From DNA opening to DNA sequence

Rémi Monasson
(Laboratoire de Physique Théorique, Ecole Normale Supérieure, Paris)

Collaborators :

Valentina Baldazzi	(doctorante, Rome)
Carlo Barbieri	(doctorant, Paris)
Serena Bradde	(stagiaire M2, Trieste)
Simona Cocco	(Lab. Phys. Stat. ENS, Paris)
Thomas Jorg	(Lab. Phys. Theor. ENS, Paris)
Enzo Marinari	(Dip. Fisica, La Sapienza, Rome)
Francesco Zamponi	(Lab. Phys. Theor. ENS, Paris)

Aussois June 2009

Overview

1. Introduction :

- *reminder on the DNA molecule and its opening*
- *micro-manipulation experiments*
- *unzipping of DNA/RNA at fixed force or extension*
- *modeling of the opening for a given sequence: relationship with random walks in random media*

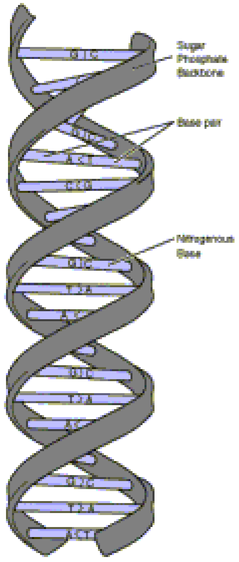
2. « Dynamical » inference

3. « Static » inference and perspectives

DNA sequence

DNA = double helix of 2 complementary strands A/T, G/C

3’ 5’

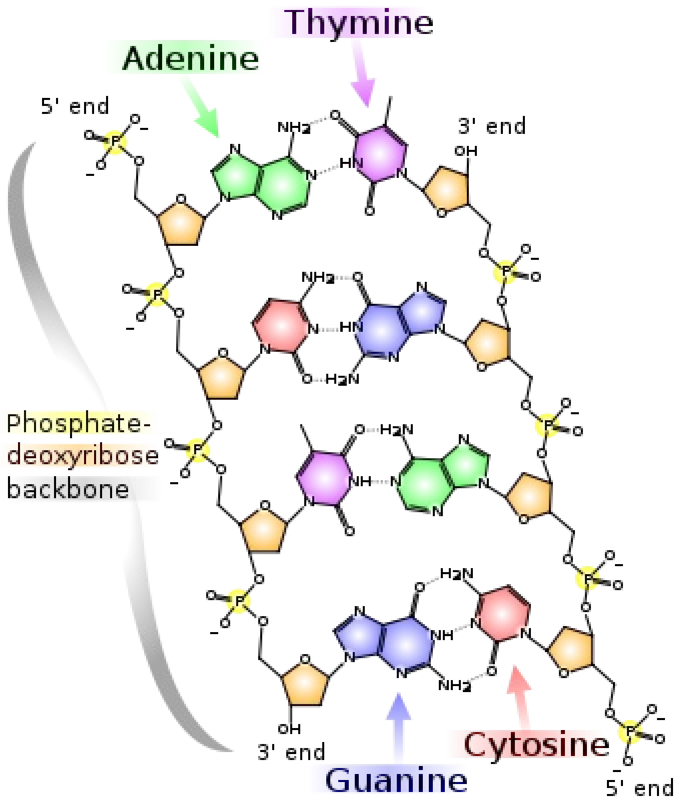


5’ 3’

5’---- ATGTGGTTAG --- 3’

3’---- TACACCAATC --- 5’

sequence = genetic information

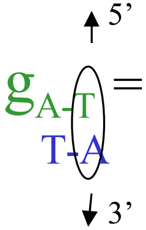


Base pairing and stacking (free) energies

Typical energy: $1\ k_BT \approx 4\ \mathrm{pN} \times \mathrm{nm}$ (at temperature T=25 °C)

H-bonds: $g_{A\text{-}T} \approx 1\ k_BT$ $g_{C\text{-}G} \approx 3\text{-}4\ k_BT$

Stacking:



$$g_{\substack{A\text{-}T\\T\text{-}A}} = 1.06\ k_BT$$

$$g_{\substack{A\text{-}T\\A\text{-}T}} = 1.78\ k_BT$$

$g(b_n,b_{n+1})$

b_{n+1}

b_n

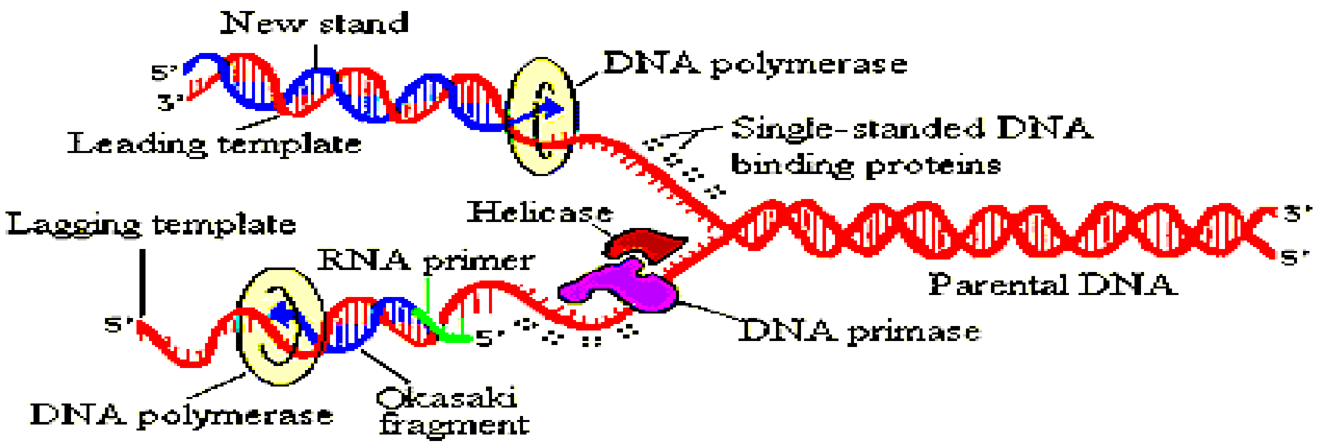
	A	T	C	G
A	1.78	1.55	2.52	2.22
T	1.06	1.78	2.28	2.54
C	2.54	2.22	3.14	3.85
G	2.28	2.52	3.90	3.14

Only 10 different values

at room temperature and standard solution conditions
(150 mM Na Cl, neutral pH)

