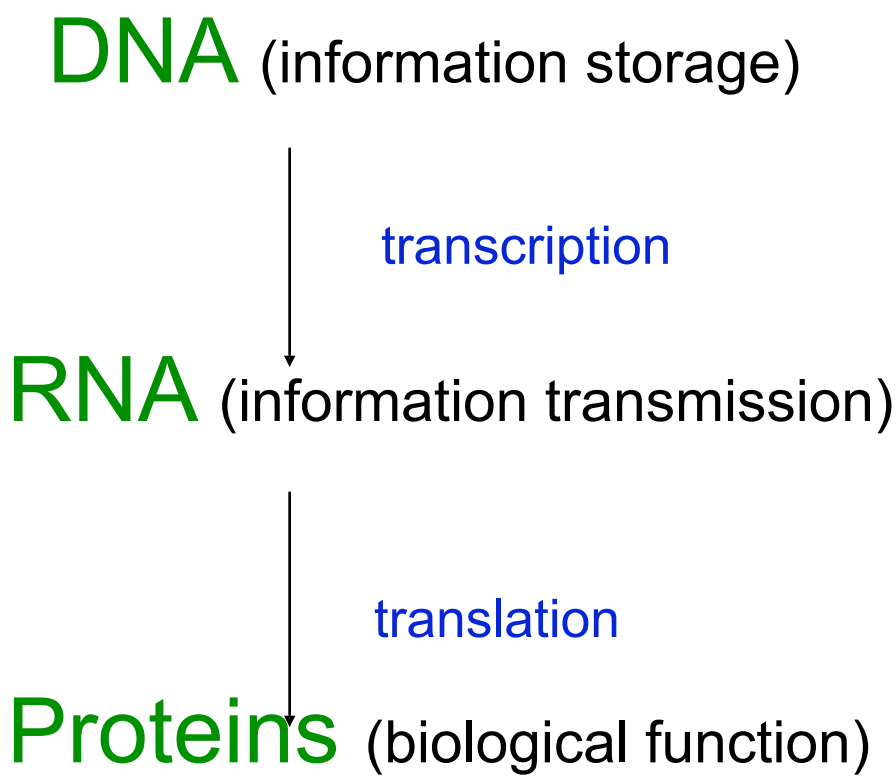
Review of basic properties of DNA

- DNA is a biopolymer
  - RNA (length ~ 70 – 2000)
  - DNA (length ~ 10^6 – 10^9)
  - Proteins (length ~ 10^2)
  - Polysaccharides (length ~ 10^3)

Composition of Cell (in weight)

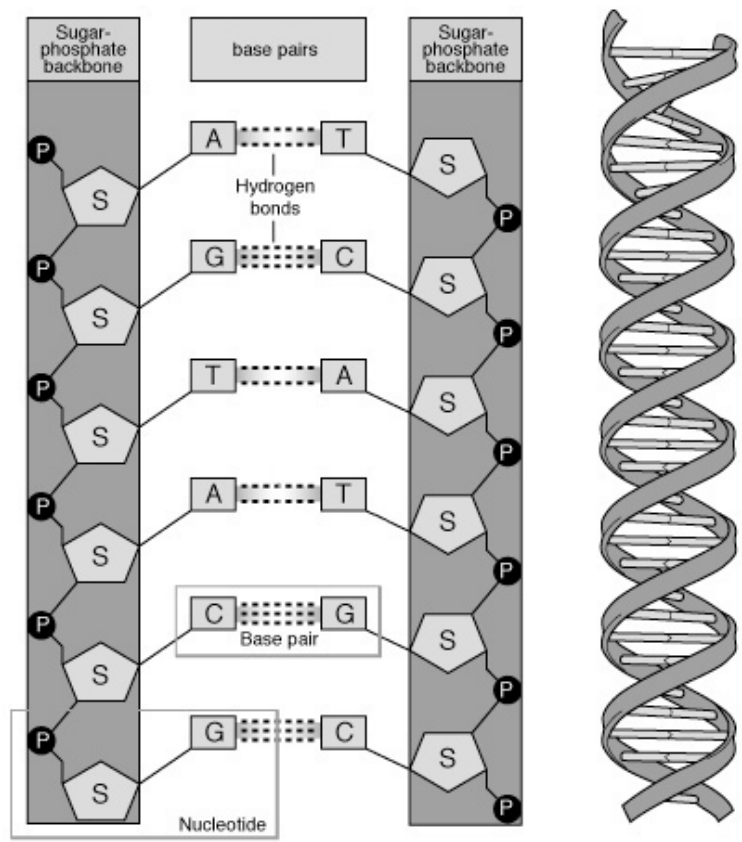
- Water 70%
- Proteins 15%
- DNA 1%
- RNA 6%
- Polysaccharides 3%
- Lipids 2%
- Mineral ions 3%
- Etc…

