DNA-protein electrostatic recognition: analysis of Protein Data Bank structures of DNA-protein complexes

A. G. Cherstvy[1], A. B. Kolomeisky[2], and A. A. Kornyshev[3]

[1] *Institut für Festkörperforschung, Theorie-II, Forschungszentrum Jülich, D-52425 Jülich, Germany*

[2] *Department of Chemistry, Rice University, Houston, Texas 77005 , USA*

[3] *Department of Chemistry, Faculty of Natural Sciences, Imperial College London, SW7 2AZ, London, UK*

Introduction to protein-DNA interaction and recognition

• DNA-protein recognition is vital for many biological processes (e.g., gene expression and regulation)

• Extreme diversity of proteins: humans ~500 000 proteins, ~ 25 000 genes.

• Protein classes: gene regulatory (transcription factors), repair proteins, structural proteins (**histones**), processing proteins (RNA Poly), etc.

• Main interactions: hydrogen bonding (HB), **electrostatic** (DNA/proteins), hydrophobic, van der Waals forces.

• Protein recognition motifs: helix-turn-helix, zinc finger, leucine zipper.

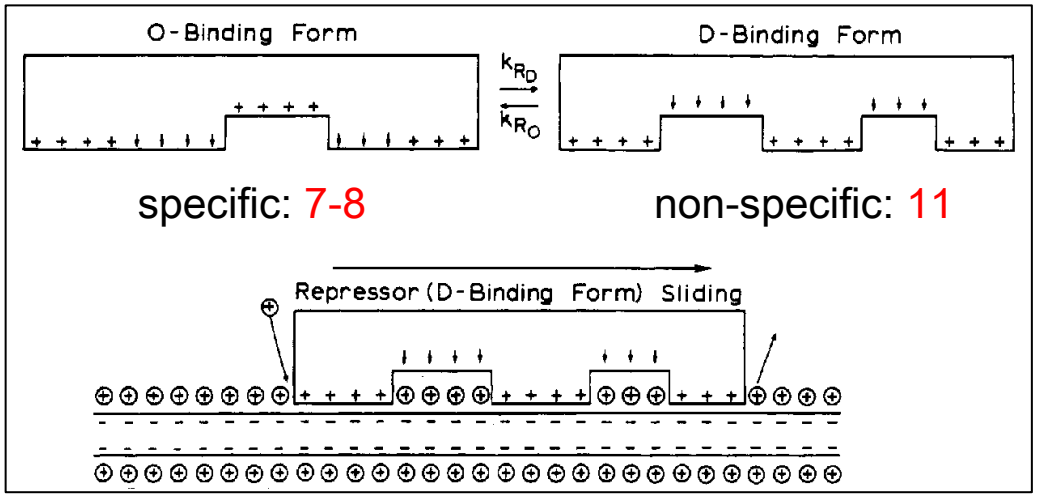
• Complex and rather probabilistic code of DNA-protein recognition.

• Protein-DNA binding affinity: **DNA sequence**, pH, [salt], *T*, helper proteins, DNA 3D conformation, etc.

• Physical mechanisms behind electrostatic DNA-protein interactions.

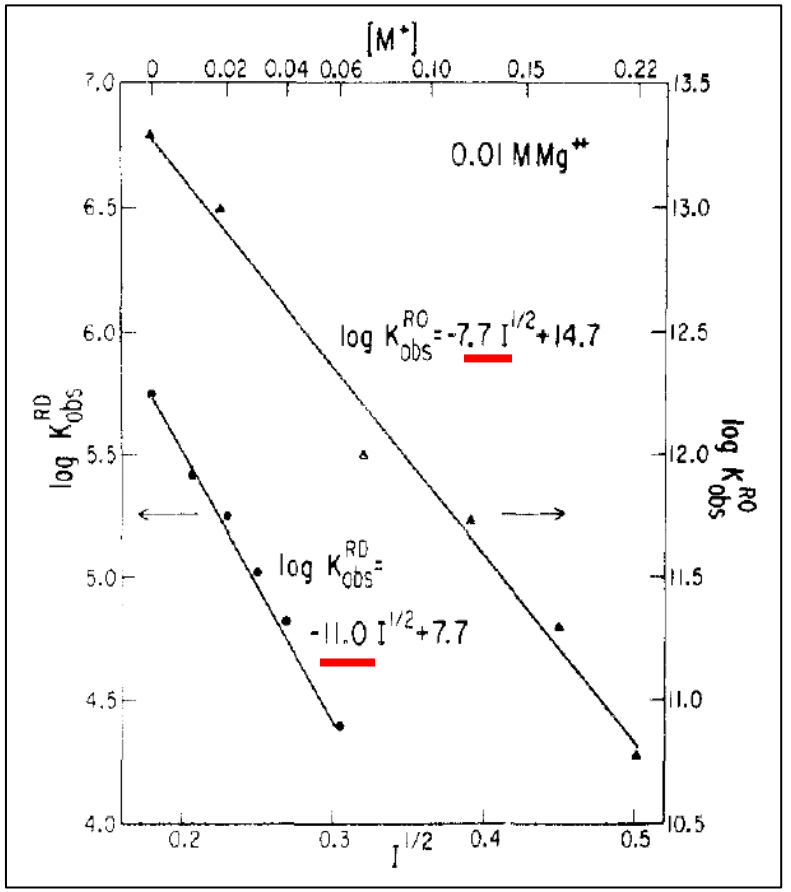
Electrostatic DNA-protein interactions: *lac* repressor

R B. Winter et al., Biochem., 20 6961 (1981)



Upon sliding, condensed cations are removed in front and they bind back on DNA behind the protein.

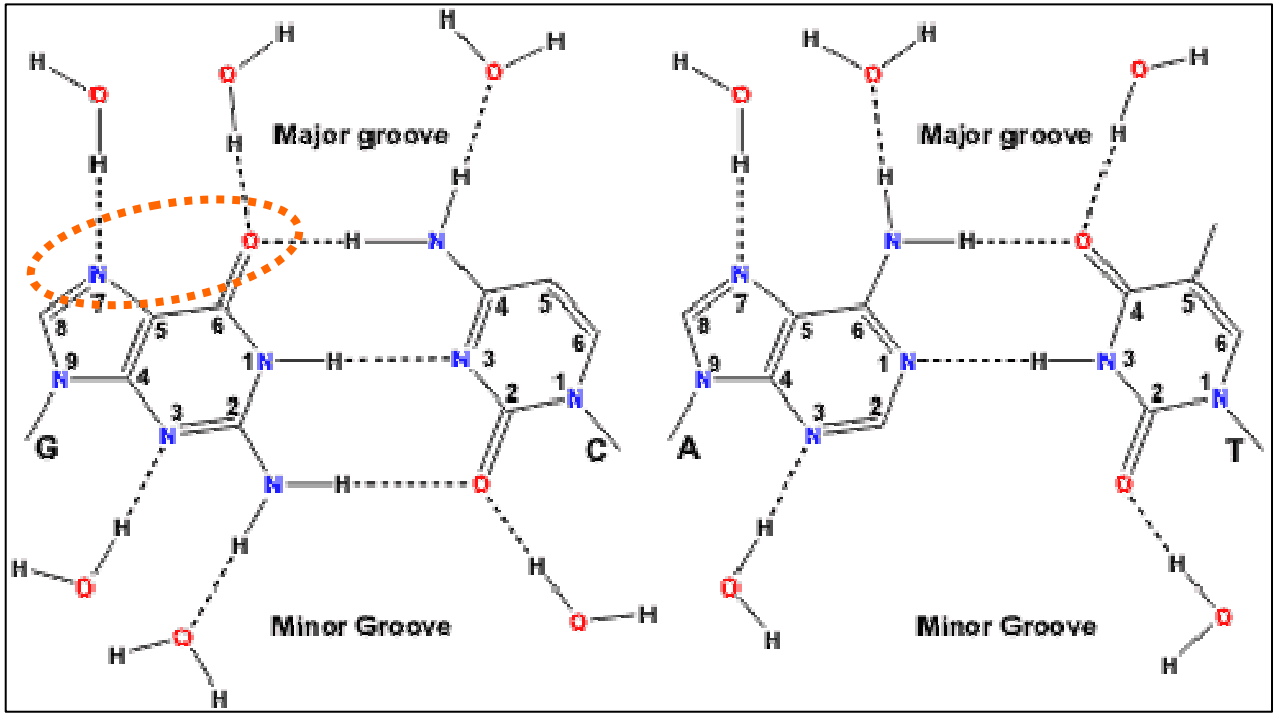
M.T. Record et al., Biochem., 16 4791 (1977)



Enormous dependence of *lac* repressor association binding constant *K* on [salt]

Winter & von Hippel: Electrostatic DNA-protein interactions are largely sequence non-specific !?

Hydrogen bonds with DNA bases: DNA-protein recognition code



• HB donors and acceptors determine the unique code of DNA-protein HB interactions; HB strength is 1-5 k_BT

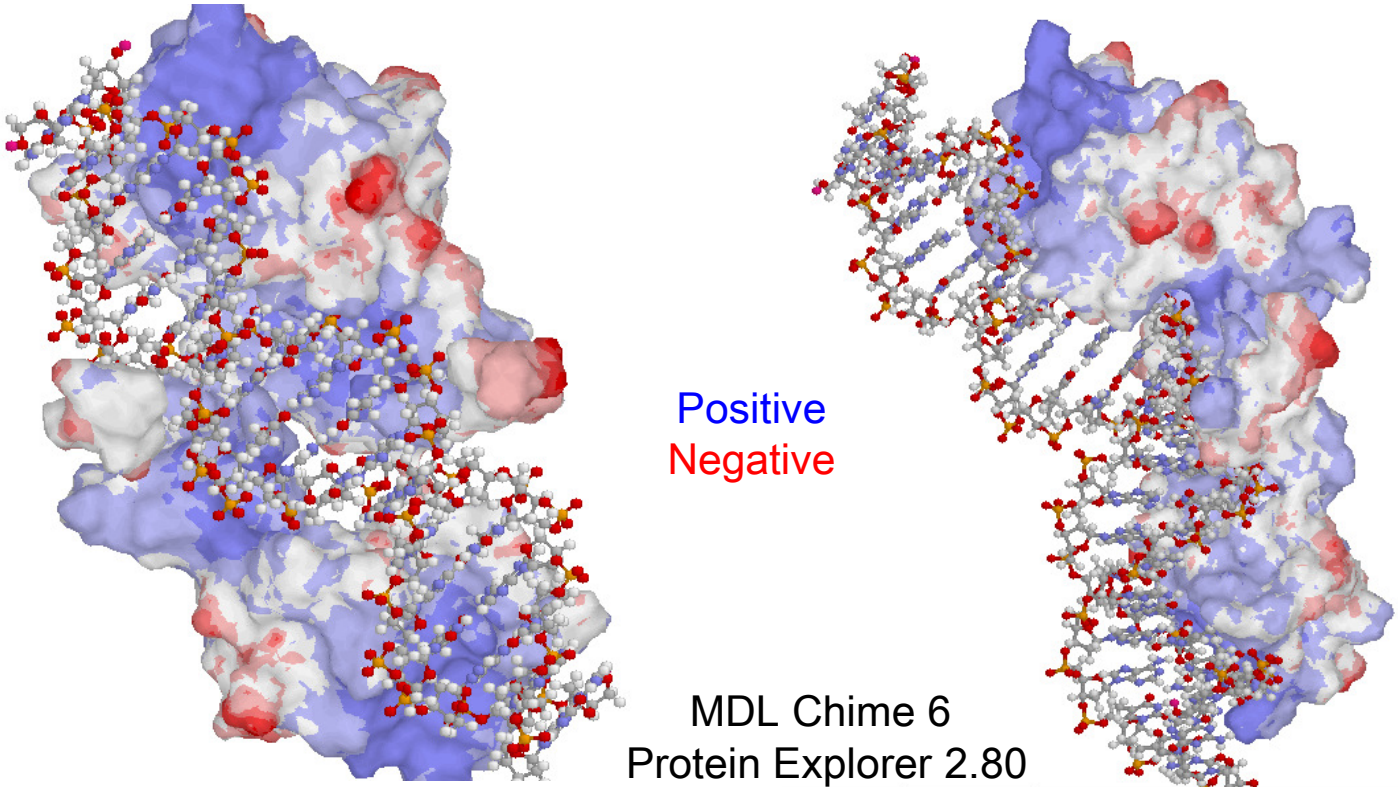
• HB formation preferences in DNA-protein complexes: Arg NH1/NH2 and Lys NZ with O6 and N7 of Guanine, Asn and Gln with Adenine, Glu and Asp with Cytosine.

N. M. Luscombe et al., Nucl. Acids Res., 29 2860 (2001)

Electrostatic potential of *lac* repressor

Non-specific: 1osl.pdb

Specific 1l1m.pdb



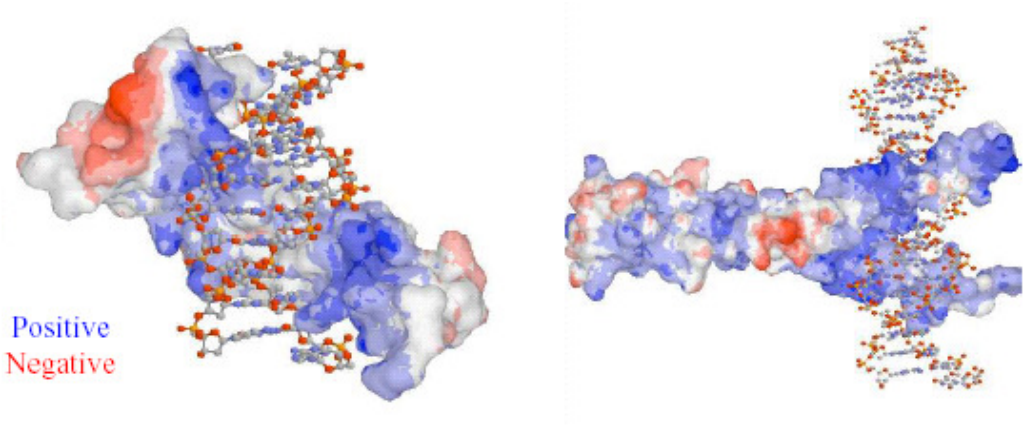
Positive
Negative

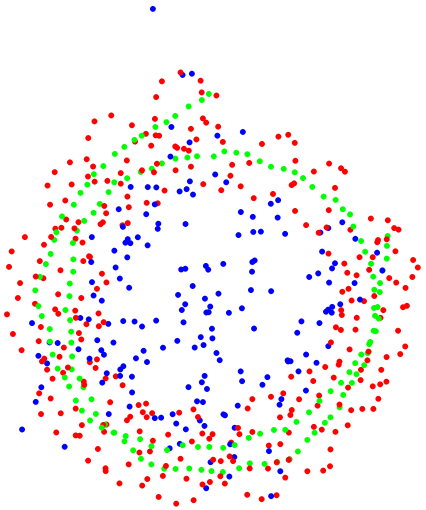
MDL Chime 6
Protein Explorer 2.80

• Protein residues Lysine (pK_a=10), Arginine (12), Histidine (6.5) are in close proximity to DNA phosphates

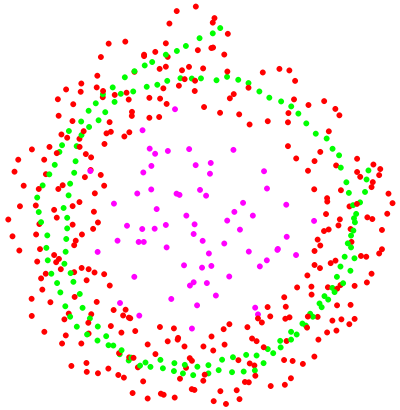
• DNA-induced charge patterns on proteins that are recognized by DNA?