Strand separation

In vivo opening : replication, transcription, …



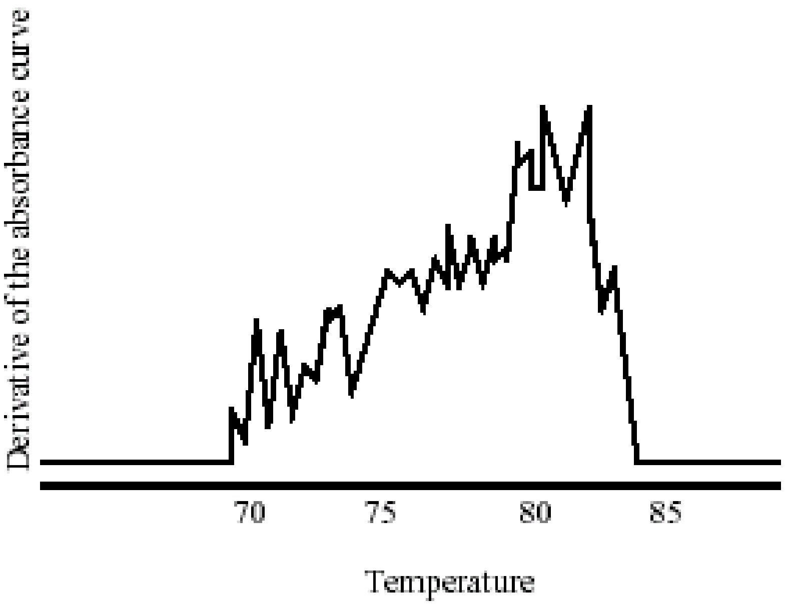
In vitro opening:

- in a bulk : thermal denaturation
- single molecule experiments

Nanopore unzipping
DNA digestion by an Exonuclease
Force Unzipping

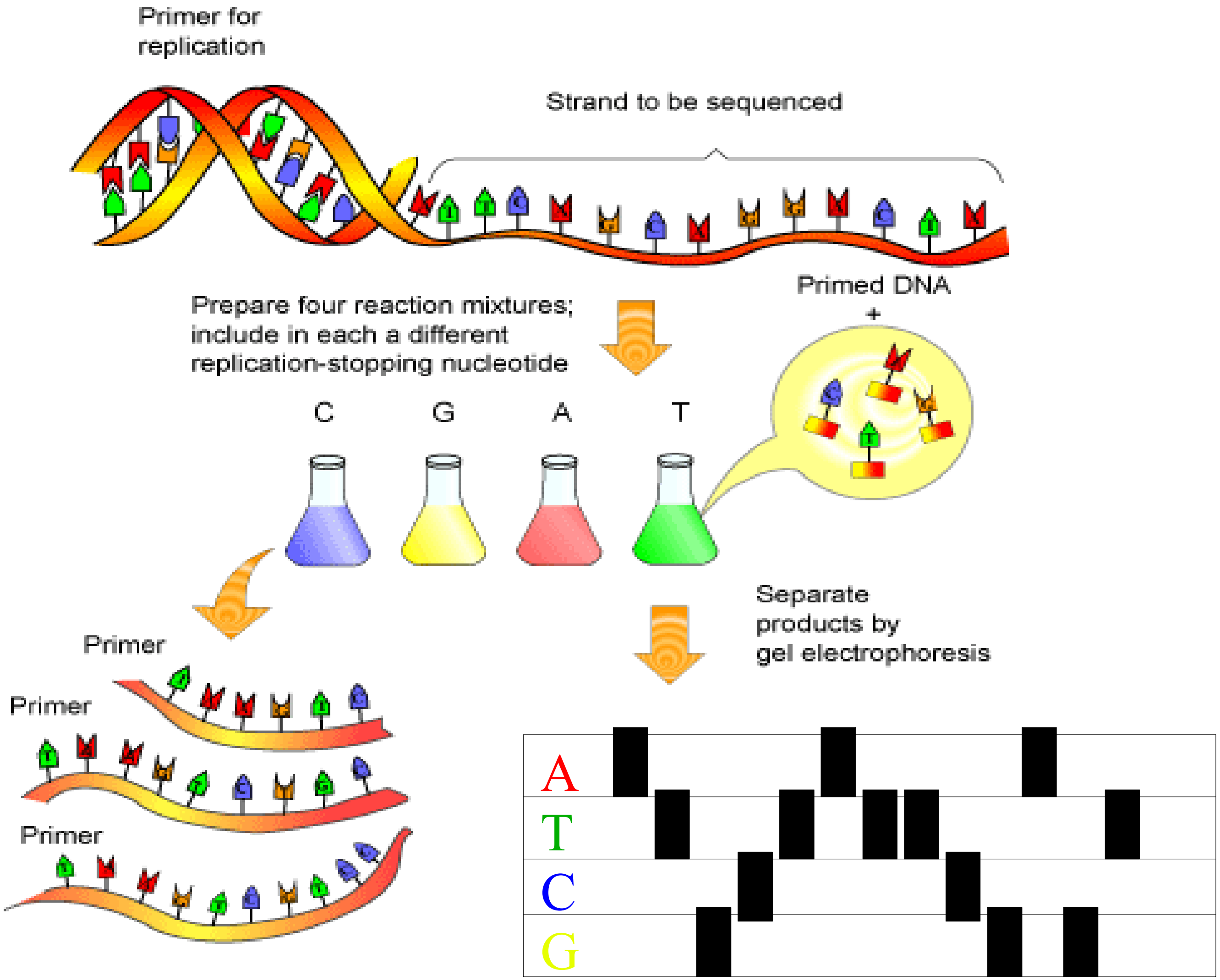
Thermal denaturation

Repeated sequences denaturate at a well defined temperature T_c (important for DNA amplification through Polymerase-Chain Reaction).
By synthetizing all 10 possible repeated sequences and measuring T_c the base pair free energies g have been evaluated.
Conversely, from T_c and the table of g, one could extract the sequence …