Central dogma of Biology

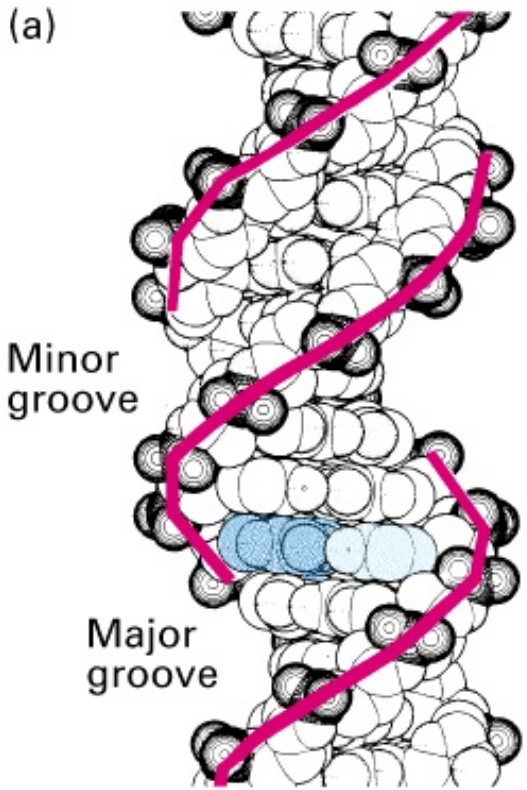


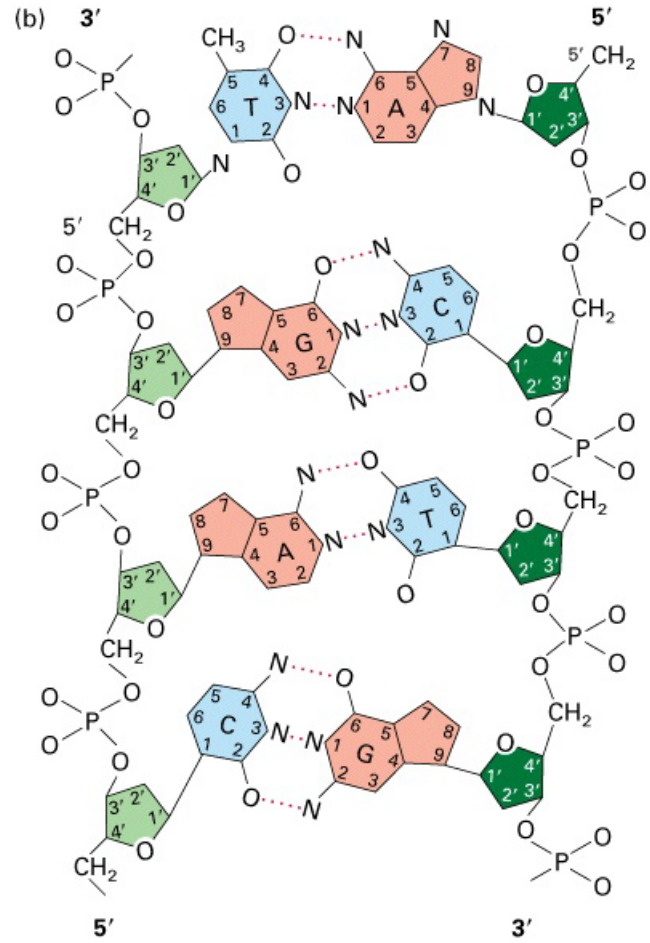
DNA structure

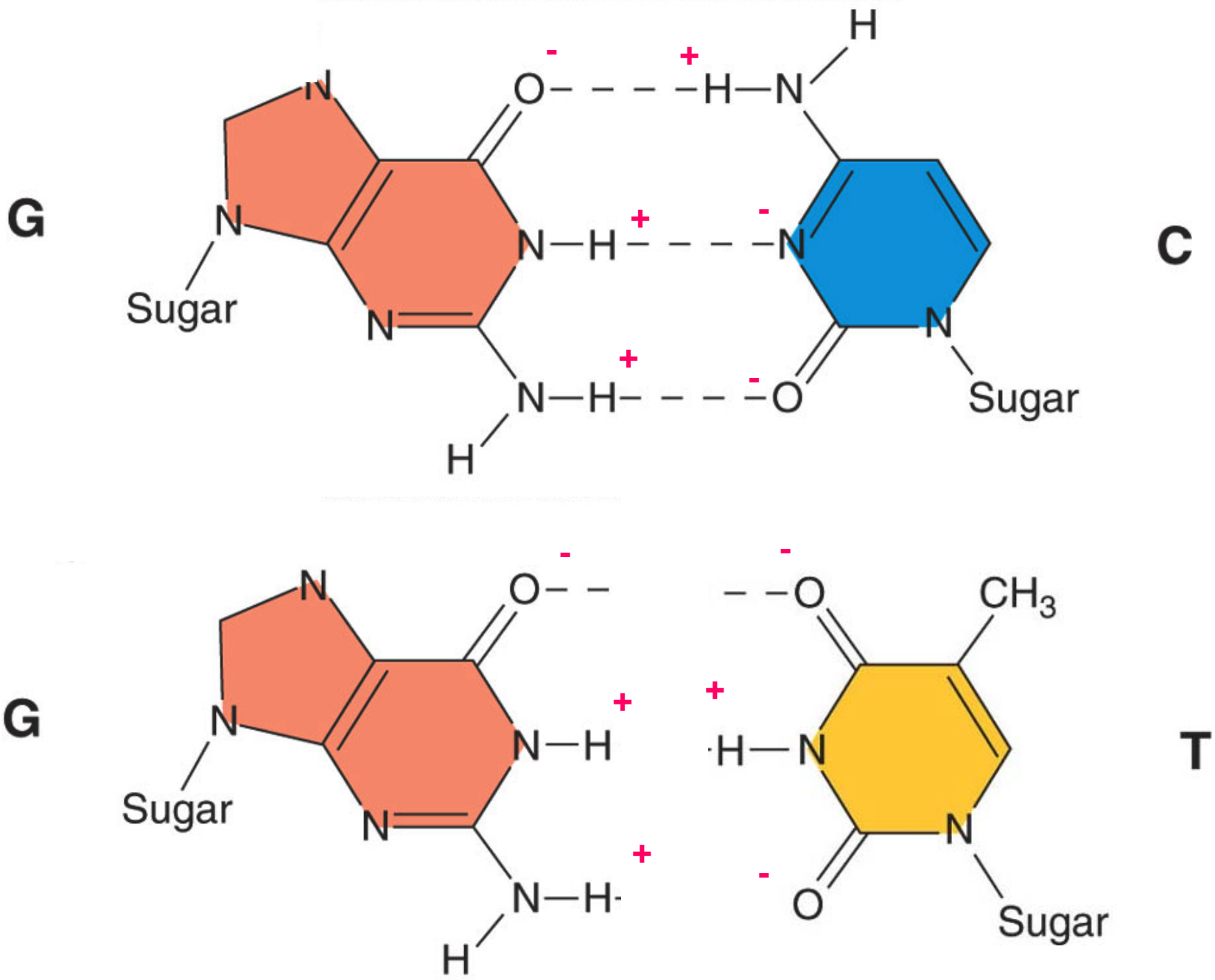
- DNA is a double stranded polymer
- Made of 4 bases:
  - adenine
  - guanine
  - cytosine
  - thymine



Native DNA is a double helix of complementary antiparallel chains



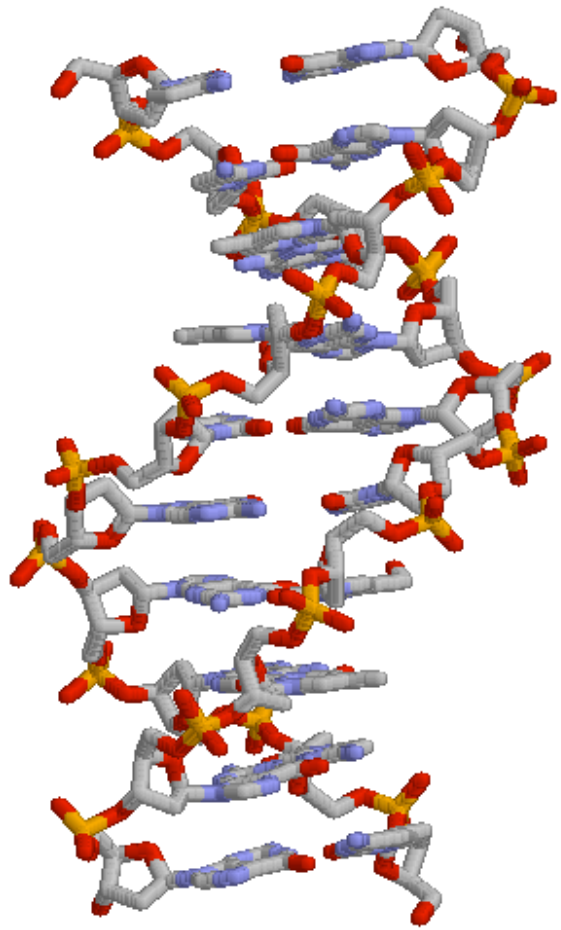



G
C
Sugar
Sugar
G
T
Sugar
Sugar

- Watson-Crick pairing
  - G -- C : 3 Hydrogen bonds , about 3 kCal/mole (5 kT)
  - A -- U : 2 Hydrogen bonds , about 2 kCal/mole (3 kT)
- The 2 strands are complementary
- The length of a DNA ranges from few thousands to few billions.