Positive protein charges “love” DNA: sequence specificity of interactions?

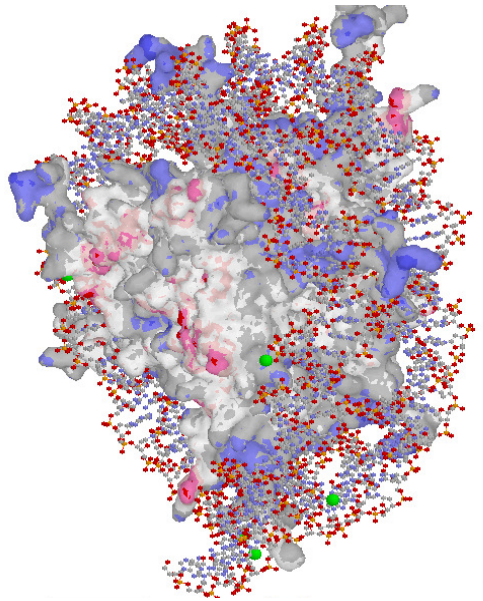




Nitrogens NZ on Lys, NH1/NH2 on Arg, and ND1 on His



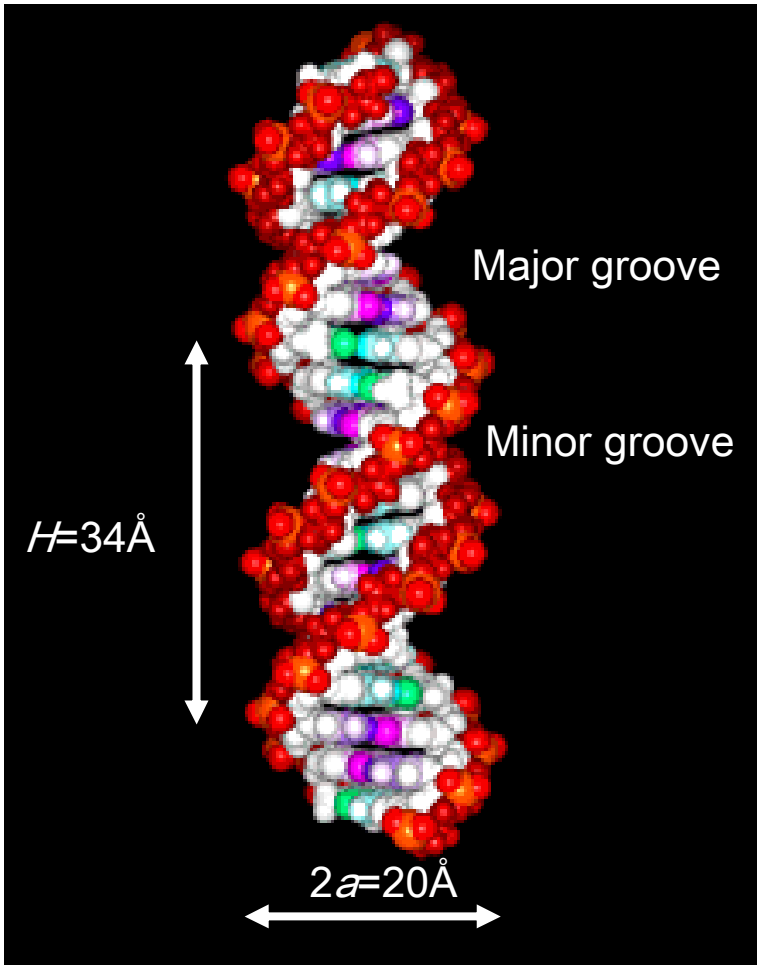
Oxygens OD1/OD2 of Asp and OE1/OE2 of Glu.



NCP stability ([salt])

B-DNA charge and structure non-ideality

-1 e_0 per each 1.7 Å along DNA axis



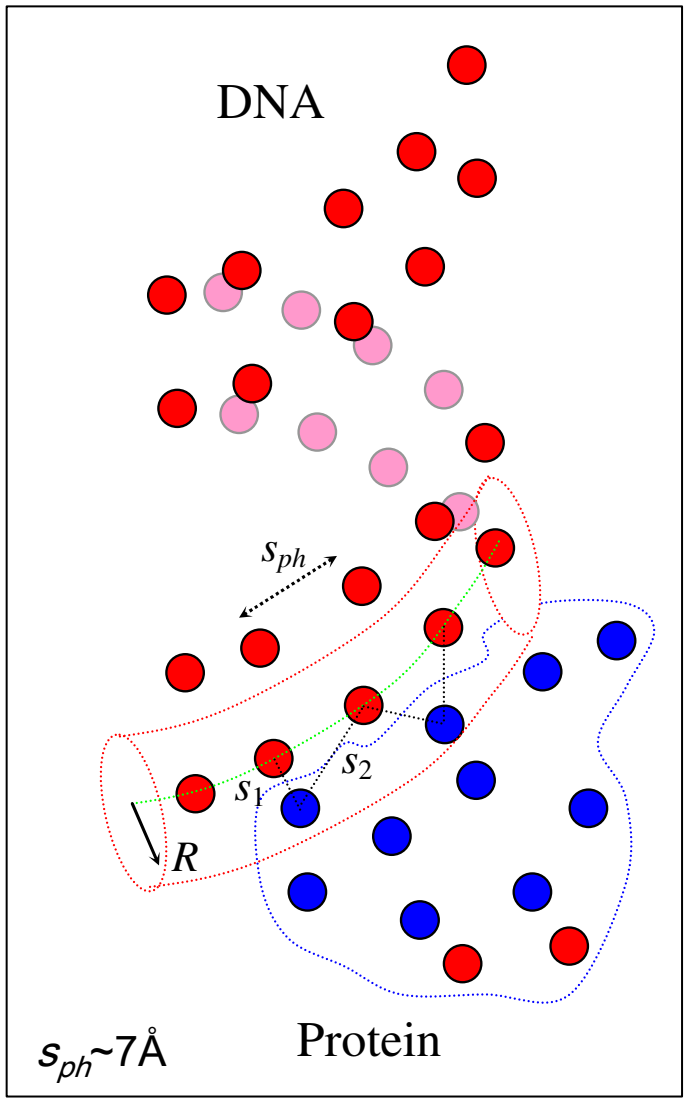
The 10 Twist Angles of B-DNA

Dinucleotide	Twist Angle (h)
(AA) · (TT)	35.6 ± 0.1
(AC) · (GT)	34.4 ± 1.3
(AG) · (CT)	27.7 ± 1.5
(AT) · (AT)	31.5 ± 1.1
(CA) · (TG)	34.5 ± 0.9
(CC) · (GG)	33.7 ± 0.1
(CG) · (CG)	29.8 ± 1.1
(GA) · (TC)	36.9 ± 0.9
(GC) · (GC)	40.0 ± 1.2
(TA) · (TA)	36.0 ± 1.0

W. Kabsch et al., Nucl. Acids Res., 10 1097 (1982)
W. K. Olson et al., PNAS, 95 11163 (1998)

DNA corrugated structure is recognized by proteins

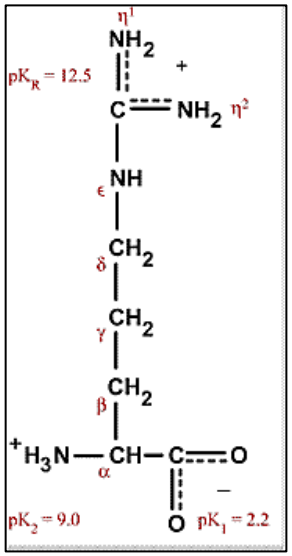
Model



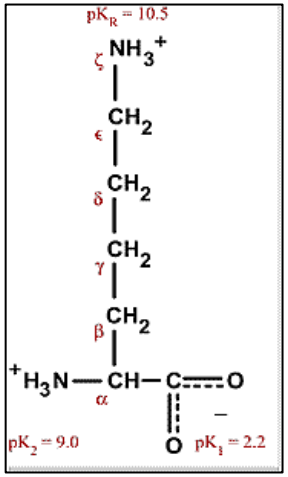
- Extract atomic coordinates from PDB files of protein-DNA complexes (Math 6)
- Identify closest protein N^+ charges, $R \sim l_B \sim 7$Å
- $s_{1,2}$ on the **same** DNA strand; DNA direction
- Histogram of s_1-s_2 distribution
- If uniform distr. $\Rightarrow$ no DNA sequence specificity
- Two-peaks distr. $\Rightarrow$ protein N^+ follow DNA P^-
- As 3D DNA structure is sequence specific, individual P^- are tracked by Lys and Arg
- Complementary DNA-protein interaction lattices
- Sequence-specific electrostatic interactions

Protein positive residues and DNA negative charges

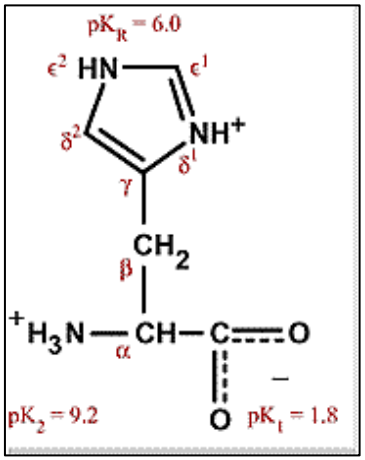
Arginine: N CA C O CB CG CD NE CZ NH1 NH2
10th and 11th atoms are N

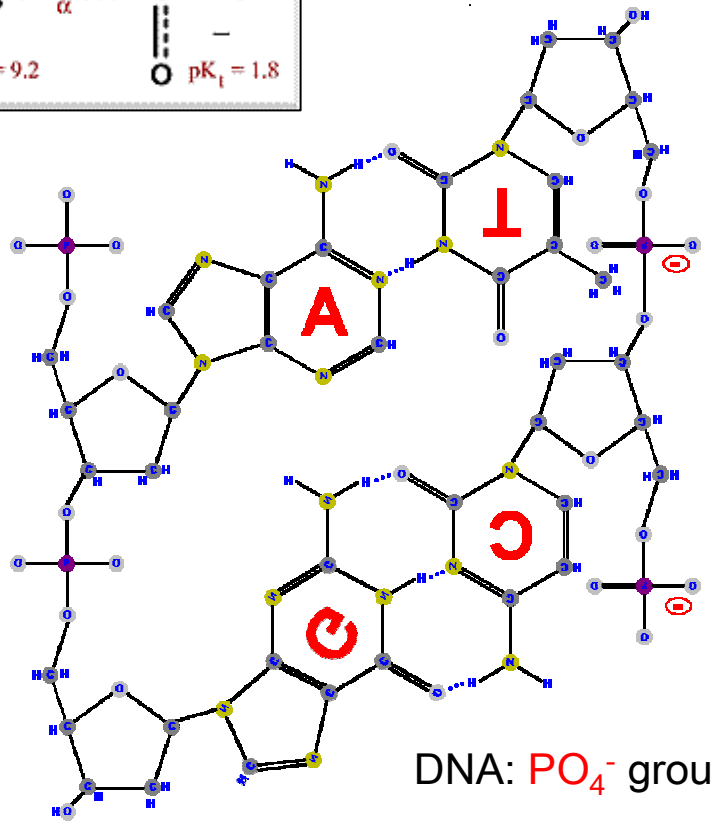


Lysine: N CA C O CB CG CD CE NZ
9th atom is N



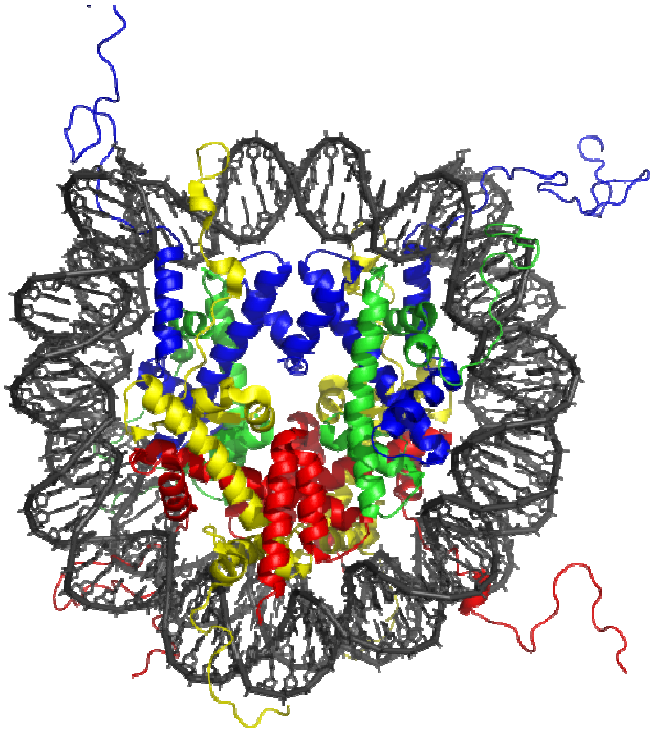
Histidine: N CA C O CB CG ND1 CD2 CE1 NE2
7th atom is N, charged or neutral