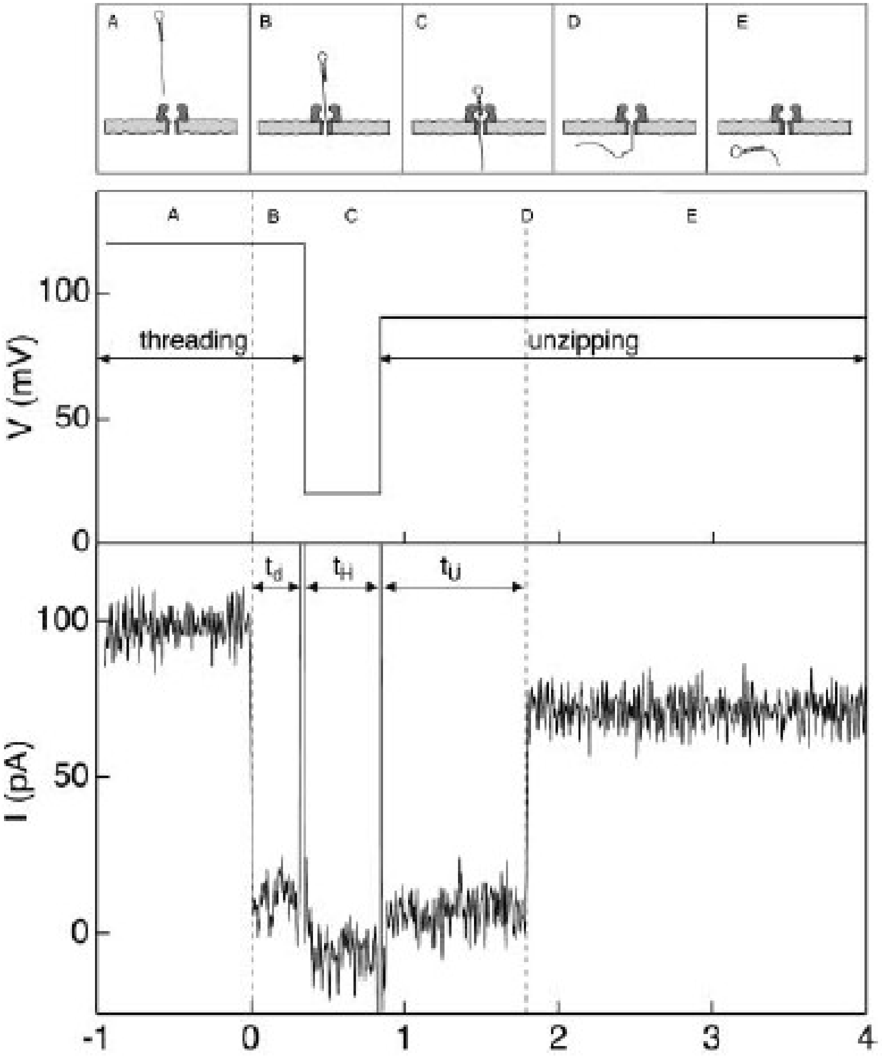


For heterogeneous sequences, AT vs GC rich region but where?

Methods for DNA sequencing



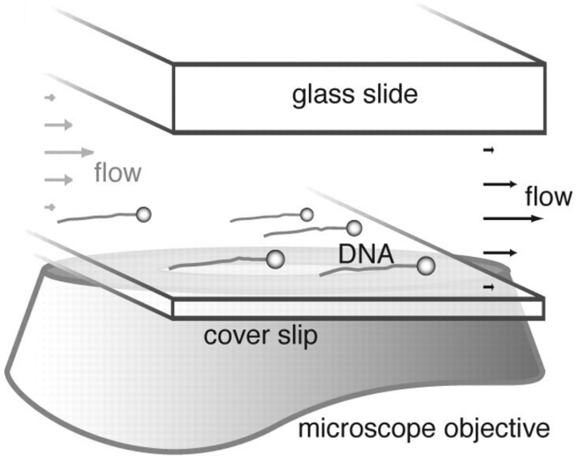
DNA translocation through a nanopore

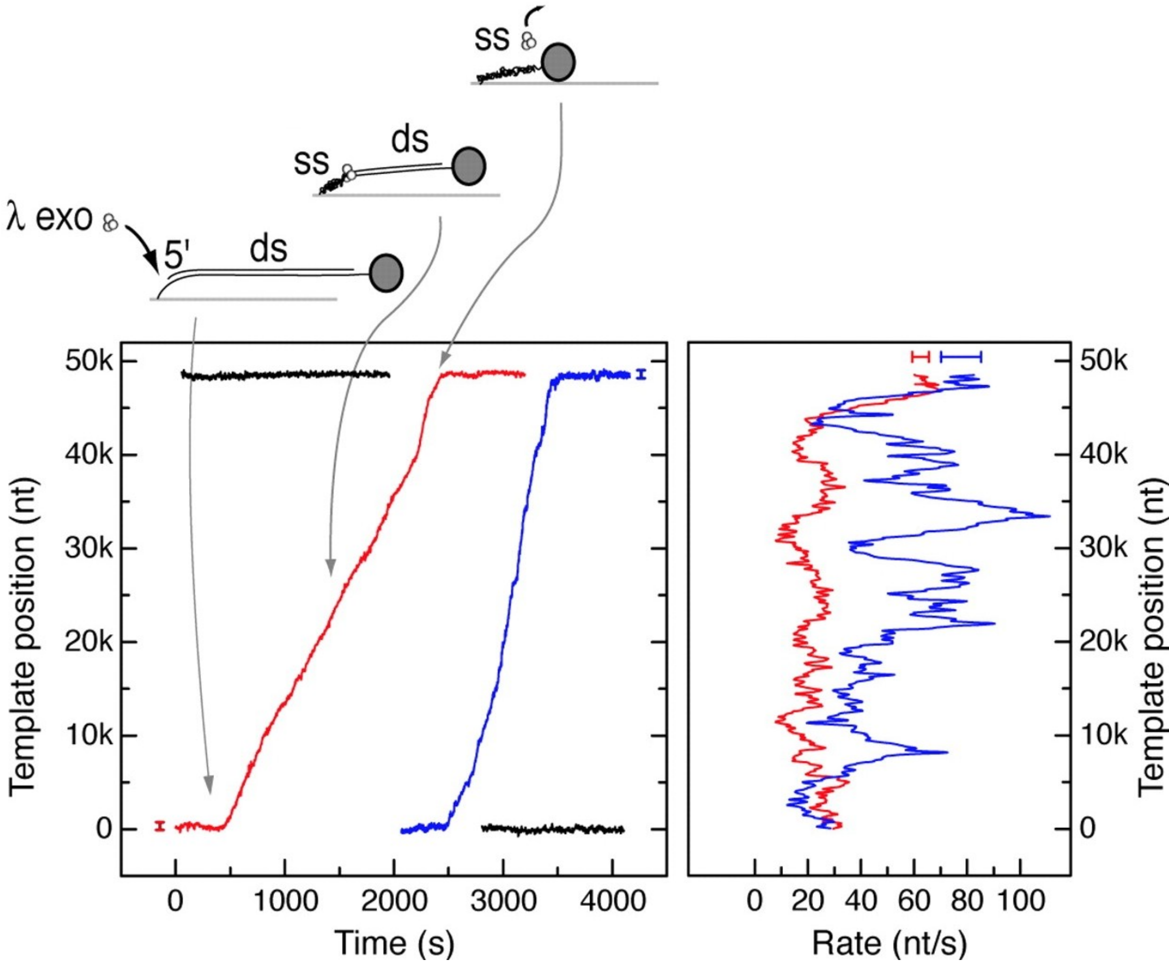


Mathé et al. (2004)