- In addition, there are Stacking Energies.
- Nucleic acids are charged ⟹ DNA is soluble
- The organic rings of the bases are Hydrophobic ⟹ bases have a tendency to cluster: Stacking energies

Electrostatics



DNA is strongly negatively charged due to Phosphate groups

Usually, there are Mg++, Na+ and Cl- ions ⟹ Screening

Debye-Huckel interaction between the charged monomers of DNA

$$\kappa_{DH}^2 = 4\pi l_B c_I$$

l_B is the Bjerrum length $7\mathring{A}$

c_I is the total ion concentration

Screening length is $4\mathring{A}$ in 1M

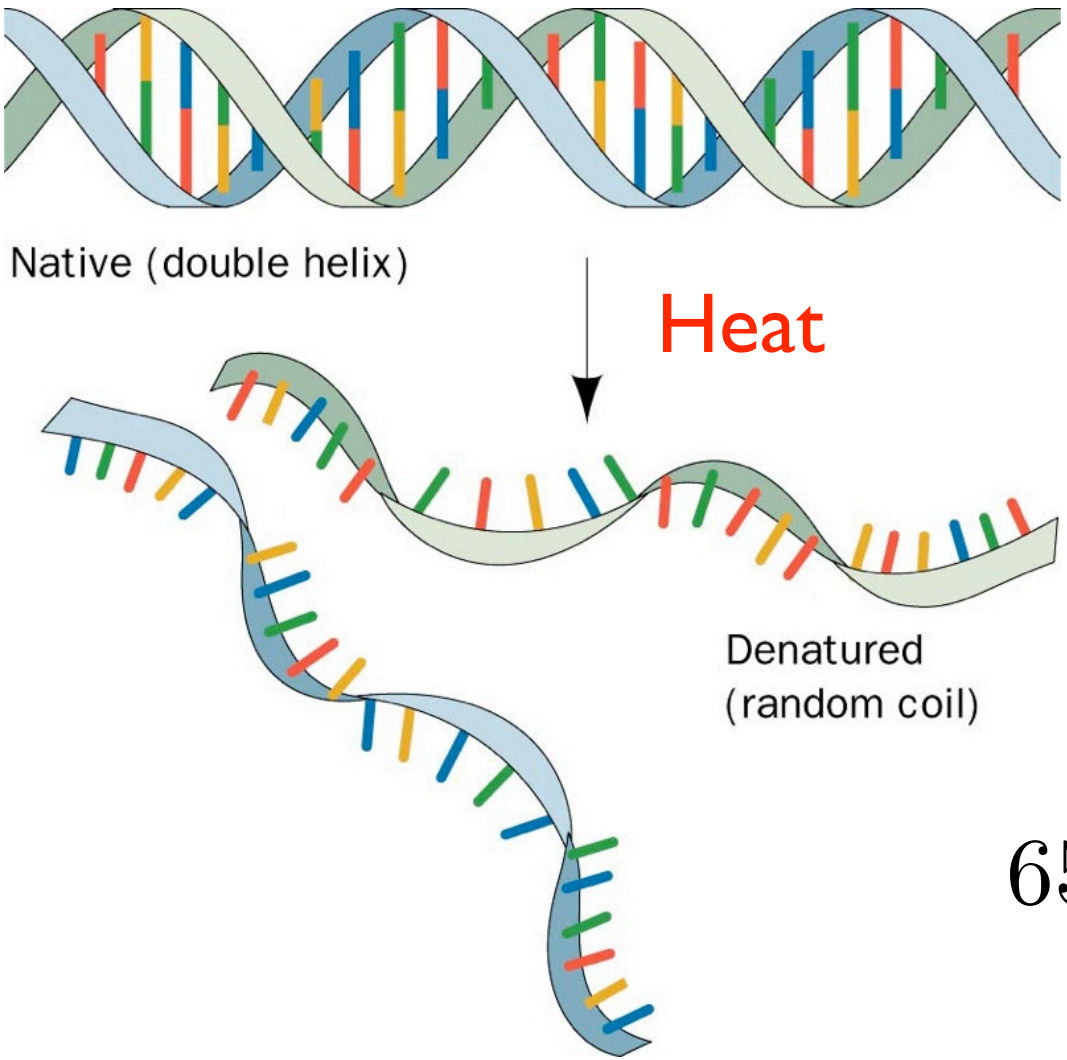
$20\mathring{A}$ in 0.075 M

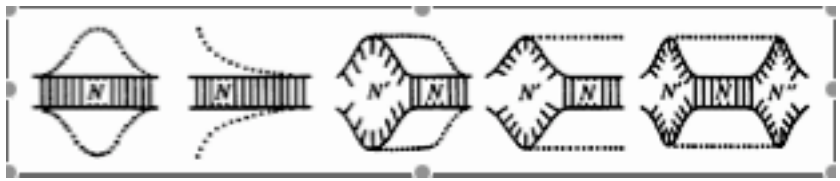
⟹ $v_{DH}(r) = \frac{l_B}{r} e^{-\kappa_{DH} r}$