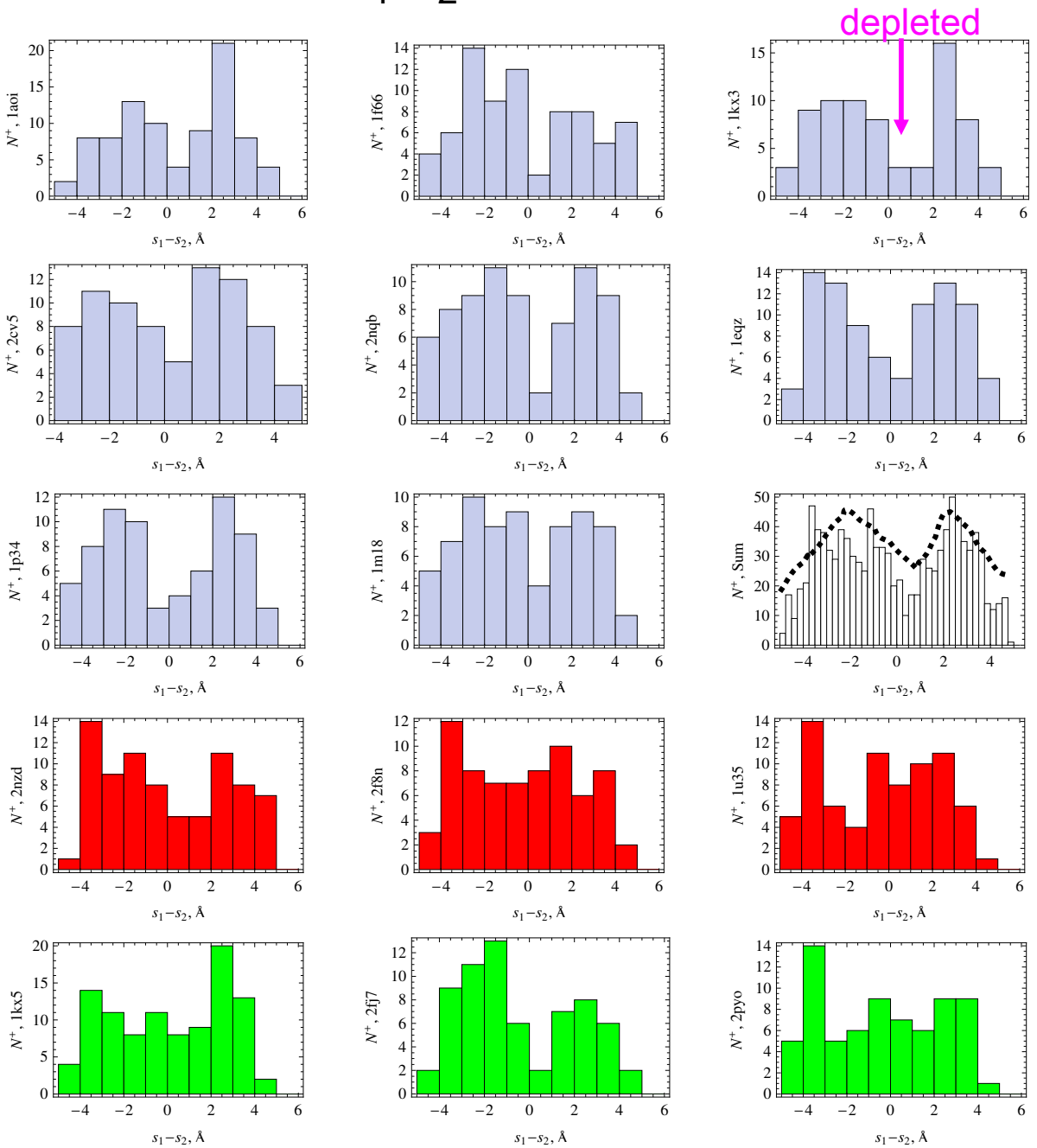
DNA: PO_4^- groups

Nucleosomes: DNA-wrapping proteins of eukaryotes



K. Luger et al., Nature, 389 251 (1997)

Results for s_1-s_2 in NCPs: 75-100 N⁺ close, 160-230 in total



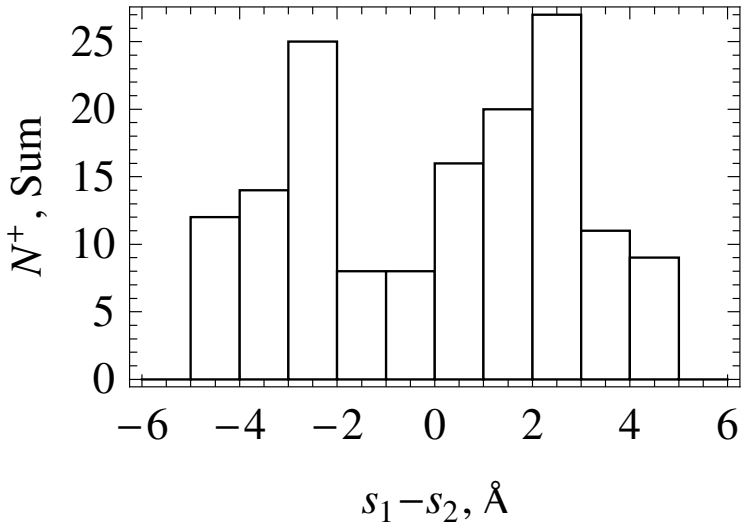
Canonical 146 bp NCP:
1aoi, 1f66, 1kx3, 2cv5,
2nqb, 1eqz. 1p34, 1m18

Sum of all complexes: frog, human, fruit fly, chicken NCPs

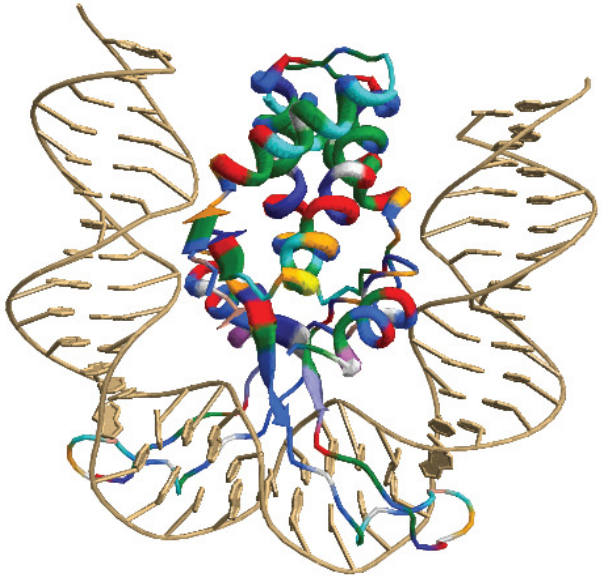
145 bp: 1nzd, 2f8n, 1u35

147 bp: 1kx5, 2fj7, 2pyo

Prokaryotic DNA-bending proteins also reveal two peaks

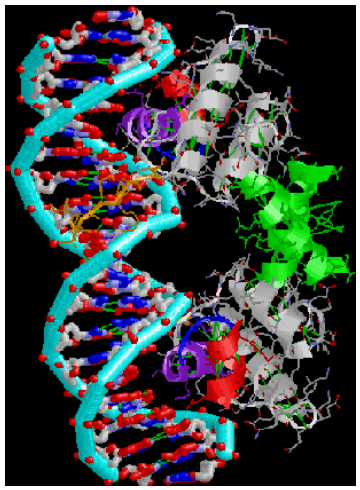


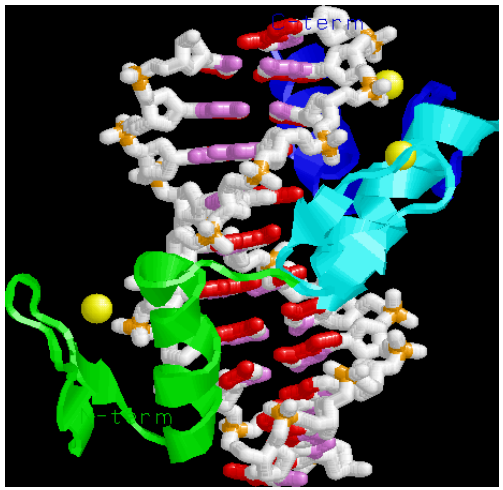
Complexes analyzed:
2np2, 1ihf, 1p51,1p71,
1p78, 1ouz, 1owf, 1owg

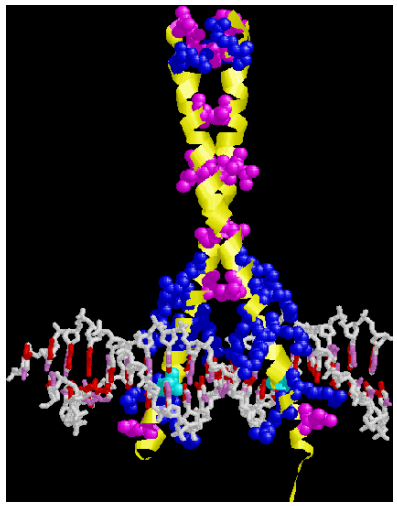


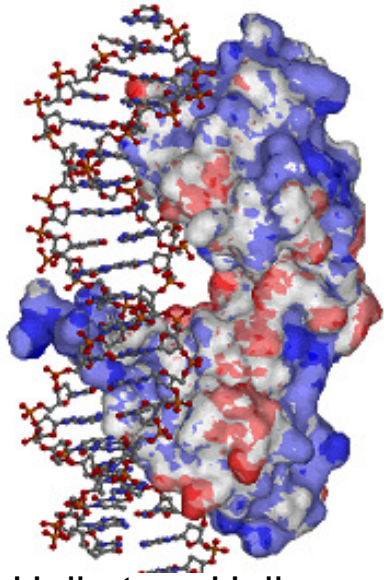
U-turn like severe
bending of DNA

Main DNA-binding motifs of proteins

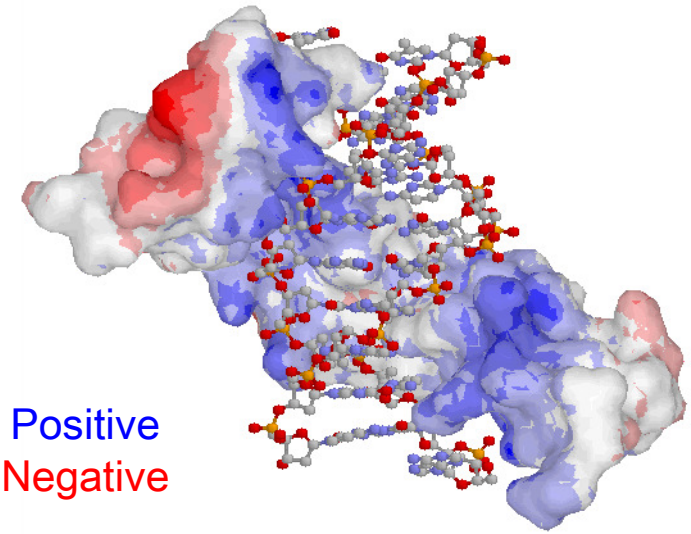


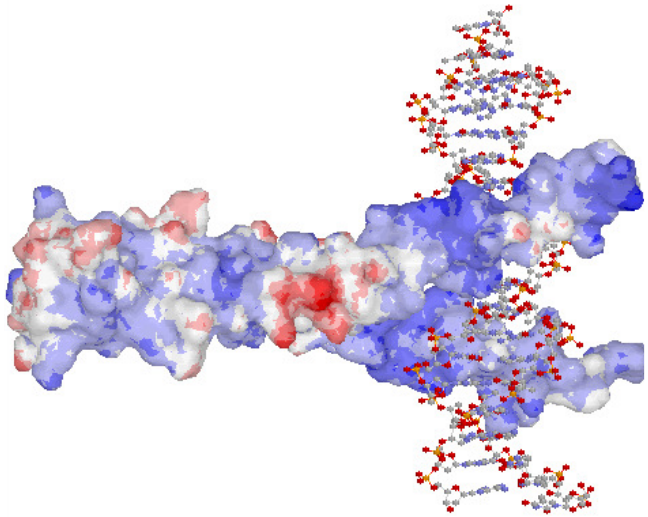






Positive
Negative



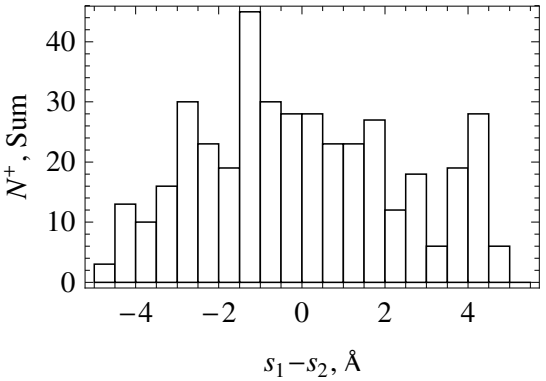


Helix-turn-Helix,
λ repressor, 1lmb.pdb:
2 α-helices in major groove,
HB with DNA bases,
ES with phosphates

Zinc finger, **Zif268**, 1aay.pdb:
3 α-helices in major groove,
each finger recognizes 3 bps,
HB+ES

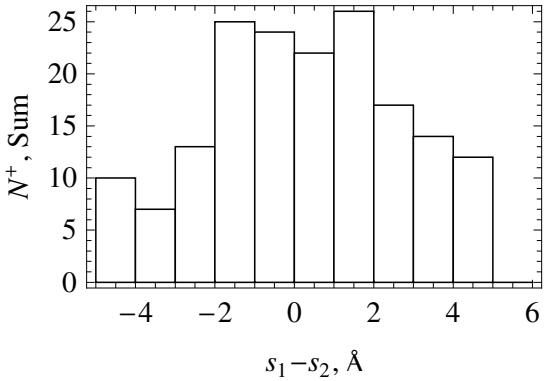
Leucine zipper, **GCN4**, 1ysa.pdb:
2 consecutive major grooves are
recognized by 2 long bound α-
helices, HB+ES