DNA digestion by the λ-exonuclease

Van Oijen et al., *Science* (2003)

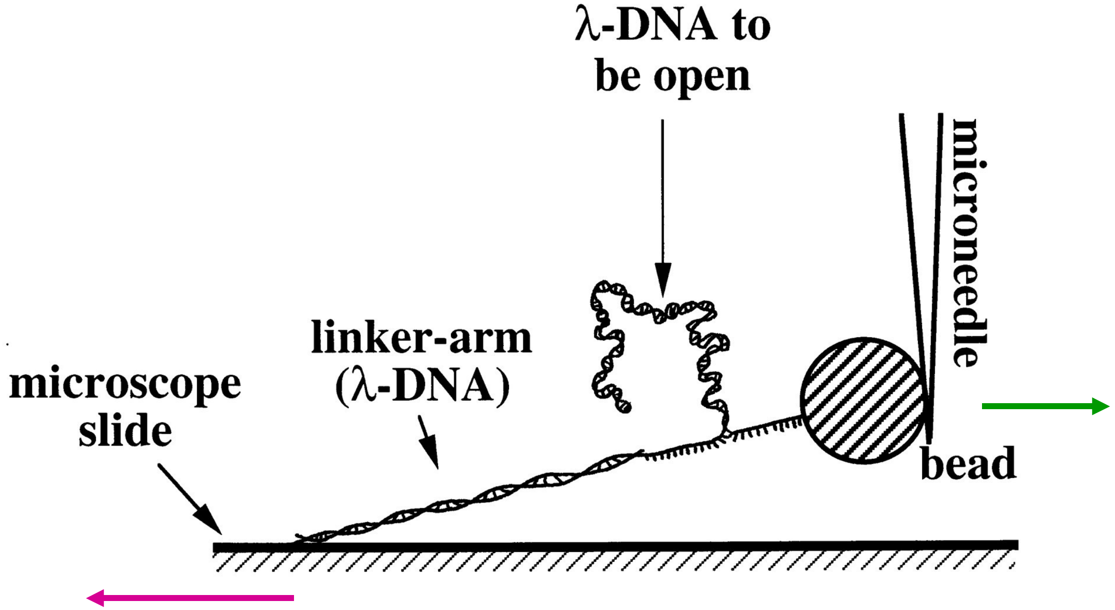




Mechanical Unzipping of DNA

Lambda-DNA : virus,
48502 base pairs

Essevaz-Roulet,
Bockelmann,
Heslot (1997)



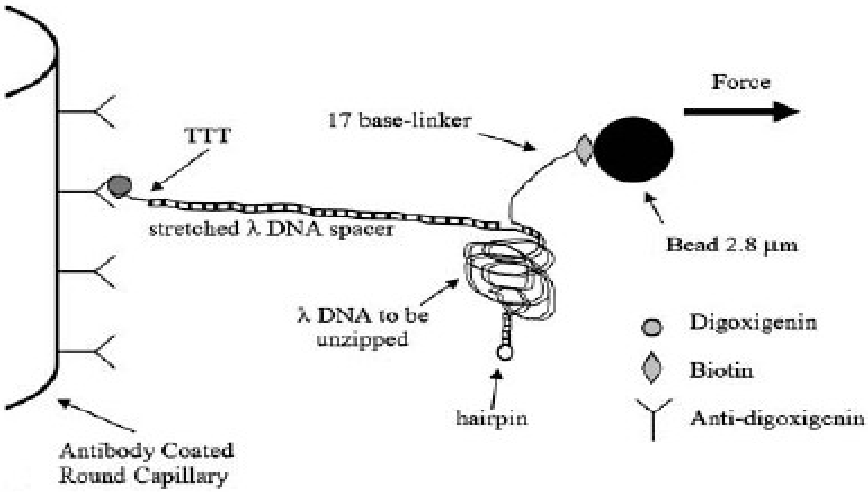
f (measured)

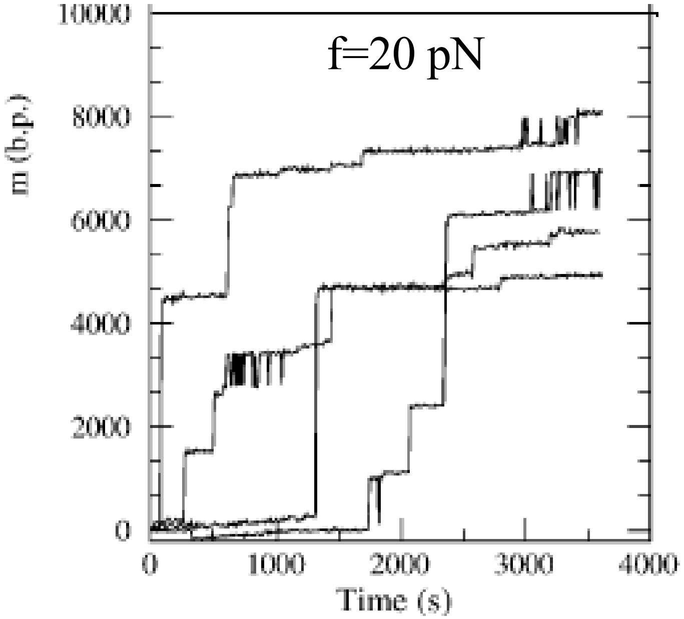
v or x (imposed)