Bending and persistence

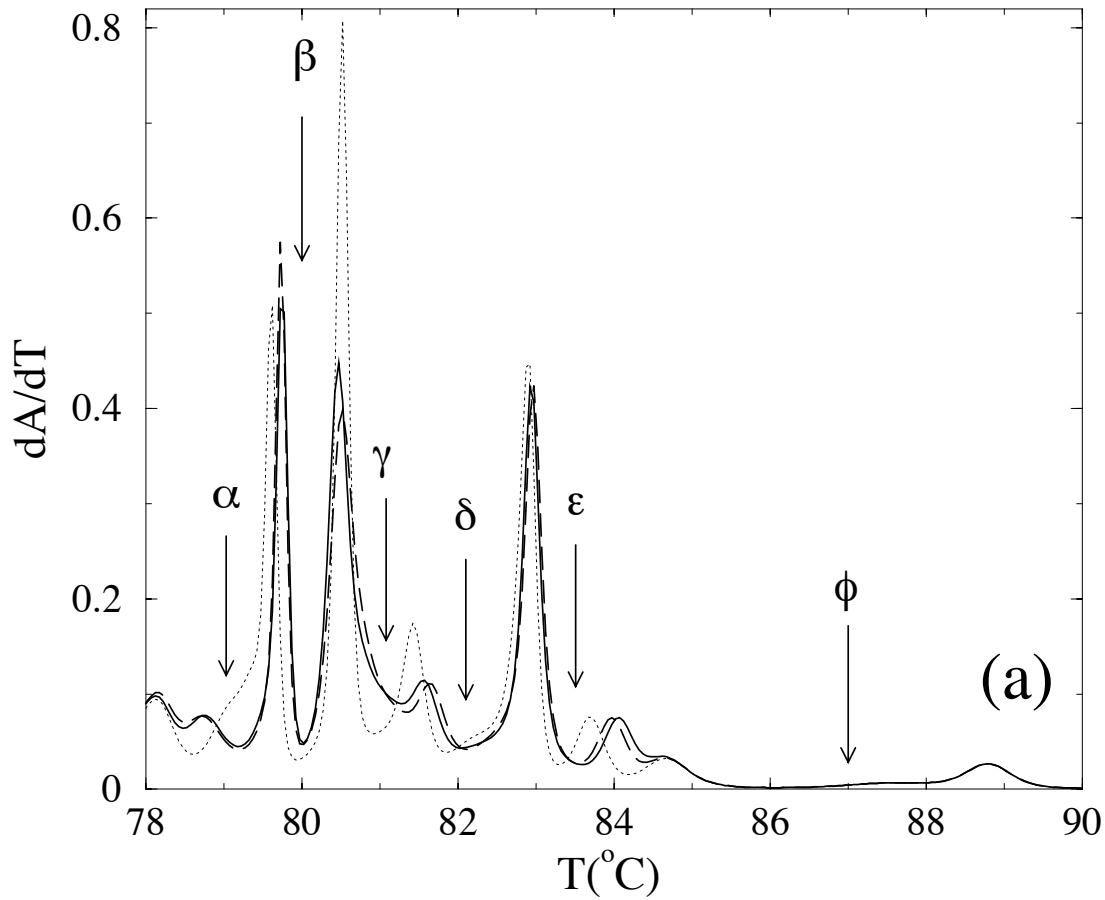
- Bending energy is characterized by the persistence length (correlation length of tangents)
  - $l_p \approx 150$ bp for double-stranded DNA
    $\approx 750 \mathring{A}$
  - $l_p \approx 15$ bp for single-stranded DNA
    $\approx 75 \mathring{A}$

DNA denaturation





$$65\ C \leq T \leq 110\ C$$



A is the number of bound pairs. Measured by looking at UV adsorption at 260 nm

$$65\ C \leq T \leq 110\ C$$

Sharp peaks: very cooperative phenomenon

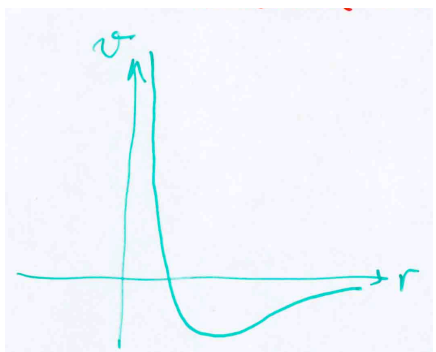
Transition looks discontinuous (1st order)

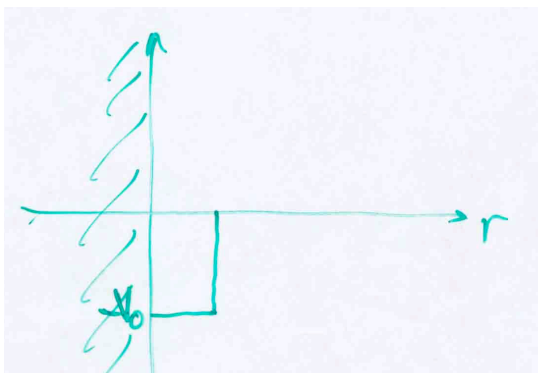
Why is it interesting?

- It is a nice statistical physics problem
- Allows to understand molecular recognition
- Allows to discriminate between coding and non coding regions of DNA
- Nice and clean experiments
- Relevant to DNA chips

A simple polymer model

- Assume 2 chains of elastic beads
- Model the H-bonds by a short range attraction between complementary bases





- Forget about Excluded Volume effects.

$$Z = \int dr_i^{(1)} dr_i^{(2)} e^{-\frac{3}{2a^2}\sum_{i=1}^{N-1}(r_{i+1}^{(1)} - r_i^{(1)})^2 - \frac{3}{2a^2}\sum_{i=1}^{N-1}(r_{i+1}^{(2)} - r_i^{(2)})^2}$$

$$\times e^{-\beta \sum_{i=1}^{N} v(r_i^{(1)} - r_i^{(2)})}$$

- where $a = l_p$ is the persistence length of single stranded DNA
- For this model, all binding energies are identical.

- Take the continuous limit ⟶ Feynman path integral

$$Z = \int \mathcal{D}\vec{r}_1(s)\,\mathcal{D}\vec{r}_2(s)\,\exp\left(-\frac{d}{2a^2}\int_0^N ds\,\left(\left(\frac{d\vec{r}_1}{ds}\right)^2+\left(\frac{d\vec{r}_2}{ds}\right)^2\right)-\beta\int_0^N ds\,v(\vec{r}_1(s)-\vec{r}_2(s))\right)$$

- Make the change of variables

$$\vec{R}(s) = \frac{\vec{r}_1(s)+\vec{r}_2(s)}{2}$$

$$\vec{r}(s) = \vec{r}_1(s)-\vec{r}_2(s)$$

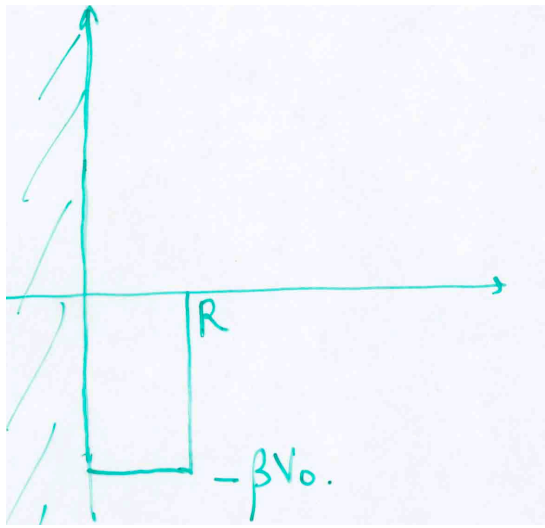
$$Z = \int \mathcal{D}\vec{r}(s)\,\exp\left(-\frac{d}{4a^2}\int_0^N ds\,\left(\frac{d\vec{r}}{ds}\right)^2-\beta\int_0^N ds\,v(\vec{r}(s))\right)$$

- One recognizes a Quantum Mechanical matrix element

$$Z = \int dr < r|e^{-NH}|0 >$$

- where the Hamiltonian is given by

$$H = -\frac{4a^2}{3}\nabla^2 + \beta v(r)$$



- In the limit $N \to \infty$, only the ground state dominates: Ground State Dominance.
- At high temperature, β small, the Hamiltonian has no bound state: Denatured phase
- At low temperature, β is large, the Hamiltonian has a bound state: Helical phase