Basic DNA-binding protein motifs: uniform distributions and no sequence-specificity



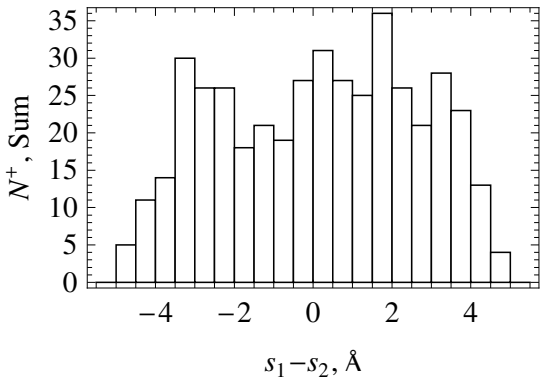
zinc fingers

1aay, 1a1l, 1p47, 1jk1, 1jk2, 1a1f, 1a1g, 1a1j, 1a1k, 1a1h, 1a1i, 1zaa, 1g2f, 1g2d, 1f2i, 1llm, 1mey, 1ubd, 1tf3, 2jp9, 2gli, 3dfx



leucine zippers

1ysa, 2c9l, 2c9n, 2h7h, 1d66, 1fos, 1gu5, 1hjb,



lac-, lambda-, 434-, cro-, arc-repressor-like complexes

Repressors (1osl, 1l1m, 2bjc, 1cjg; 1lmb, 3bdn, 6cro, 1lli, 1rio; 1par, 1bdt, 1bdv, 2bnz, 2cax; 1au7, 2or1, 1per, 3cro, 1rpe, 2p5l, 1gt0, 1hf0, 1ic8, 1o4x, 2r1j) and CAP proteins (1cgp, 1zrc, 1zrd, 1zre, 1zrf, 1o3q, 1o3r, 1o3s, 1j59, 1run, 2cgp),

Conclusions and outlook

• For **large** DNA-protein complexes, NCP and HU, tracking of individual DNA phosphates.
• DNA-induced s_{ph}~7 Å charge periodicity along DNA-protein interfaces.
• Up to 100 charge-charge contacts, large 10-30 $k_B T$ energy profit due to complementarity of DNA-protein charge lattices.
• Recognition of native and strongly bound DNA sequences.
• Nucleosome positioning on DNA, together with sequence-specific DNA bending code.
• 146 vs. 145/147 bp DNA NCPs. Different DNA affinities to histones? No data.

• For **small** complexes, with 5-10 ES contacts, no statistical preference and weak/no sequence specificity of ES interactions.
• Electrostatics is weak and other interactions contribute to recognition (HB).

• Interplay of HB+ES : future research.

- A. G. Cherstvy, A. B. Kolomeisky, and A. A. Kornyshev, J. Phys. Chem. B, 112 4741 (2008).

- A. G. Cherstvy, J. Phys. Chem. B, 113 4242 (2009).

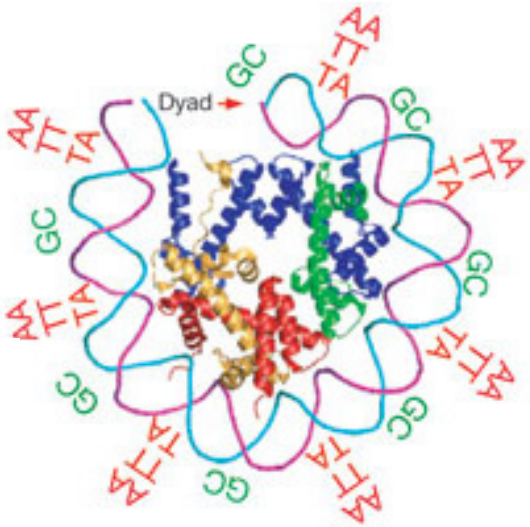
Thank you

Deutsche Forschungsgemeinschaft DFG

Grant CH 707/2-1

JÜLICH FORSCHUNGSZENTRUM

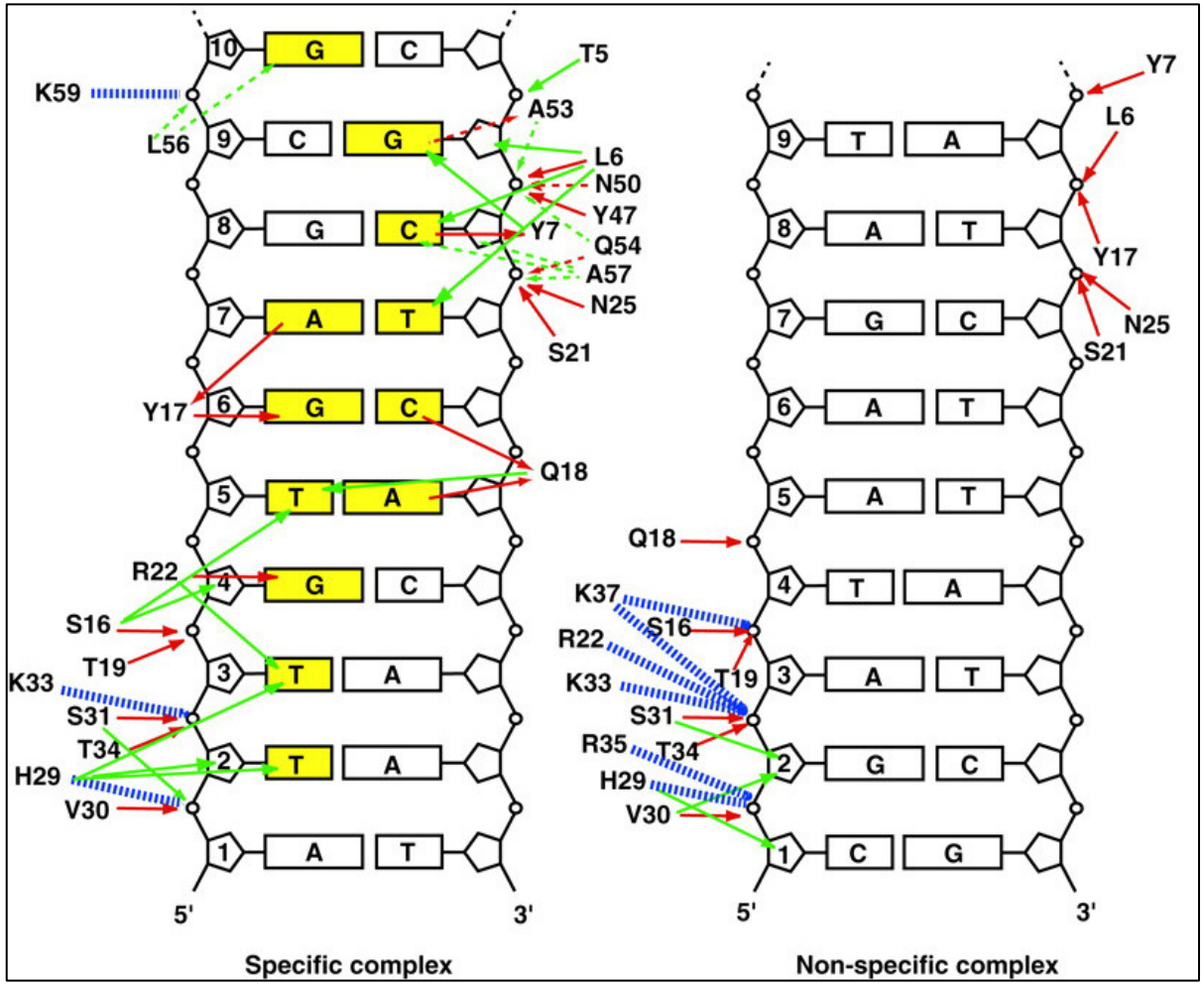
NCP positioning code



E. Segal et al., Nature 442 772 (2006)

Lac repressor contacts with DNA

hydrogen bonding hydrophobic electrostatic

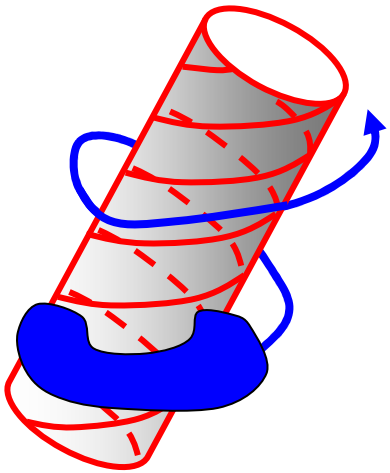


Electrostatic contacts are believed to be sequence-nonspecific,
while hydrogen bonding is highly sequence specific

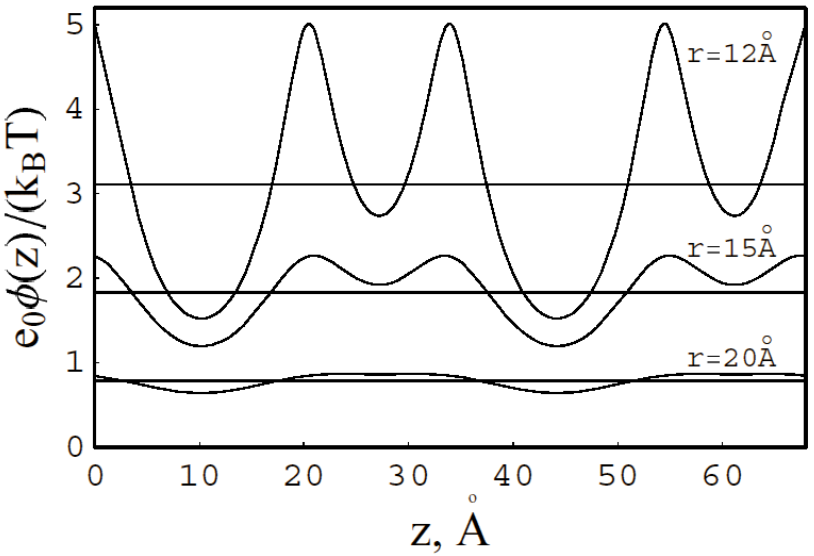
C. G. Kalodimos et al., Science, 305 386 (2004)

Sliding *vs.* Spiraling
electrostatic barriers *vs.* hydrodynamic friction

Spiraling RNA Polymerase:
protein binding requires DNA
groove tracking



K. Sakata-Sogawa et al.,
PNAS, 101 14731 (2004)



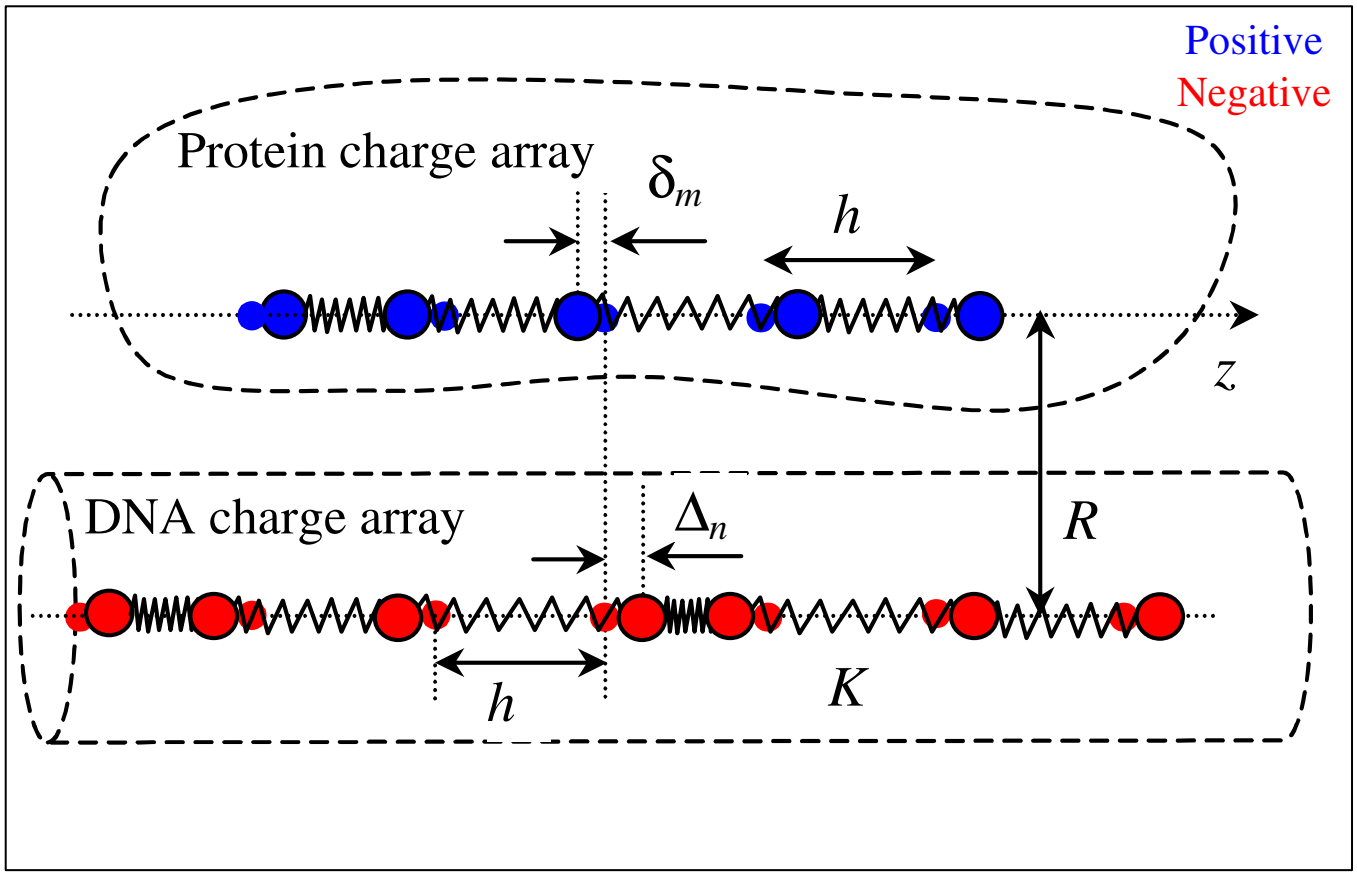
A.G. Cherstvy and R.G. Winkler,
J. Chem. Phys., 120 9394 (2004)