Optical trap:
Spatial resolution ~ 1 nm
Temporal resolution ~ 1 kHz

Unzipping at constant force

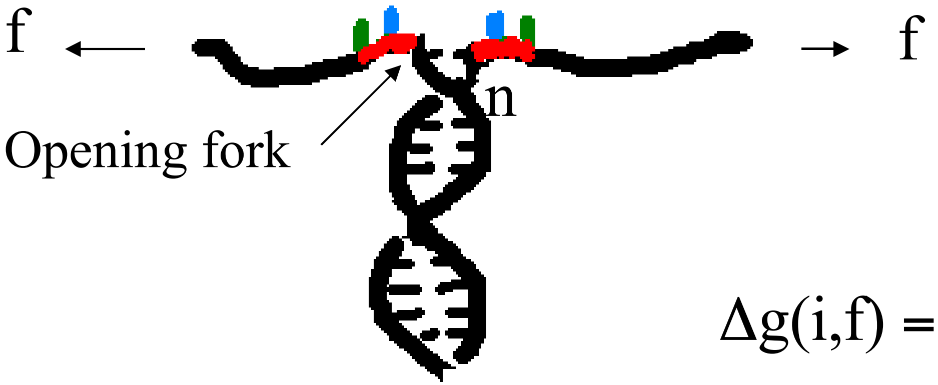
Weeks et al (2005)
Danilowitz et al. (2004)





Lionnet, Bensimon, Croquette (ENS): hairpin 200 bases

Thermodynamics of Unzipping



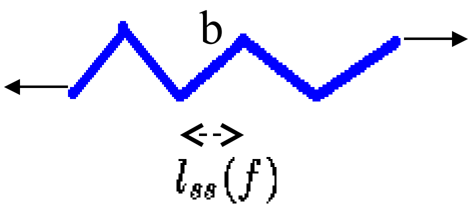
$$\Delta g(i,f) = g(b_i,b_{i+1}) - 2\ g_s(f)$$

- $g\ (b_i,b_{i+1})$ = base pair energy

- $g_s(f)$ = single strand DNA extension (ssDNA elasticity)

Elasticity of single-strand DNA (ssDNA)

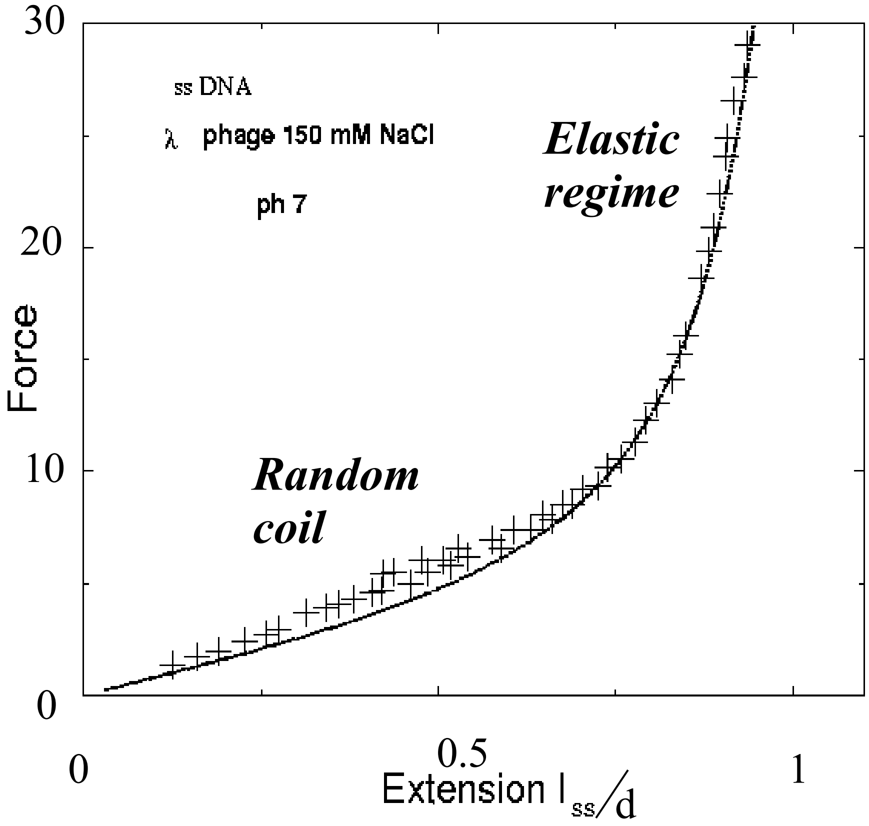
Freely-jointed chain model



$$x_{ss} = n\, l_{ss}(f)$$

$$l_{ss}(f) = d\left[Coth\left(\frac{f\,b}{k_B T}\right) - \frac{k_B T}{f\,b}\right]\left[1 + \frac{f}{\gamma_{ss}}\right]$$

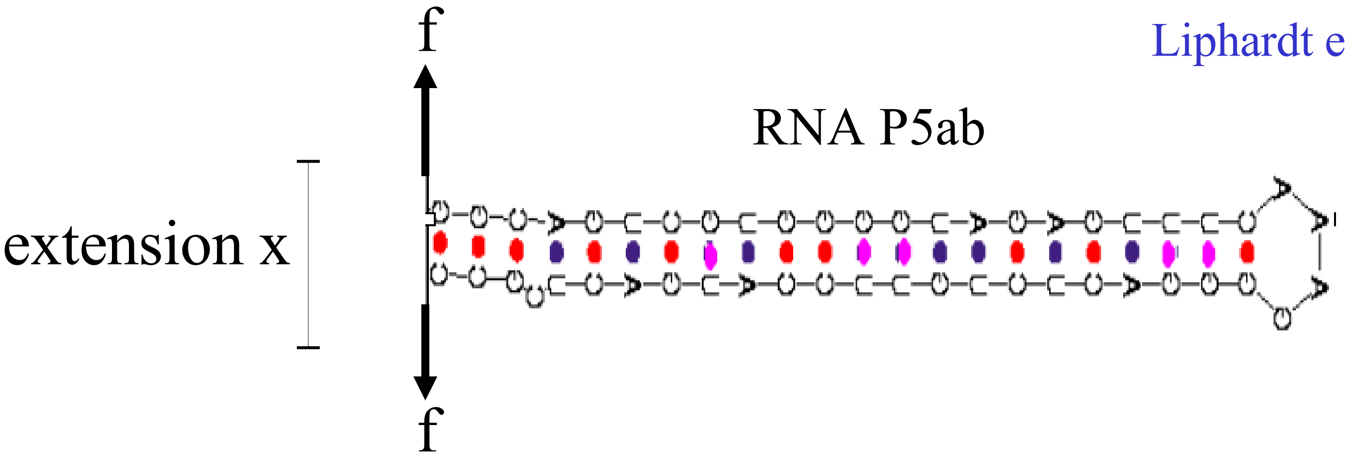
Kuhn length: b = 1.4 nm
monomer length: d = 0.56 nm
Young modulus: γ_{ss} = 800 pN