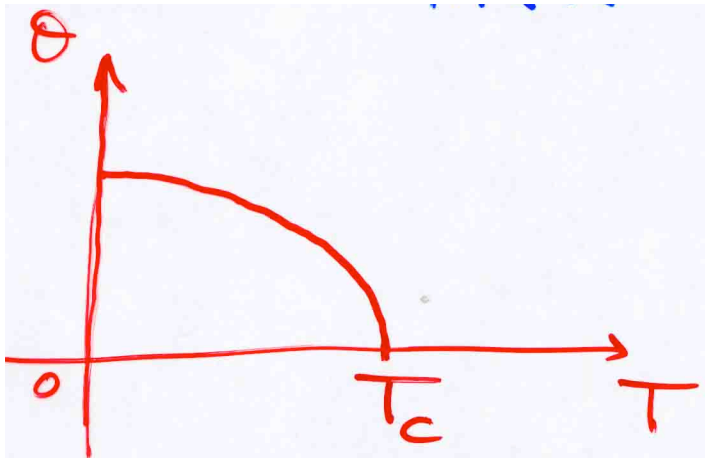
$$Z \approx e^{-NE_0} \int dr \Psi_0(r) \Psi_0(0)$$

- where

$$H|\Psi_0> = E_0|\Psi_0>$$

- There is a phase transition at a temperature T_c
- It is a second order (continuous) unbinding transition

$$\theta = \frac{N_{bound}}{N} = \int_{|r|<R} dr \ \Psi_0^2(r)$$

- 

- Model for Unzipping with a force

$$Z = \int \mathcal{D}\vec{r}(s)\, \exp\left(-\frac{d}{4a^2} \int_0^N ds \left(\frac{d\vec{r}}{ds}\right)^2 - \beta \int_0^N ds\, v(\vec{r}(s)) - \beta f.r(N) \right)$$

- which can be written as

$$Z = \int dr < r|e^{-NH} e^{-\beta f.r}|0 >$$

The Peyrard-Bishop model

- The original Peyrard-Bishop model is exactly the Schrodinger model in one dimension, with a Morse potential

$$v(r) = V_0(e^{-4cr} - 2e^{-2cr})$$

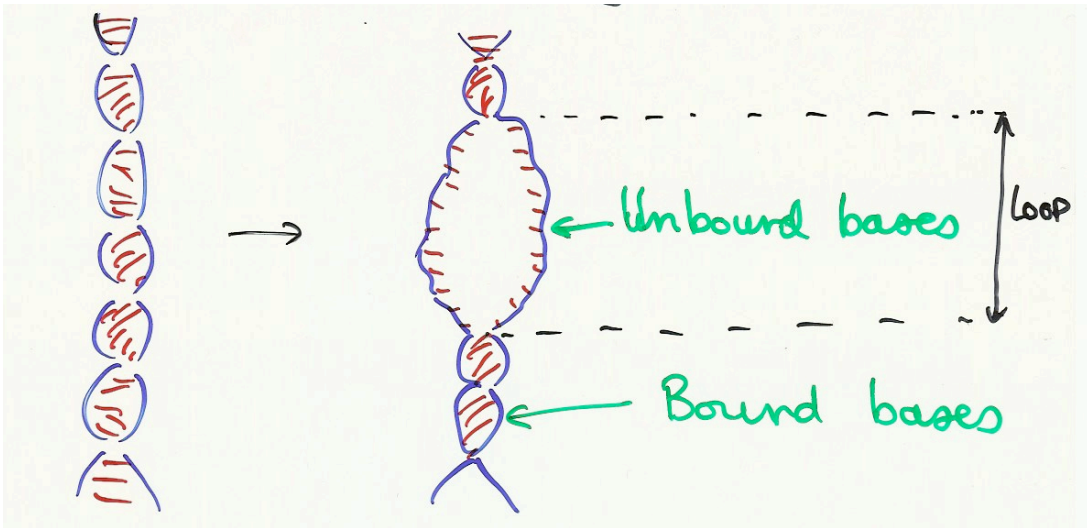
- The Peyrard-Bishop model can be modified to include stacking energies.
- The discrete form is given by

$$Z = \int \prod_{n=1}^{N} dy_n e^{-\frac{\beta k}{2}(1+\rho e^{-a(y_n+y_{n-1})})(y_n-y_{n-1})^2-\beta D(e^{-ay_n}-1)^2}$$

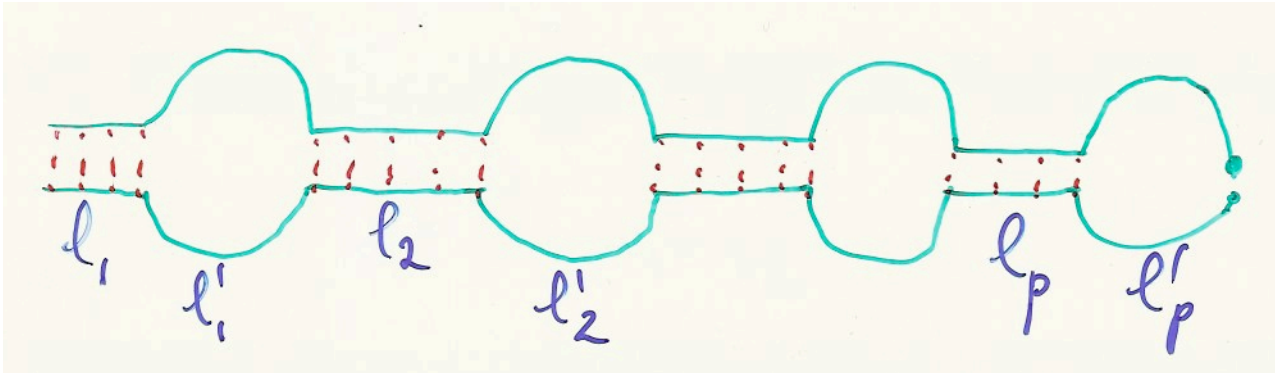
The Poland-Scheraga model

- the bases of DNA are modeled as points
- two types of energies between bases
  - Hydrogen bonds
  - Stacking energies: due to hydrophobicity of the rings
- the entropy of loops is calculated using polymer theory

Basic model



- only one type of energies: $\varepsilon_i = -\varepsilon$
- unbound loops have entropy: $\Omega(2l)$
- can be easily solved



$$Z_N = \sum_{p=1}^{\infty} \sigma^p \sum_{l_1=1}^{\infty} \sum_{l'_1=1}^{\infty} \dots \sum_{l_p=1}^{\infty} \sum_{l'_p=1}^{\infty} \delta(\sum_{i=1}^{p} (l_i + l'_i) - N) w^{l_1} \Omega(2l'_1) \dots w^{l_p} \Omega(2l'_p)$$

where

σ is the fugacity of loops (due to stacking) $\approx 10^{-5}$

$w = e^{\beta\varepsilon}$

Loop Entropy

From Polymer theory (Flory, de Gennes, des Cloizeaux, ...) we know that

$$\Omega(2l) \sim \frac{s^{2l}}{l^c}$$

where c is a universal critical exponent

for Brownian walks: $c = 1.5$

for a Self-Avoiding walk: $c = 3\nu \approx 1.75$

for interacting Self-Avoiding loops: $c \approx 2.15$

Grand partition function

Define the grand canonical partition function

$$Z(z) = \sum_{N=0}^{\infty} z^N Z_N$$

one gets

$$Z(z) = \sum_{p=1}^{\infty} \sigma^p \sum_{l_1=1}^{\infty} \dots \sum_{l_1'=1}^{\infty} (zw)^{l_1} z^{l_1'} \Omega(2l_1') \dots (zw)^{l_p} z^{l_p'} \Omega(2l_p')$$