- Theory of Schurr for *lac* spiraling: 100 times stronger hydrodynamic drag and smaller D_1: $D_1=5\times10^{-9}$ cm^2/s

- Old experiments (Blomberg): $D_1=3\times10^{-9}$ cm^2/s

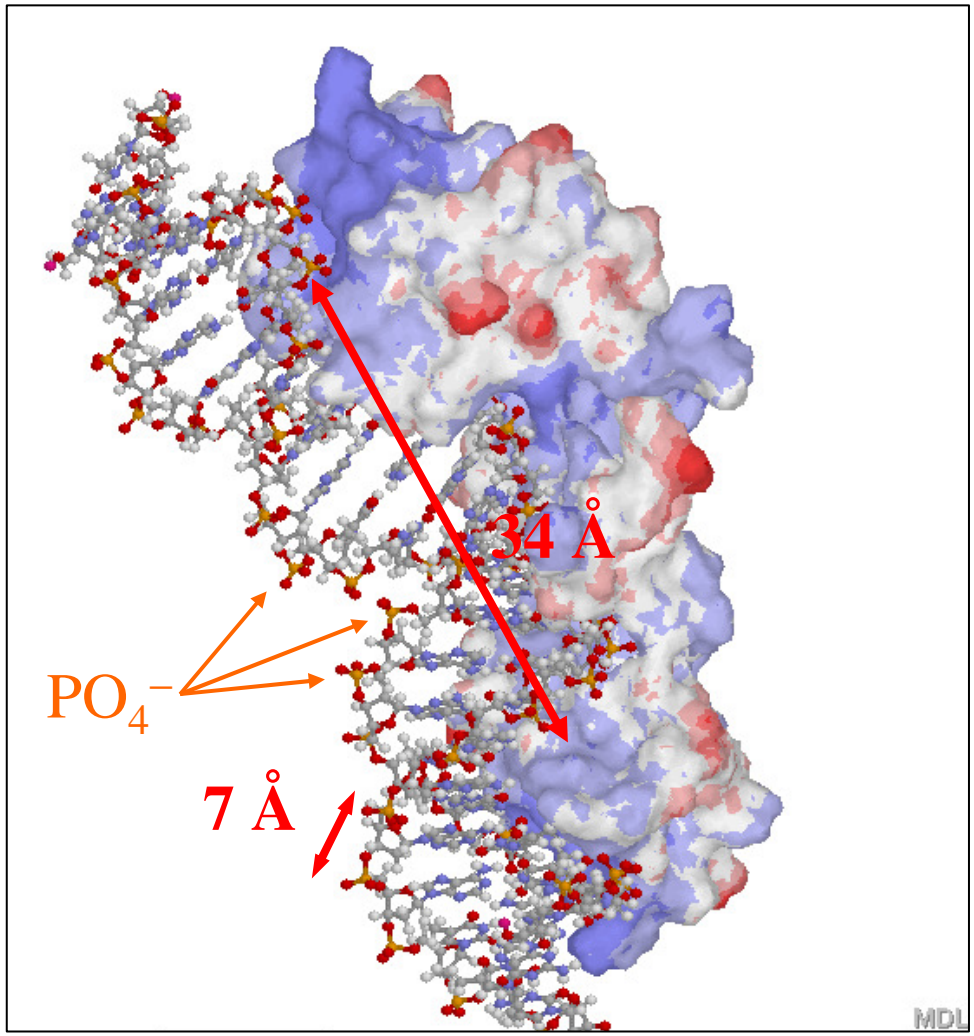
M. J. Schurr, Biophys. Chem., 9 41 (1979)

Model of DNA-protein recognition: charge complementarity



- Random charge displacements mimic bp specific nonideality of DNA/protein structure
- Long-range correlations $z_n = nh + \Delta_n$
- Recognition region -- similar charge variations $\boldsymbol{\Delta_n = \delta_m}$ -- stronger DNA-protein attraction
- Potential well near the homology region

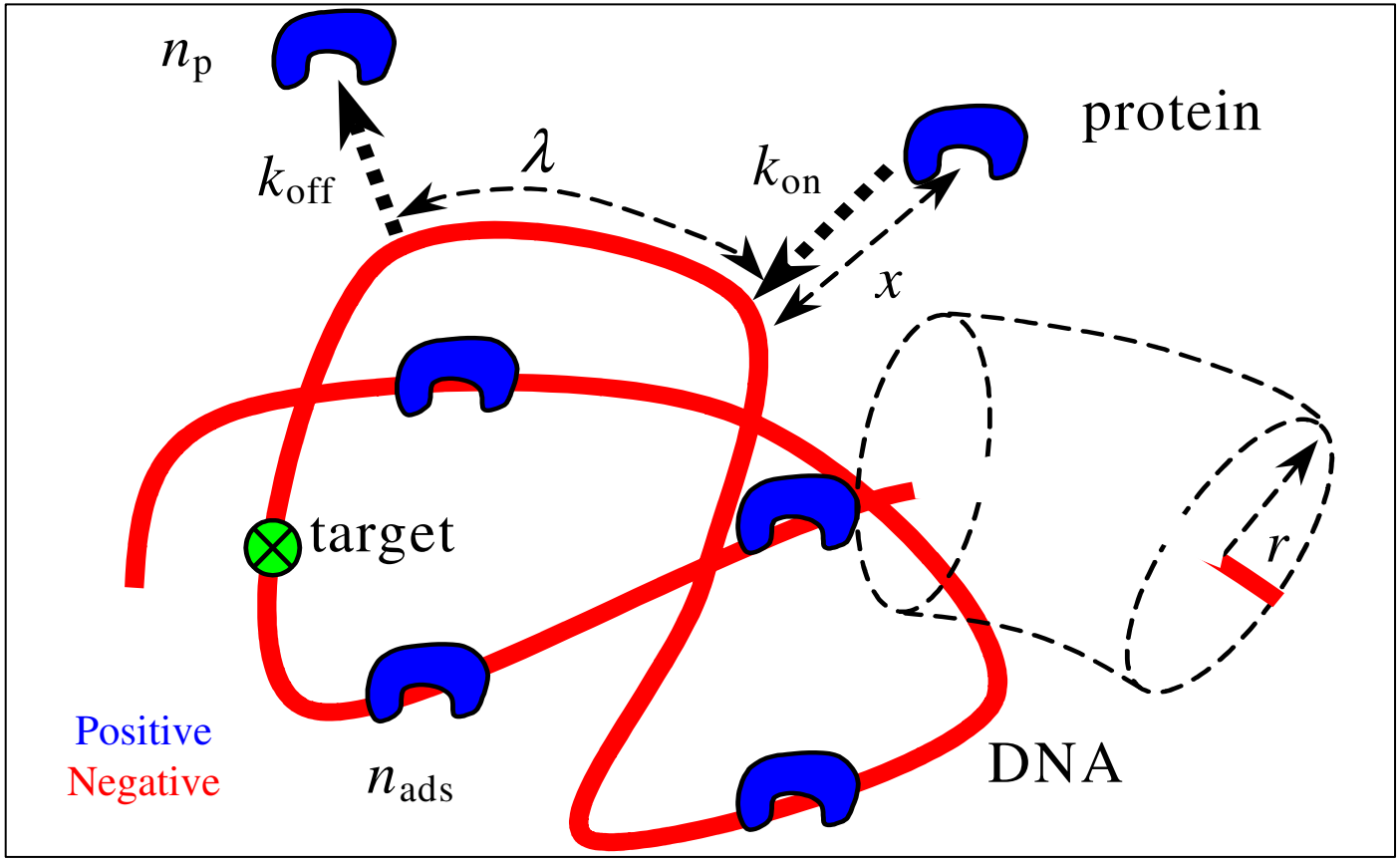
Artificial charge periodicity in protein DNA-binding domains



• Periodicity of ≈ 7 Å and ≈34 Å is expected from PDB data analysis.

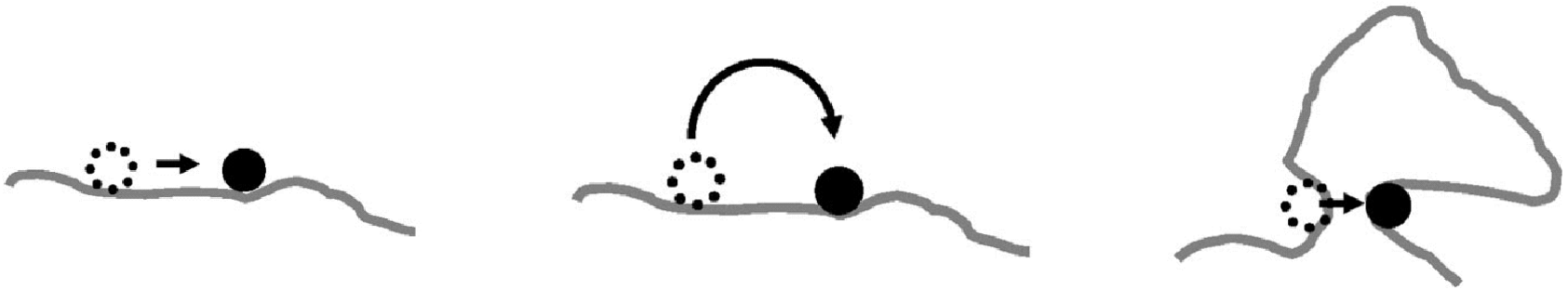
• Next step: backbone elasticity + DNA helicity +
PDB files analysis + computer simulations of protein diffusion

Macroscopic *qualitative* model of protein diffusion in DNA coil



- Every cycle: 3D diffusion in solution + 1D sliding along DNA
- [Protein] in solution $c_p=n_p/V$ and on DNA $c_{ads}=n_{ads}/V$
- Volume of DNA coil $\sim Lr^2$

Mechanisms of protein diffusion on DNA



Sliding/1D diffusion

Hopping

Inter-segmental transfer
(loop-facilitated process)

Actual diffusion is a combination of these basic steps

J. Marko et al., Nucl. Acids Res., 32 3040 (2004)

Time of target search: 3D + 1D

$$\tau_c = \int_0^{x+\lambda} \frac{\exp[\beta G(z)]}{D(z)} dz \int_0^z \exp[-\beta G(z')] dz'$$

van Kampen: Mean First Passage Time for 1 cycle

$$D(z) = \begin{cases} D_3, & 0 < z < x \\ D_1, & x < z < x + \lambda \end{cases}$$

Diffusion coefficient profile

$$y_{eff} = \frac{k_{on} n_p}{k_{off} n_{ads}} = y \frac{c_p}{c_{ads}} = \exp\left(\frac{E_{eff}}{k_B T}\right)$$

Non-equilibrium protein adsorption constant on DNA;
equilibrium: $y = k_{on}/k_{off}$

$$G(z) = \begin{cases} 0, & 0 < z < x \\ -E_{eff}, & x < z < x + \lambda. \end{cases}$$

Free energy profile: no DNA bp specificity

$$\tau_c = \frac{x^2}{2D_3} + \frac{\lambda^2}{2D_1} + \frac{x\lambda}{D_1 y_{eff}}$$

3D + 1D + correlation term (missing previously)
protein unbinding before travelling length λ on DNA

$$\tau \sim \left(\frac{L}{\lambda n_{ads}}\right)^{1/\alpha} \tau_c$$

Total search time along DNA of length L:
$\alpha = 1$: random protein attachment every step
$\alpha > 1$: super-diffusion

Total search time *vs.* Smoluchovski time

$$n_p L x^2 = L r^2 = V$$

L scales out

$$x = r / \sqrt{n_p}$$

Length of 3D path

$$k_{on} c_p = \frac{1}{\tau_{free}} = \frac{2D_3}{x^2}, \quad k_{off} c_{ads} = \frac{1}{\tau_{ads}} = \frac{2D_1}{\lambda^2}$$

Rates of protein binding and unbinding

$$d = \frac{D_1}{D_3} << 1$$

$$\lambda = \frac{r\sqrt{yd}}{\sqrt{n_{ads}}}$$

Optimal sliding length λ

$$\tau = \frac{Lr}{2D_3 n_p}\left(\frac{r}{\lambda}\frac{1}{n_{ads}} + \frac{\lambda}{r}\frac{n_p}{n_{ads}}\frac{1}{d} + \frac{2}{\sqrt{n_p}\, yd}\right)\left(\frac{L}{\lambda n_{ads}}\right)^{\frac{1}{\alpha}-1}$$

$$\tau_S = \frac{1}{2D_3 a c_p} = \frac{Lr^2}{2D_3 a n_p}$$

Smoluchovski 3D diffusion rate to a drain of radius a

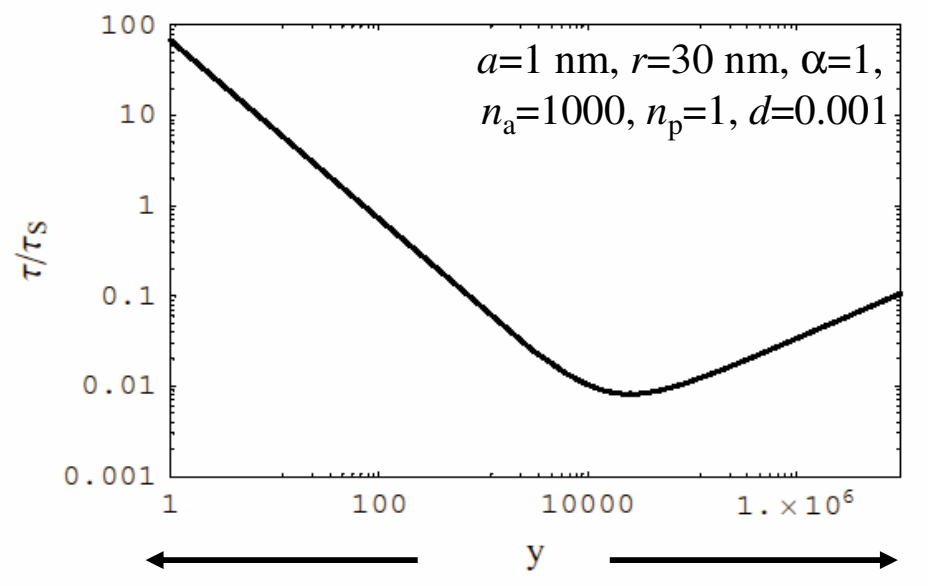
$$\frac{\tau}{\tau_S} = \frac{a}{r}\left(\frac{1}{\sqrt{n_{ads}\, yd}} + \frac{n_p\sqrt{y}}{n_{ads}^{3/2}\sqrt{d}} + \frac{2}{\sqrt{n_p}\, yd}\right)\left[\frac{L}{r\sqrt{n_{ads}\, yd}}\right]^{\frac{1}{\alpha}-1}$$