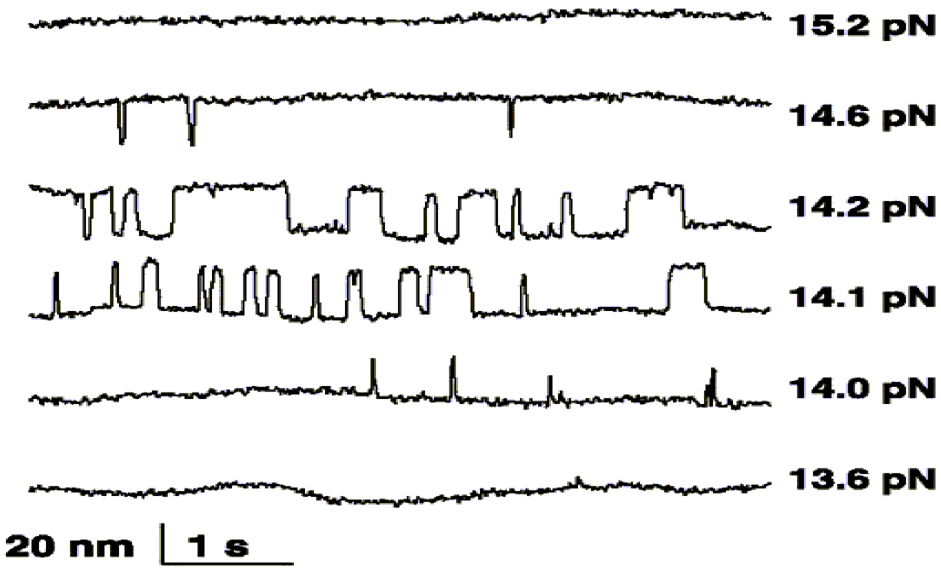


Bustamante et al. (1996)

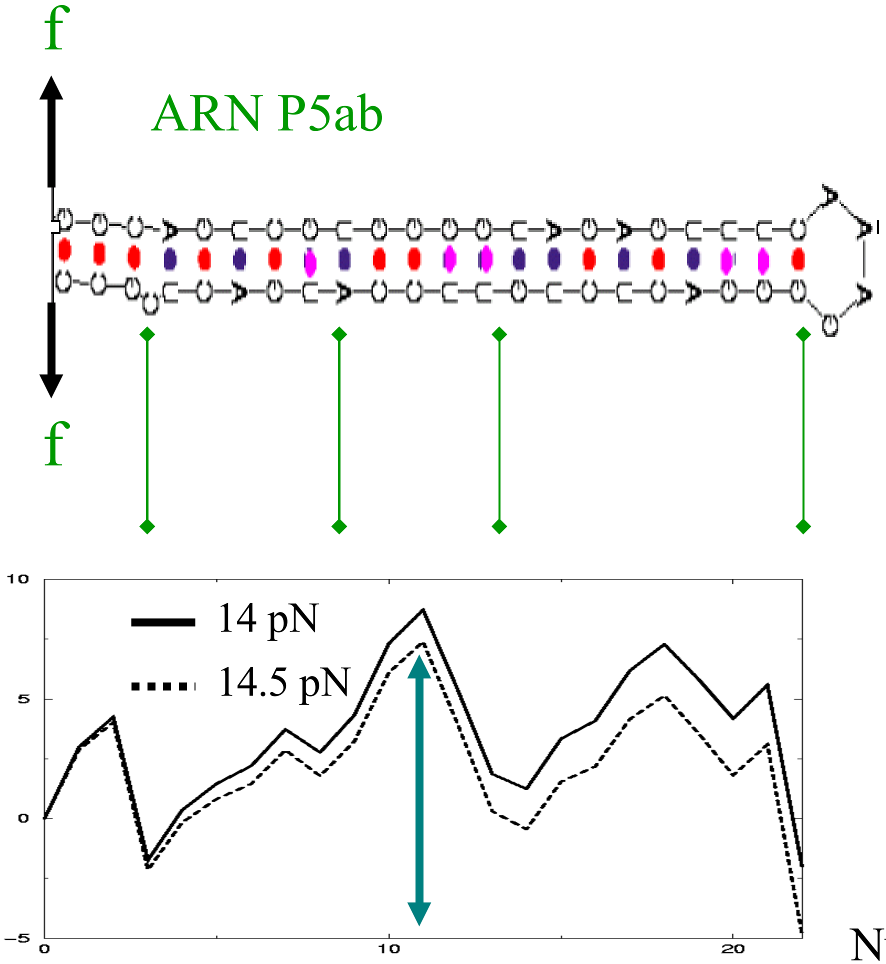
Example: opening a RNA double-helix

Liphardt et. al. (2001)





Sequence-dependent Landscape



Free energy

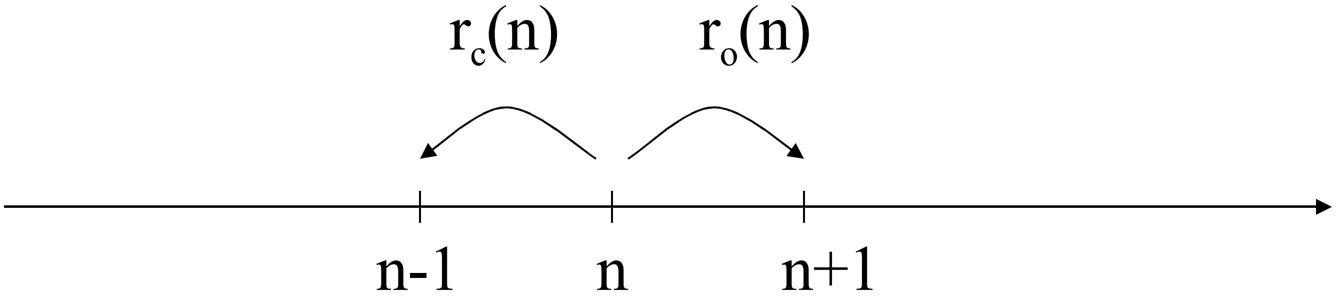
$$G(n) = \sum_{i<n} \Delta g(i,f)$$

barrier

Number n of open bases

Only two configurations can be seen in practice ...