The sum is now decoupled and one obtains

$$Z(z) = \sum_{p=1}^{\infty} \sigma^p \left(\sum_{l=1}^{\infty} z^l w^l \right)^p \left(\sum_{l=1}^{\infty} z^l \Omega(2l) \right)^p$$

$$Z(z) = \frac{\sigma U(z) V(z)}{1 - \sigma U(z) V(z)}$$

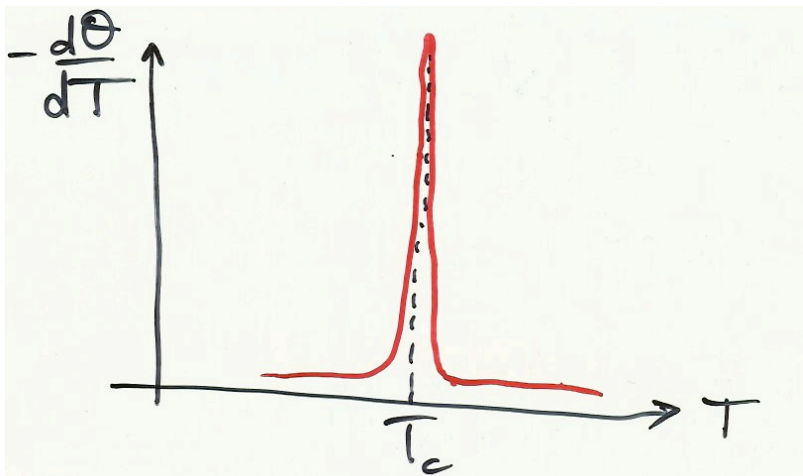
where

$$U(z) = \frac{zw}{1 - zw}$$

and

$$V(z) = \sum_{l=1}^{\infty} \frac{s^{2l} z^l}{l^c}$$

is the polylogarithm function

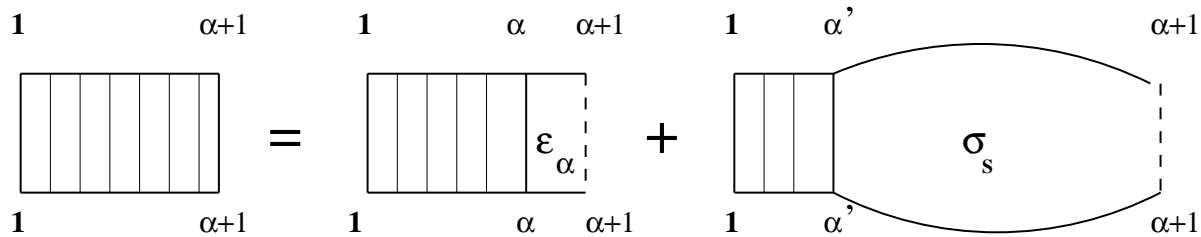


Results

- the phase diagram depends on the loop exponent c
- If c<1, the 2 strands are always bound
- if 1<c<2, there is a continuous unbinding transition
- if 2<c, there is a discontinuous unbinding transition
- since σ is so small ($\approx 10^{-5}$) even if 1<c<2, the transition looks discontinuous.

Inhomogeneous sequences

- One can reformulate the problem in terms of recursion relations.
- Define the partition function $Z(\alpha)$ of the non homogeneous fragment of length α



$$Z_f(\alpha+1) = e^{-\beta\varepsilon_\alpha}\ Z_f(\alpha) + \sigma_S \sum_{\alpha'=1}^{\alpha-1} Z_f(\alpha')\mathcal{N}(2(\alpha+1-\alpha'))$$

- Algorithm scales like N^2
- Limited to fairly small sizes (< 10000)
- Must improve to study full genomes

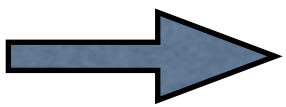
The Fixman-Freire method

Idea: represent the loop power law as a sum of exponentials

$$f(x) = \frac{1}{x^c} \simeq \sum_{i=1}^{I} a_i \; e^{-b_i x}$$

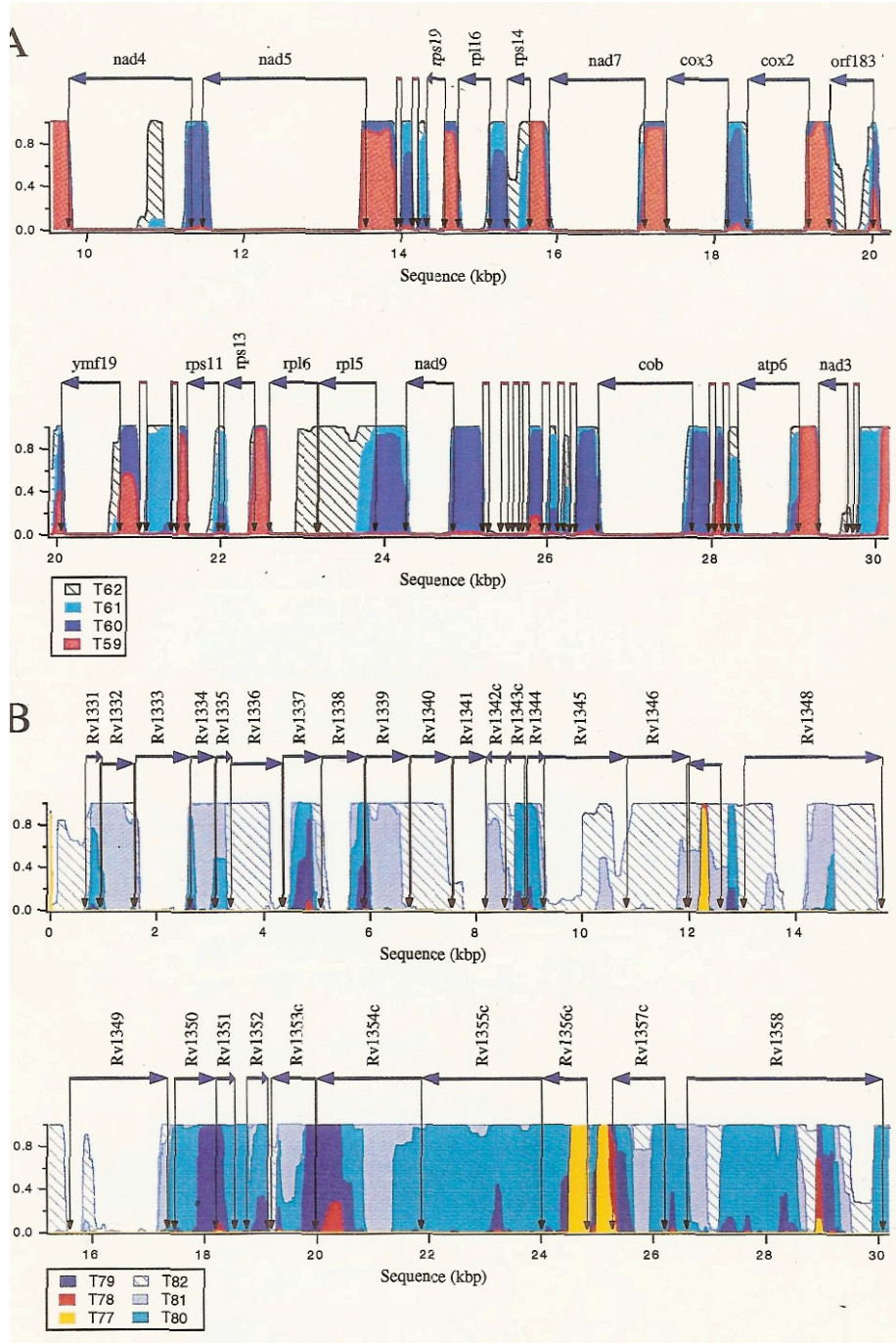
With this representation, the algorithmic complexity goes down from N^2 to NI