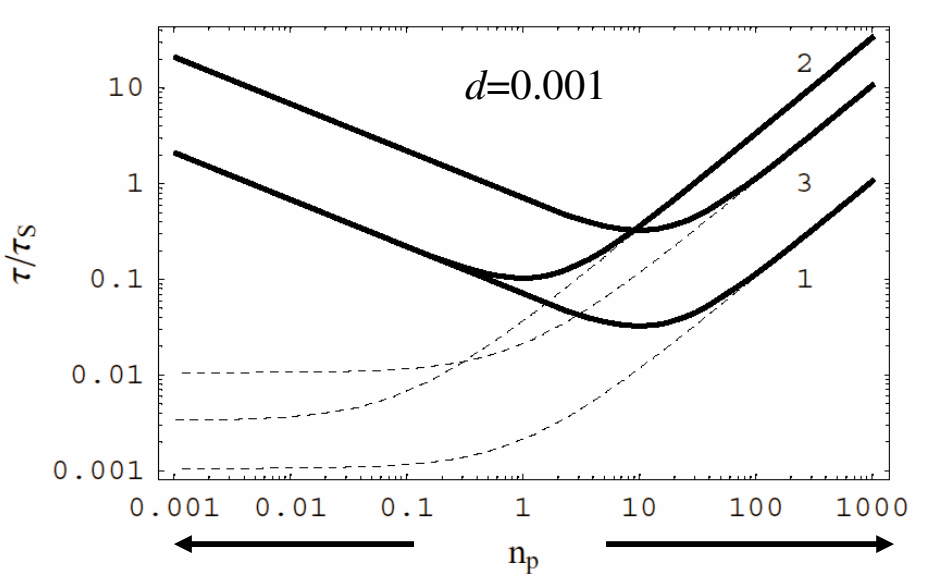
Final ratio of search times

$$\frac{\tau}{\tau_S} = \frac{a}{r}\frac{2}{y\sqrt{n_p}}\frac{\sqrt{d}+1}{d}$$

At equilibrium, y_{eff}=1, d<<1,
correlation term

Minimal search time at intermediate y and n_p values





Weak attraction to DNA

Strong attraction: long λ ineffective 1D search only

Unbinding drift is strong, λ is short

Always exist proteins close to the target in solution

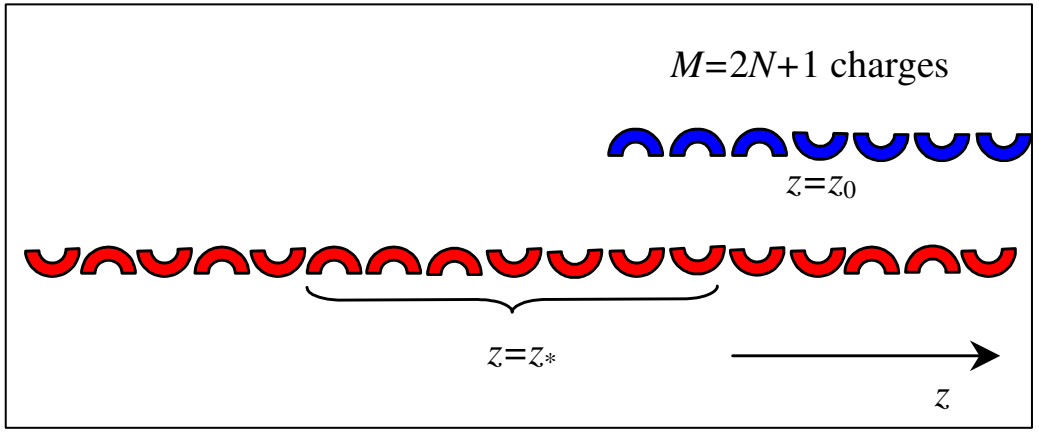
- Diffusion times faster than Smoluchovski

- As n_{ads} grows, τ decreases -- parallel search of DNA by many proteins

- Dotted curves: without correlation term -- wrong results

Part 2: Electrostatic key-lock mechanism of protein-DNA recognition

Electrostatic DNA-protein interaction and recognition energy



$$W_{el} = -\frac{e_0^2}{\varepsilon_c \pi}\int_{-\infty}^{\infty} dq\, K_0\left(\sqrt{q^2+\kappa^2}\,R\right) e^{iqz_0} \sum_{m=-N}^{N}\sum_{n=-\infty}^{\infty} e^{iqh(m-n)} e^{iq(\delta_m-\Delta_n)}$$

General electrostatic interaction energy

$$\left\langle W_{el}\right\rangle_{long-range} = -\frac{2e_0^2 M}{\varepsilon_c b}\left\{K_0(\kappa R) + 2\sum_{n=1}^{\infty} K_0\left(\sqrt{n^2g^2+\kappa^2}\,R\right) e^{-n^2g^2\Omega^2/2}\cos[ngz_0]\right\} -$$

$$\frac{2e_0^2}{\pi\varepsilon_c} M \int_0^{\infty} dq\, K_0\left(\sqrt{q^2+\kappa^2}\,R\right)\cos\left[q(z_*-z_0)\right]\left(1-e^{-q^2\Omega^2/2}\right)$$

Averaged

← Recognition energy

$$\frac{\left\langle \Delta W(\Delta z)\right\rangle_{long-range}}{k_B T} \approx -\frac{l_B M \Omega^2 \varepsilon}{2\varepsilon_c}\frac{R^2-2\Delta z^2}{\left(R^2+\Delta z^2\right)^{5/2}}$$

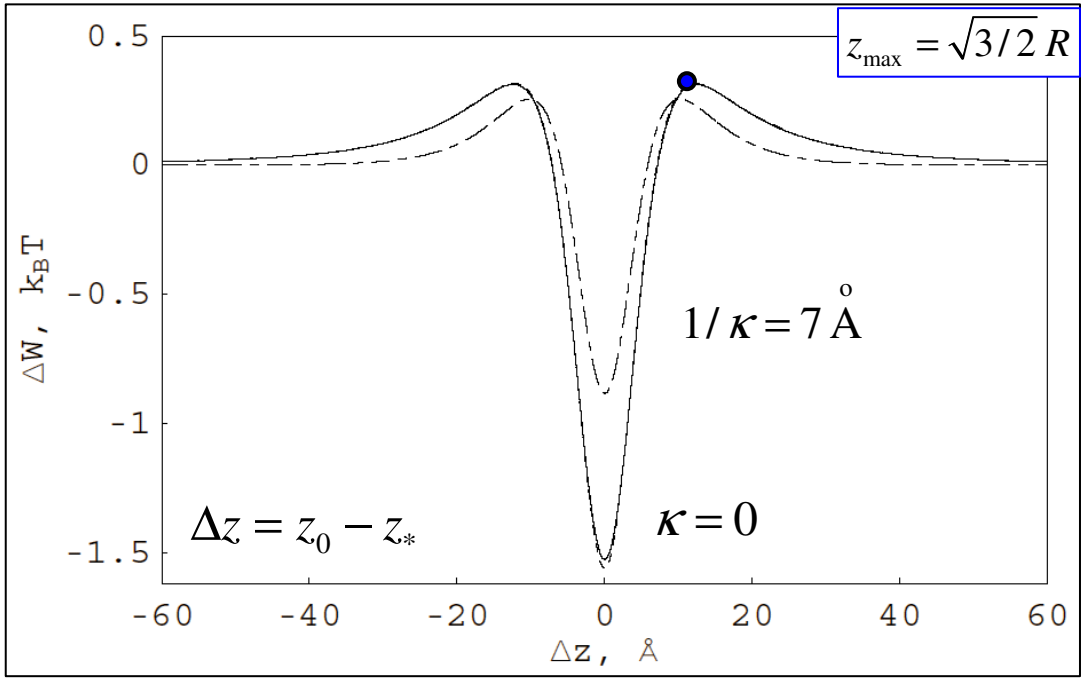
Simple form for $\Omega \ll 1$, $\kappa=0$

$g=2\pi/h$, $\langle\Delta_n^2\rangle=\Delta^2$, $\langle\delta_m^2\rangle=\delta^2$, $\Omega^2=\delta^2+\Delta^2$

$$\Delta z = z_0 - z_*$$

$$l_B = e_0^2/(\varepsilon k_B T)$$

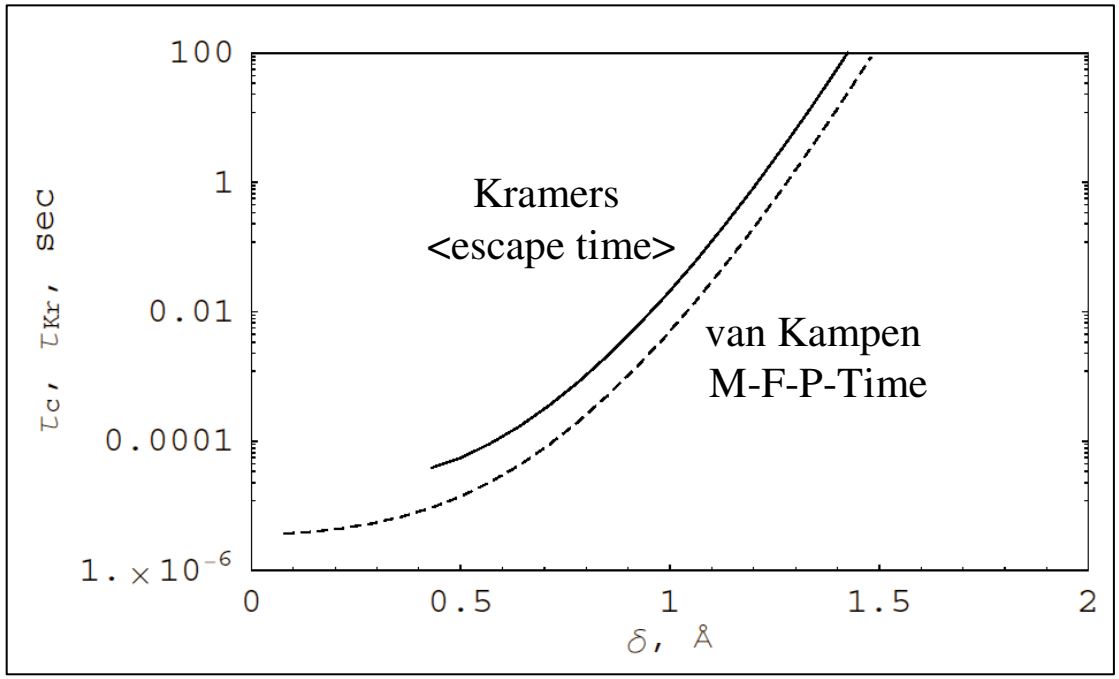
Electrostatic recognition energy ΔW

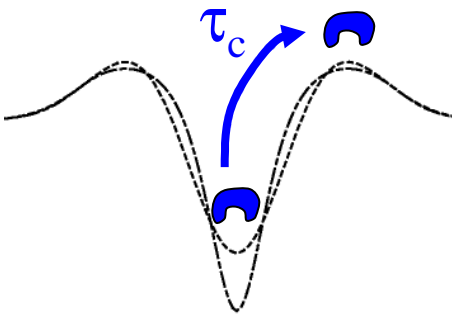


$M = 11,\ R = 10$ Å, $\varepsilon_c = 2,\ h = 3.4$ Å, $\delta^2 = \Delta^2,\ \Omega = 1$ Å

- Well is accompanied by the barriers
- Well depth is several $k_B T$
- Narrow wells: no “funnels” for protein diffusion
- Screening makes wells shallower
- Well depth d grows linearly with M
- d scales as $1/R^3$ at $\kappa \to 0$ and as $e^{-\kappa R}$ with salt

Protein residence time in the well



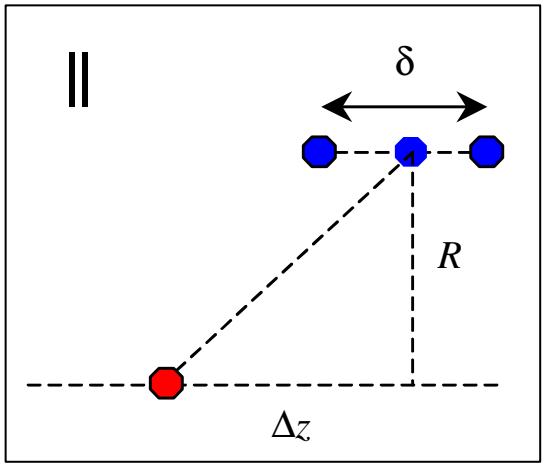


- Wells of ~k_BT in depth slow down protein diffusion
- Enough time to provoke protein conformational changes (μs- ms) and to induce stronger protein binding to DNA
- ES DNA-protein recognition is the first step of protein docking
- Stronger Hydrogen Bonding interactions can be formed afterwards