Opening dynamics: random walk in random landscape

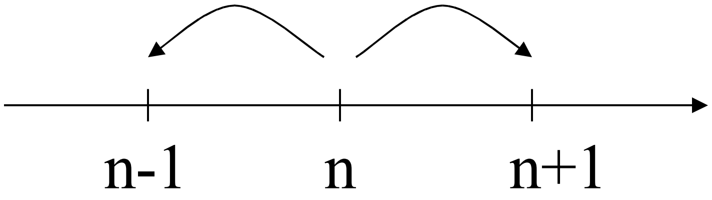


or, equivalently,

sojourn time on base n: t_n
$< t_n > \; = \; (r_c(n)+r_o(n))^{-1}$,
exponential distribution

&

$$q_n = \frac{r_c(n)}{r_c(n)+r_o(n)} \qquad 1\text{-}q_n$$



Transition matrix

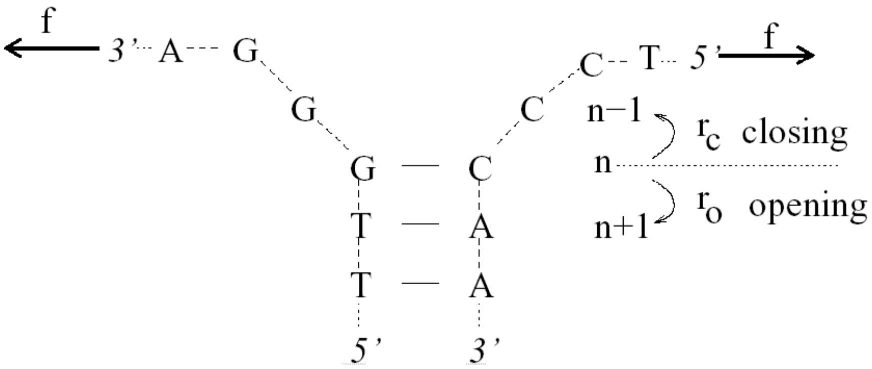
$$\text{rate matrix} = \mathbf{M} = \begin{pmatrix} -r_c(1) & r_c(2) & 0 & 0 \ldots \\ r_o(1) & -r_o(2)-r_c(2) & r_c(3) & 0 \ldots \\ 0 & r_o(2) & -r_c(3)-r_o(3) & \ldots \\ 0 & 0 & r_o(3) & \ldots \end{pmatrix}$$

$$\boxed{\frac{dP_n(t)}{dt} = \sum_{n'} M_{n,n'} P_{n'}(t)}$$

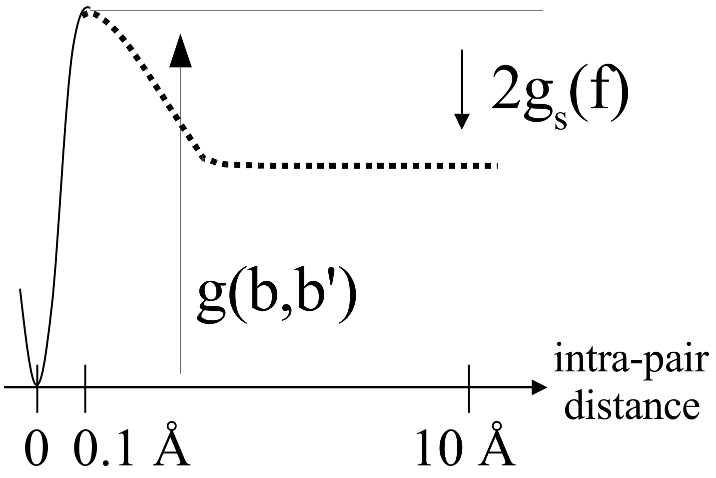
Transition matrix M?

1. equilibrium: $P_n = \exp(-G(n))/Z$
2. Rates depends on the local content of the sequence i.e. $r_o(j)$ depends on b_{j+1}, b_j …

Dynamical rates



Rates ?