For *N* = 1000000 and *I* = 14, the accuracy is better than 0.1% over the whole range of *x*.



Possibility to study sequences up to few Mbps

Some examples: gene detection

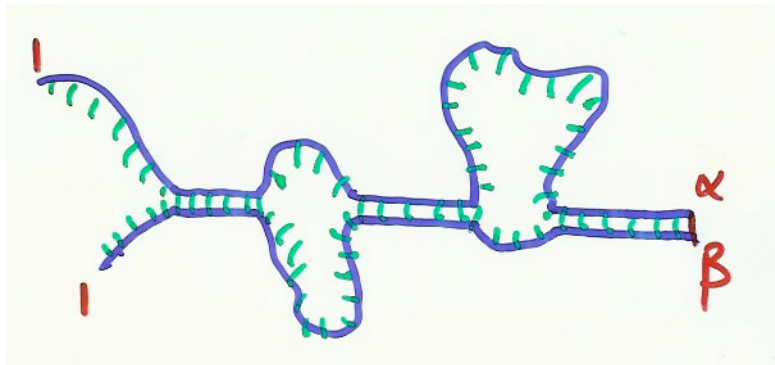


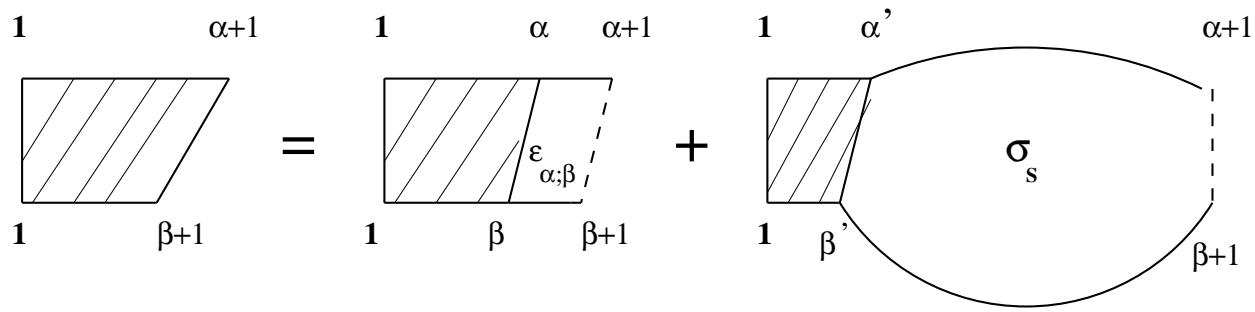
Limitations of the PS model

- strands of equal length
- complementary sequences
- no mismatches
  - generalize to a full theory of DNA hybridization

A general model for DNA hybridization

We now consider 2 strands of length N_1 and N_2 where the 2 strands are not necessarily complementary. Define $Z(\alpha, \beta)$ as the partition function of the 2 DNA strands with strand 1 going from 1 to α and strand 2 from 1 to β bound at α and β





$$Z_f(\alpha+1,\beta+1) \;=\; e^{-\beta\varepsilon_{\alpha;\beta}} Z_f(\alpha,\beta)$$

$$+\;\; \sigma_S \sum_{\alpha'=1}^{\alpha-1} \sum_{\beta'=1}^{\beta-1} Z_f(\alpha',\beta') \mathcal{N}(\alpha+1-\alpha'+\beta+1-\beta')$$

Now, any base of 1 can pair with any base of 2

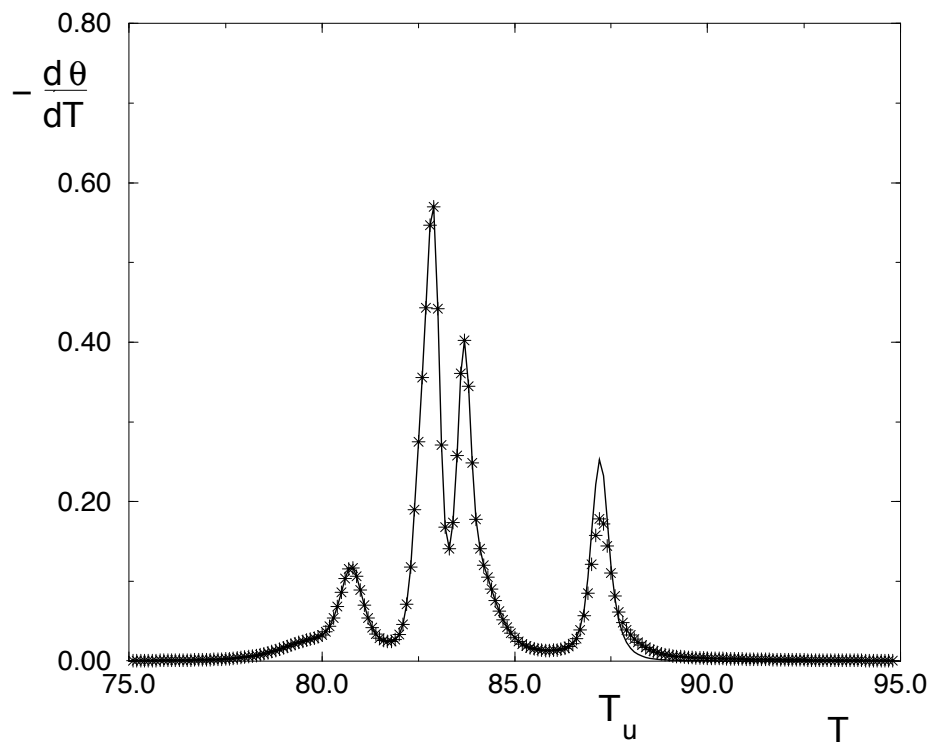
$$\varepsilon_{\alpha;\beta} = \varepsilon_\alpha \quad \text{if} \quad \beta = \bar{\alpha}$$

$$\varepsilon_{\alpha;\beta} = 0 \quad \text{otherwise}$$

- Now, algorithm scales like $N_1^2 N_2^2$: unusable!
- Use Fixman-Freire trick: it becomes $N_1 N_2 I$
- can study sequences of sizes up to 10000