Thank you

cond-mat: 0708.0021

Funny energy barriers: Coulomb case

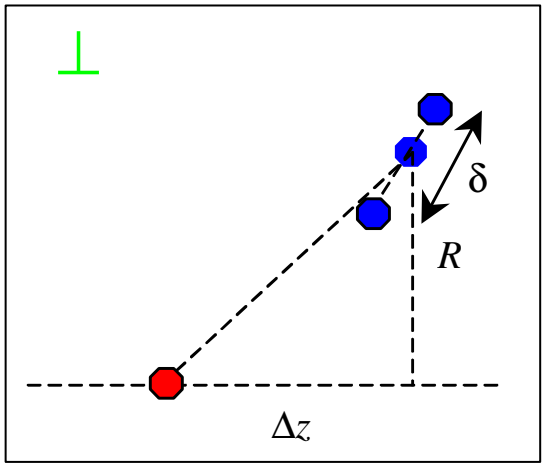


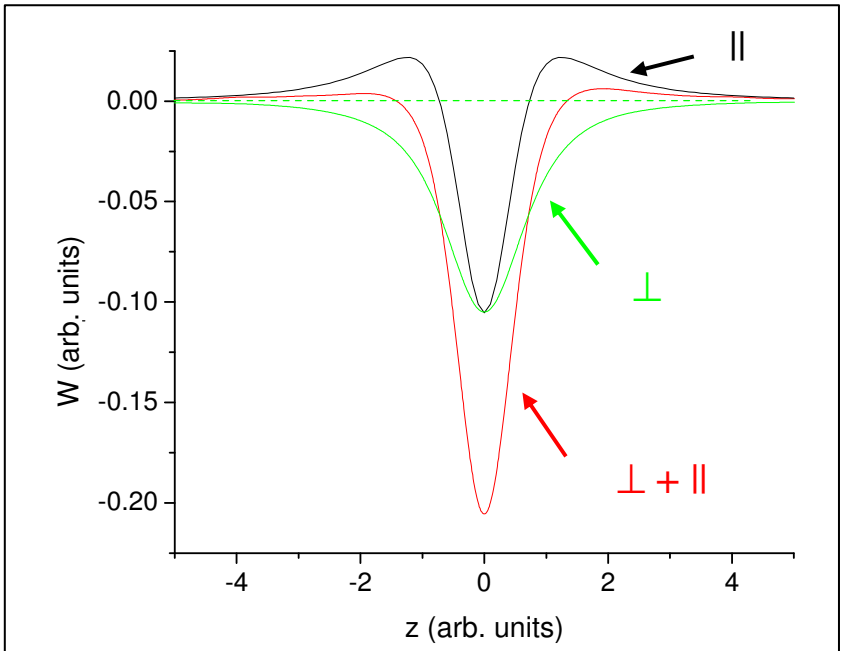
$$R(\delta) = R^2 + (\Delta z + \delta)^2$$

$$\Delta W_{el} = W_{el}(0) - W_{el}(\delta) = \frac{e_0^{\,2}}{\varepsilon_c \sqrt{R^2 + \Delta z^2}} - \frac{e_0^{\,2}}{\varepsilon_c \sqrt{R^2 + (\Delta z + \delta)^2}}$$

$$\langle \Delta W_{el} \rangle_\delta \overset{\text{Taylor Expansion}}{\approx} \cancel{\frac{e_0^{\,2} \Delta z \langle \delta \rangle}{\varepsilon_c \left(R^2 + \Delta z^2\right)^{3/2}}} + \frac{e_0^{\,2} \langle \delta^2 \rangle \left(R^2 - 2\Delta z^2\right)}{2\varepsilon_c \left(R^2 + \Delta z^2\right)^{5/2}}$$

At $R^2 > 2\Delta z^2$ fluctuations of charges always *reduce* their attraction energy

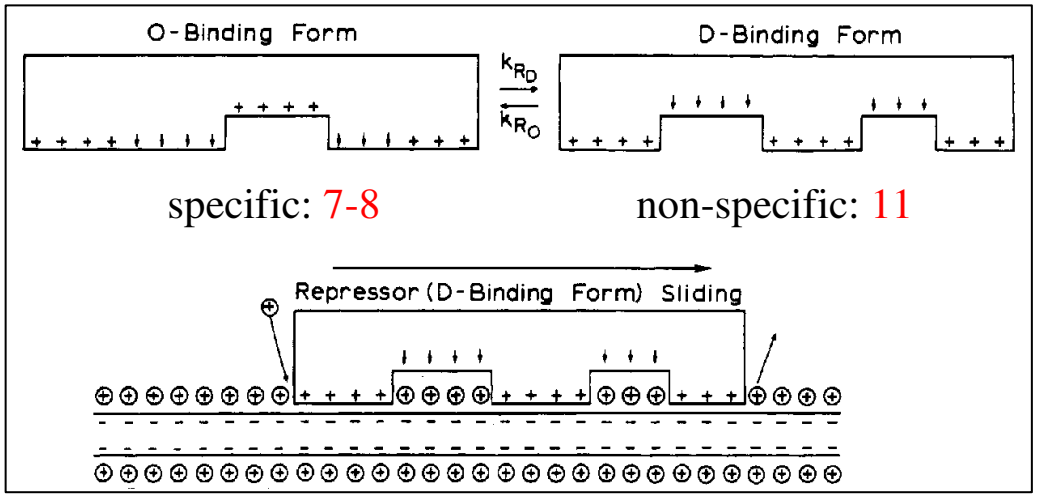




Computer Simulations

Electrostatic DNA-protein interactions: *lac* repressor

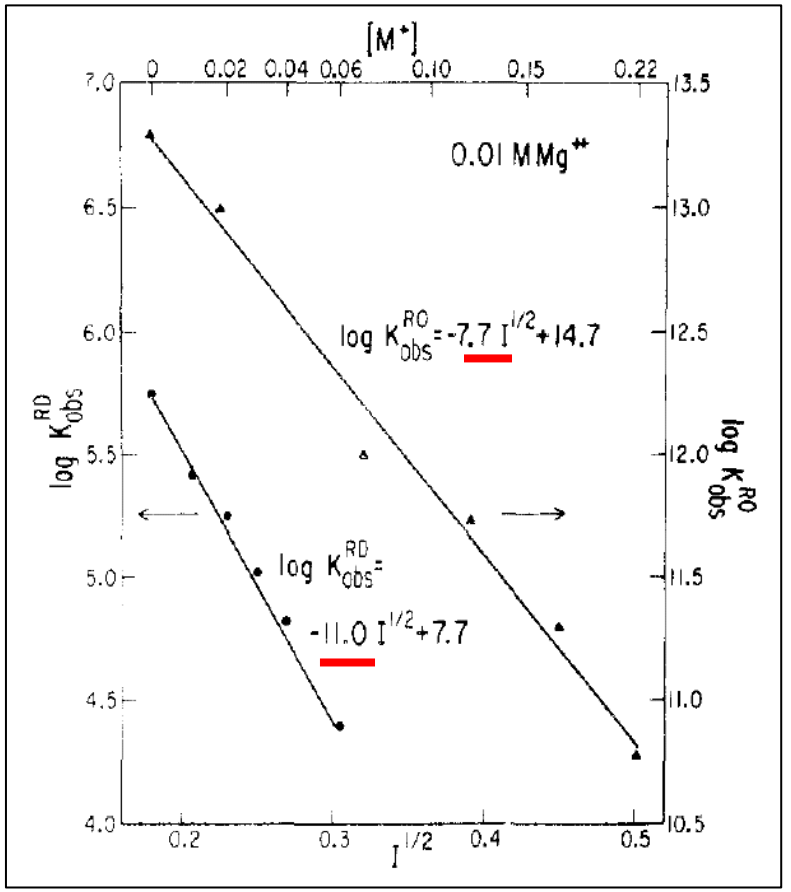
R B. Winter et al., Biochem., 20 6961 (1981)



Upon sliding, condensed cations are removed in front and they bind back on DNA behind the protein.

Electrostatic DNA-protein interactions are largely sequence non-specific?

M.T. Record et al., Biochem., 16 4791 (1977)

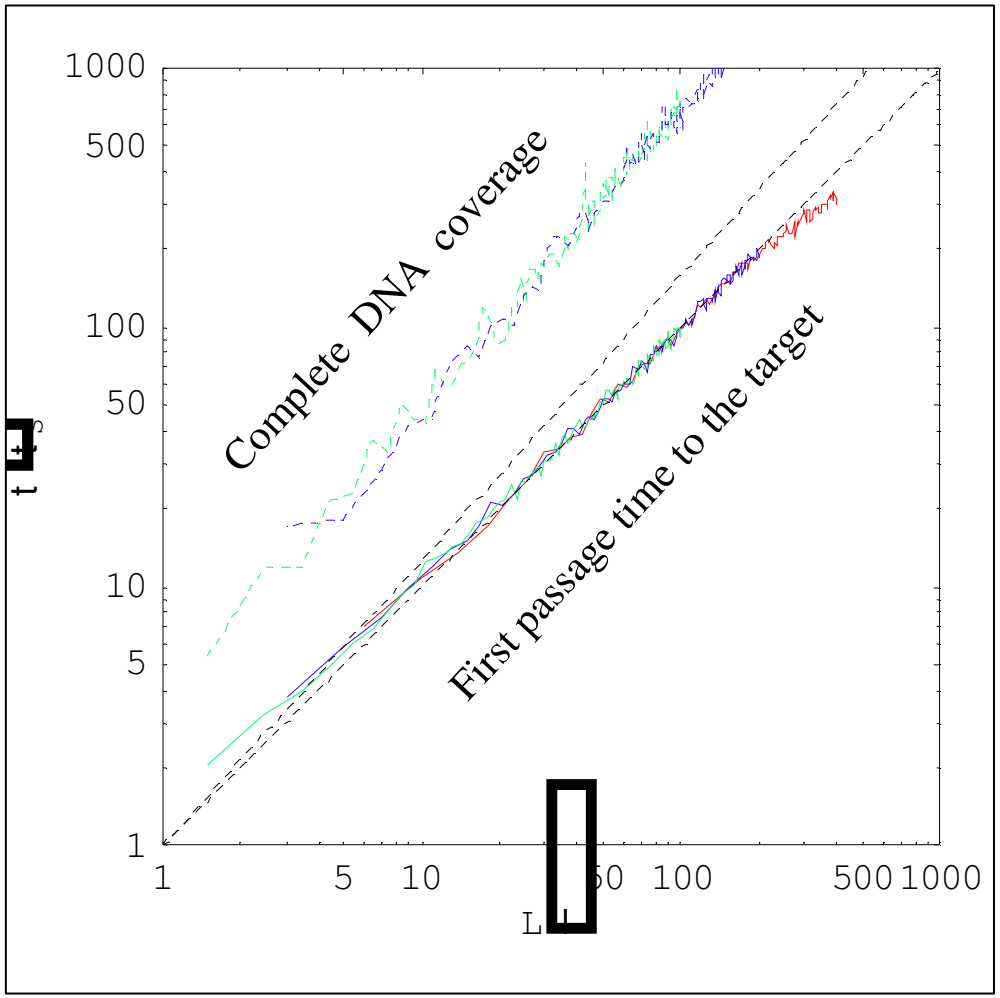


Enormous dependence of *lac* repressor association binding constant *K* on [salt]

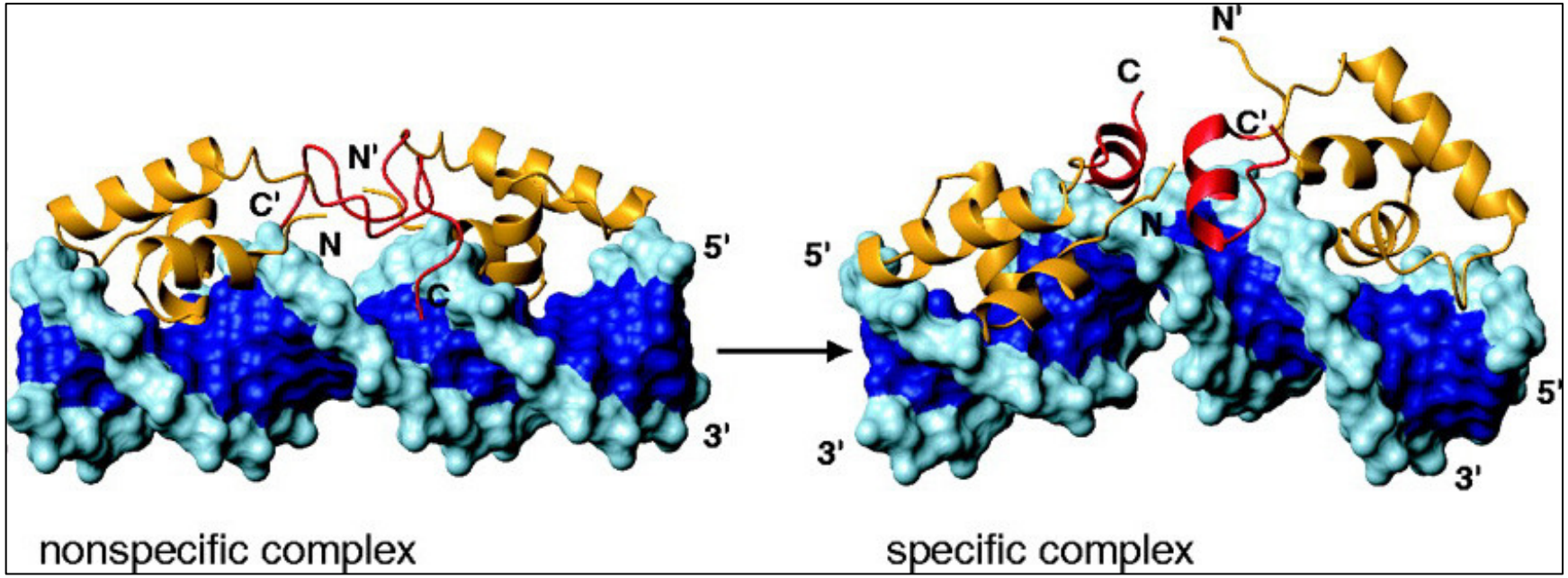
$$K = \frac{[\text{complex}]}{[\text{DNA}][\text{protein}]} \cdot \text{M}$$

Simple computer test

- single protein hopping randomly to left/right
- random target location
- random protein attachment point
- average over 5 runs
- L=20000, λ = 50, 100, 200