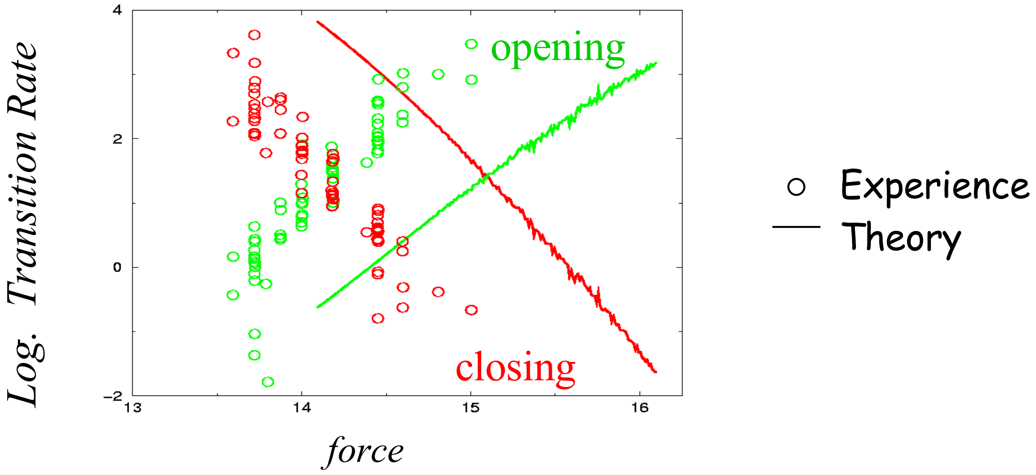


+

detailed balance

$r_o(n) = r \exp(- g(b_n,b_{n+1}))$ and $r_c(n) = r \exp(- 2g_s(f))$

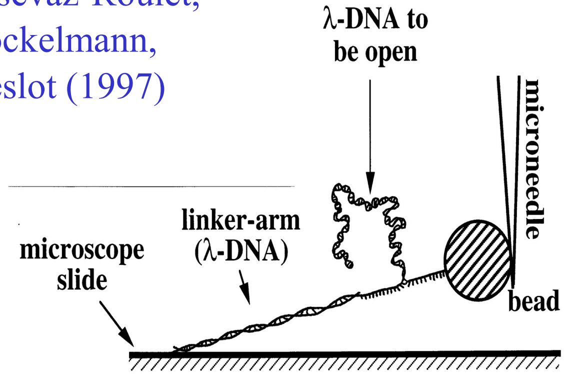
Check of the rate expressions

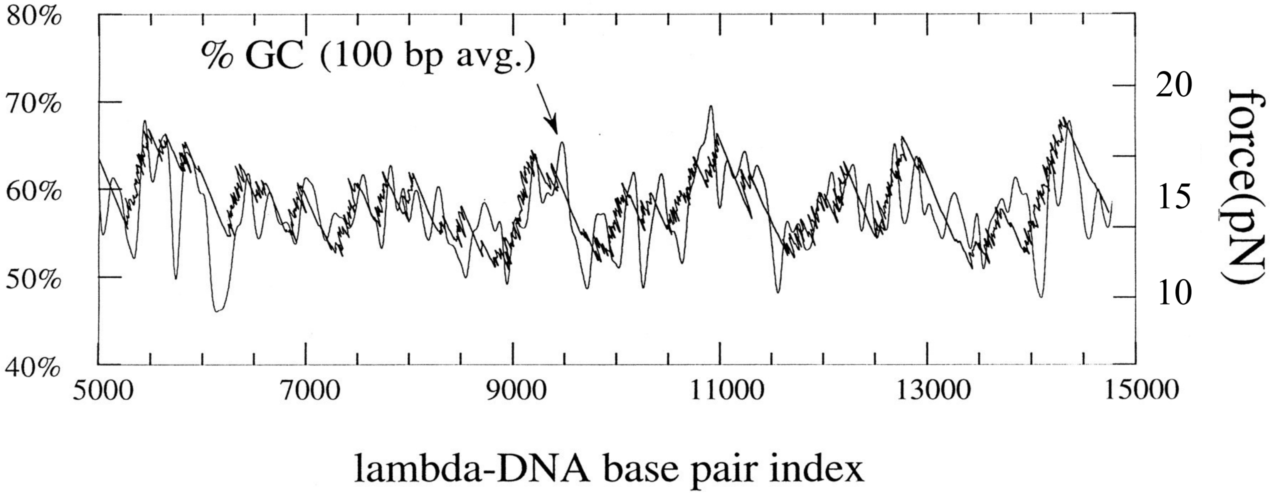


- The dependence on force seems to be ok
- Energetic parameters have to be tuned ...

Correlation between sequence and unzipping force

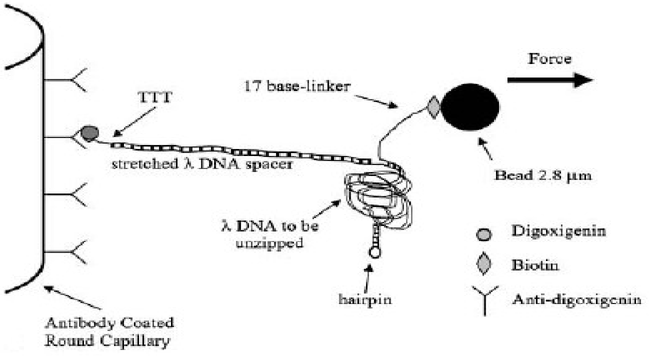
Essevaz-Roulet, Bockelmann, Heslot (1997)

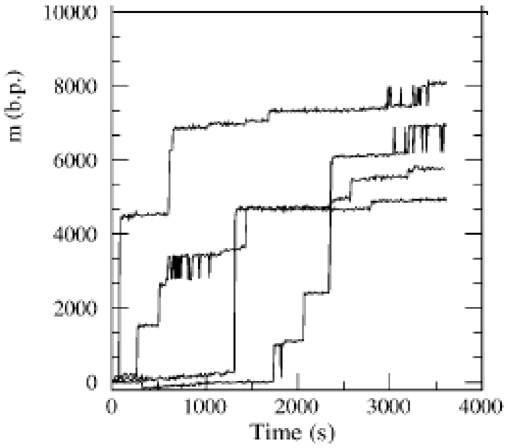


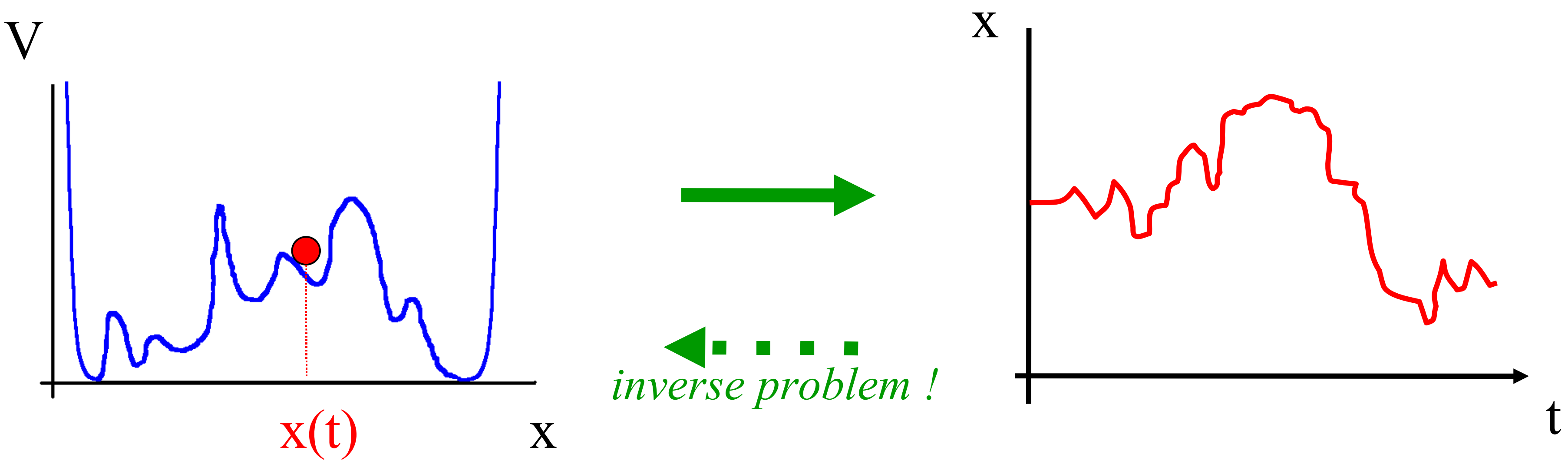


Random Walks in Random Landscape & Inverse Problem

Weeks et al (2005)
Danilowitz et al. (2004)







inverse problem !