Some results

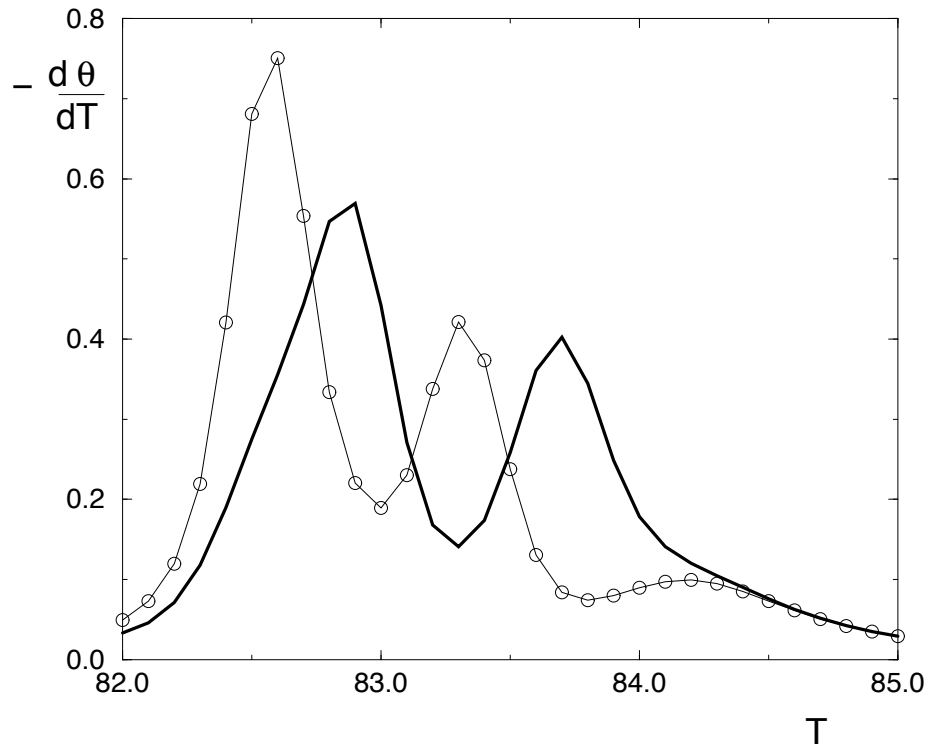
- For Homogeneous sequences (AAAAA..., TTTTTT...), one can again solve analytically. Effectively, the exponent c is replaced by c-1
- Comparison of PS and GPS for long complementary sequences



- The two curves superimpose, except near the last peak ⟹ no mismatches if the two strands are complementary. Very strong selectivity in molecular recognition.

Effect of single point mutations

Two complementary sequences of length N=1980 with a single point mutation somewhere in the center

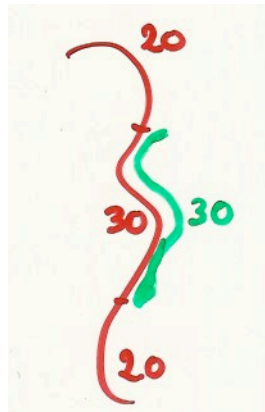


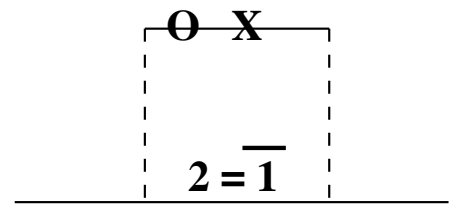
Strong sensitivity to mutations: molecular selectivity

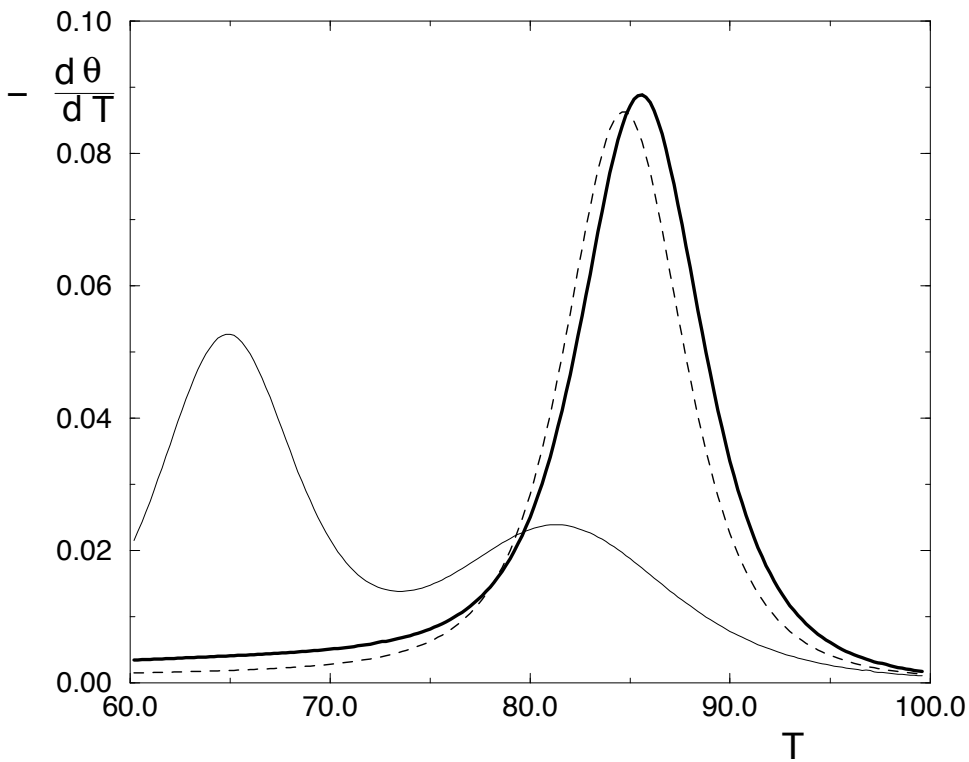
Shift of 1C for one mutation and 5-10 C for two mutations.

No more binding for 3 mutations.

Effect of mutations on short sequences: DNA chips







Very strong sensitivity to a single mutation in the central region

Very weak sensitivity near edges