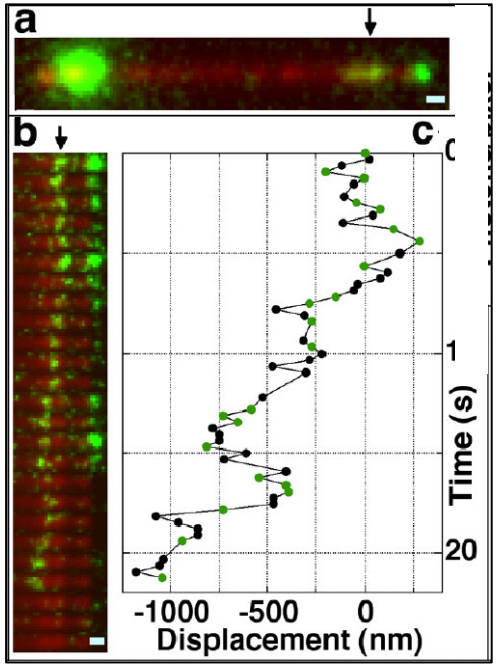


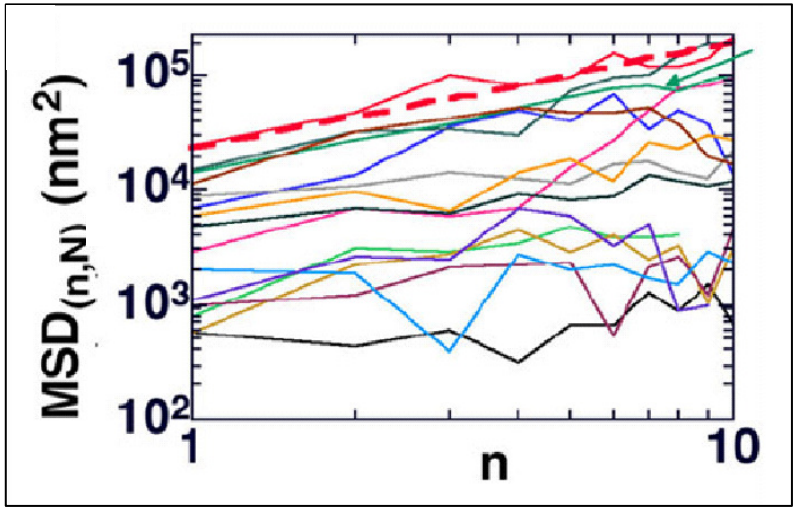
Interaction-induced folding and conformational adaptation

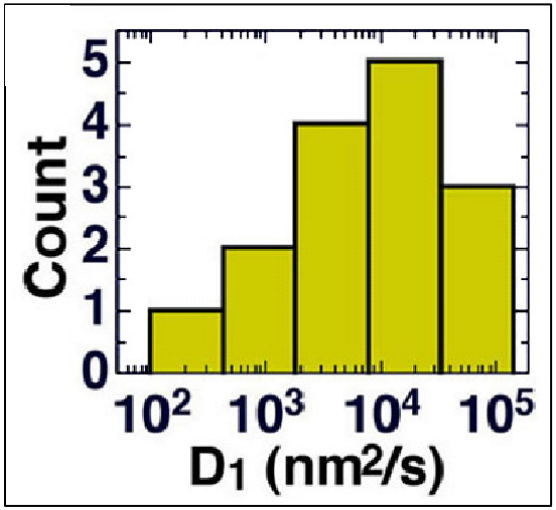


C. G. Kalodimos et al., Science, 305 386 (2004)

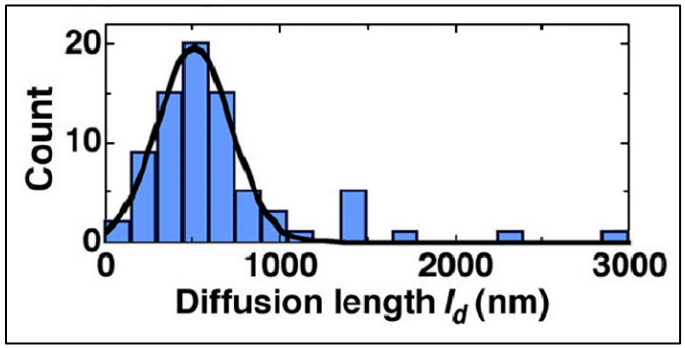
Lac repressor: $D_1 << D_3$







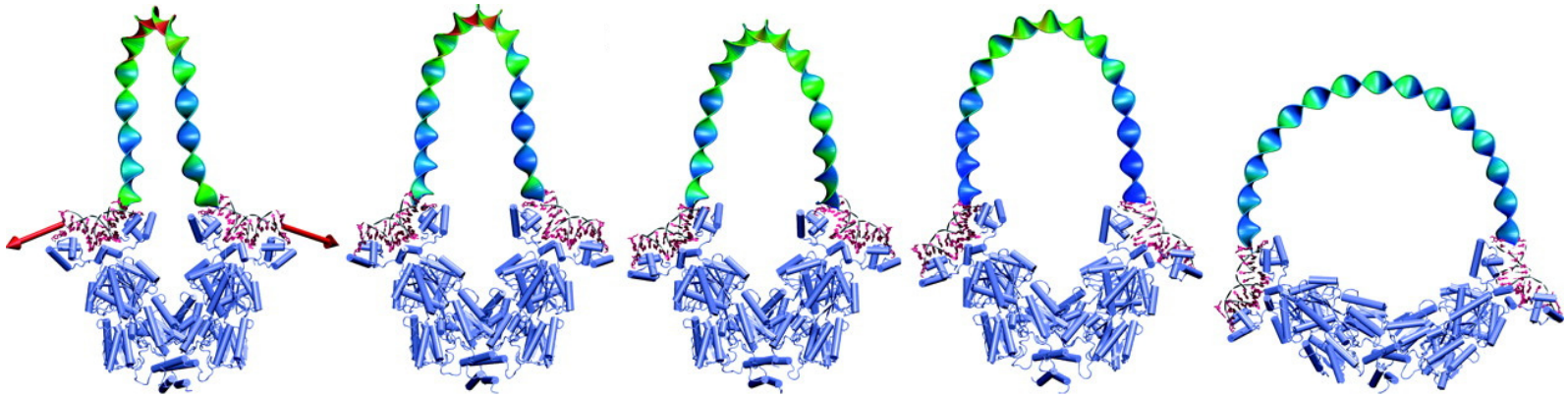
$$\mathrm{MSD}(n=1) \approx 2D_1\Delta t$$



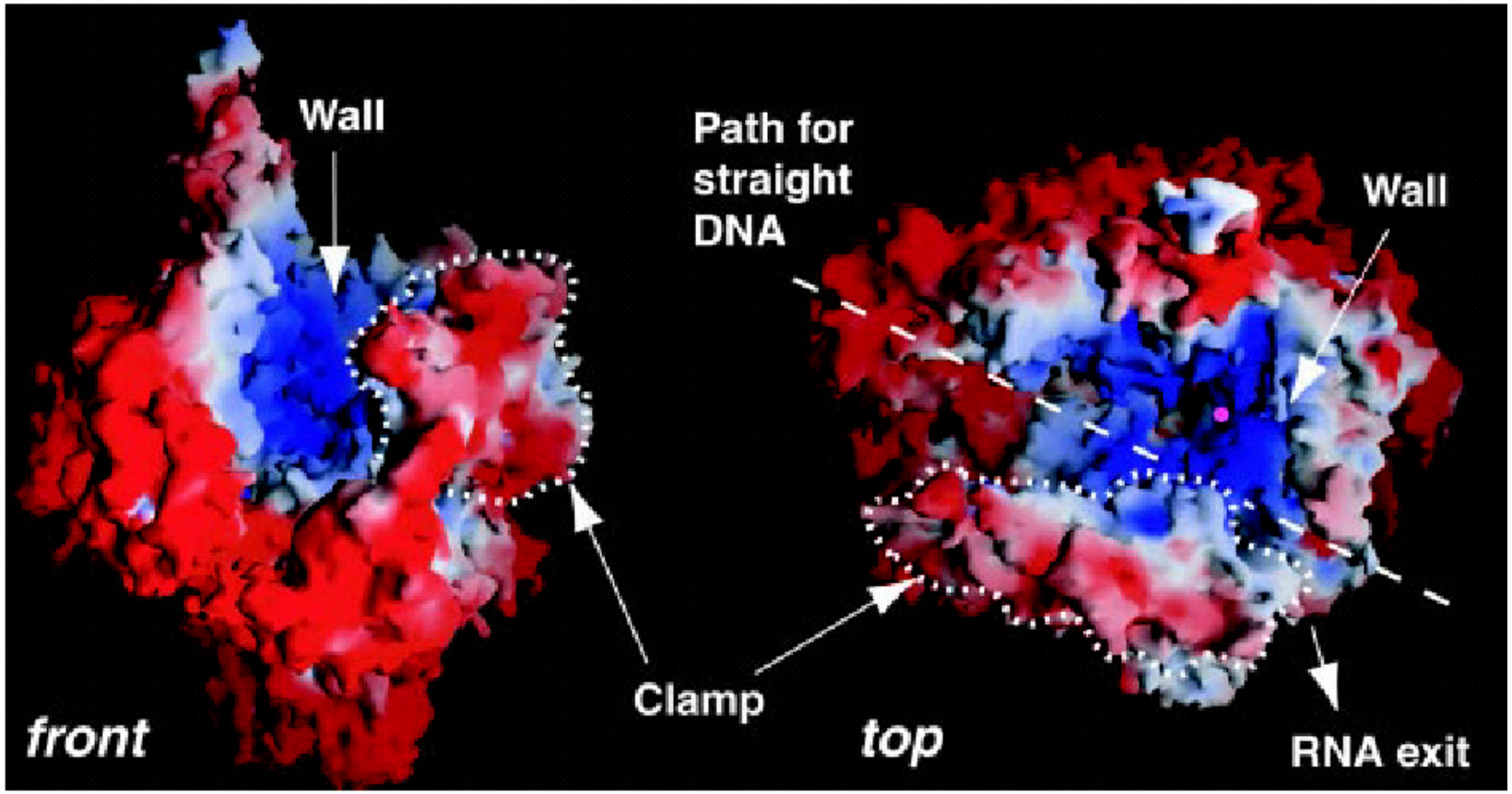
- Brownian Protein Motion with large D_1 variations
- Extract D_1 from Mean Square Displacements of proteins
- Experiment (Austin): D_1: $D_1=2\times10^{-10}$ cm^2/s
- Experiment D_3: $D_3=4\times10^{-7}$ cm^2/s

R. Austin et al., Phys. Rev. Lett., 97 048302 (2006)

DNA loops formed by *lac* repressor



Electrostatic potential of RNA Polymerase II



R. D. Kornberg et al., Science, 292 1863 (2001)

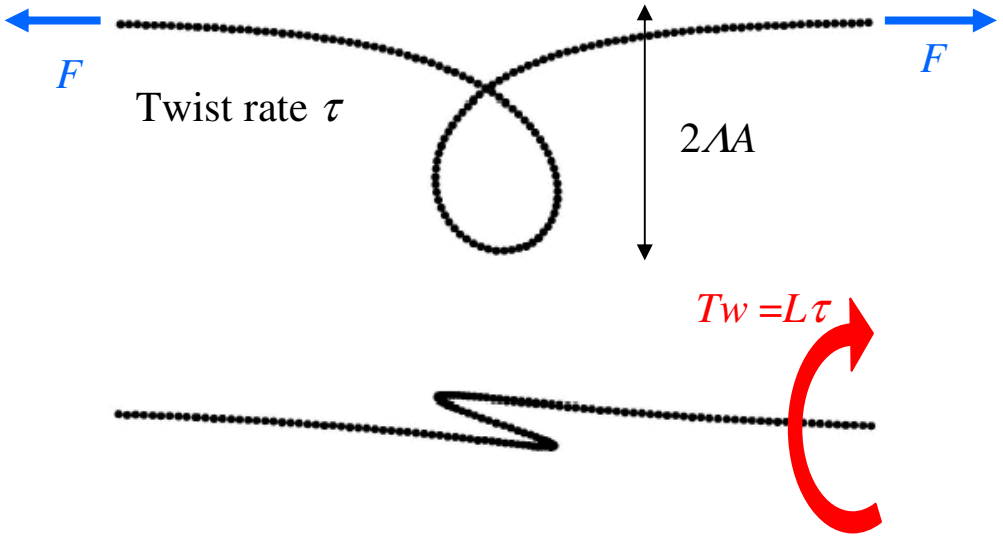
Looping uncharged elastic rods: buckling instability

Elasticity theory: 2D and 3D elastica, Euler and Kirchhoff -- local balance of forces and moments

Excess twist energy E_{tw} turns into loop bending energy E_b

Every loop removes about 2π of the excess twist Tw: $\tau = \tau_0 - 2\pi/L$

Looping of submarine cables [J. Coyne, IEEE J. Ocean. Ing., 15 72 (1990)]



$C=k_BTl_{tw}$ -- twist modulus, l_{tw}=750 Å
$B= k_BTl_p$ -- bend modulus, l_p=500 Å

$\Lambda=\sqrt{B/F}$, $A^2=1-C^2\tau^2/(4BF)$
$K^2(s)=4FA^2/\cosh[As]^2$ -- curve curvature

$E_b=4FA\Lambda$ -- loop bending energy

$F_0>C^2\tau^2/(4B)$ -- force to keep cable straight

$\Delta L=4A\Lambda$ -- cable slack upon looping

$$\vec{r}(s) = \left\{ 2A\Lambda \sin\left[\frac{s\sqrt{1-A^2}}{\Lambda}\right] / \cosh\left[\frac{As}{\Lambda}\right], -2A\Lambda\cos\left[\frac{s\sqrt{1-A^2}}{\Lambda}\right] / \cosh\left[\frac{As}{\Lambda}\right], s - 2A\Lambda\tanh\left[\frac{As}{\Lambda}\right] \right\}$$

Looping charged rods: limitations of OSF theory

$$E_{el}(r) = \frac{e^2}{\varepsilon r} e^{-\kappa r}$$

Screened interactions of charges

$$\kappa = \sqrt{8\pi l_B n_0}$$

$1/\kappa \approx 10$Å in physiological solution

Optimal loop shape in 3D is a complicated problem: non-locality, self-contacts.

E_{el} of loops with Debye-Hückel interactions:
OSF electrostatic rod stiffening works only
for large loops $R \gg 1/\kappa$ with no close contacts

$$l_p \rightarrow l_{p,el} = l_p + l_B/(4\kappa^2 h^2)$$

h is charge-charge separation

Applicability of OSF to tight DNA loops

Numerical summation of the Debye-Hückel potentials along the loop contour

$$\Delta E_{el} = E_{el}^{\text{looped}} - E_{el}^{\text{straight}}$$