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

A. G. Cherstvy\textsuperscript{1}, A. B. Kolomeisky\textsuperscript{2}, and A. A. Kornyshev\textsuperscript{3}

\textsuperscript{1} Institut für Festkörperforschung, Theorie-II, Forschungszentrum Jülich, D-52425 Jülich, Germany

\textsuperscript{2} Department of Chemistry, Rice University, Houston, Texas 77005, USA

\textsuperscript{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


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

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

Enormous dependence of *lac* repressor association binding constant \( K_{\text{on}} \) 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 $k_B T$

- 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.

Electrostatic potential of *lac* repressor

- 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?

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

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

NCP stability ([salt])
B-DNA charge and structure non-ideality

$-1 \epsilon_0$ per each 1.7 Å along DNA axis

**The 10 Twist Angles of B-DNA**

<table>
<thead>
<tr>
<th>Dinucleotide</th>
<th>Twist Angle (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(AA) · (TT)</td>
<td>35.6 ± 0.1</td>
</tr>
<tr>
<td>(AC) · (GT)</td>
<td>34.4 ± 1.3</td>
</tr>
<tr>
<td>(AG) · (CT)</td>
<td>27.7 ± 1.5</td>
</tr>
<tr>
<td>(AT) · (AT)</td>
<td>31.5 ± 1.1</td>
</tr>
<tr>
<td>(CA) · (TG)</td>
<td>34.5 ± 0.9</td>
</tr>
<tr>
<td>(CC) · (GG)</td>
<td>33.7 ± 0.1</td>
</tr>
<tr>
<td>(CG) · (CG)</td>
<td>29.8 ± 1.1</td>
</tr>
<tr>
<td>(GA) · (TC)</td>
<td>36.9 ± 0.9</td>
</tr>
<tr>
<td>(GC) · (GC)</td>
<td>40.0 ± 1.2</td>
</tr>
<tr>
<td>(TA) · (TA)</td>
<td>36.0 ± 1.0</td>
</tr>
</tbody>
</table>

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, $R \approx |b| \approx 7\text{Å}$
- $s_{1,2}$ 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

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

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

DNA: PO₄⁻ groups
Nucleosomes: DNA-wrapping proteins of eukaryotes

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

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

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

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

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.


Thank you
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**

Spiraling RNA Polymerase:
protein binding requires DNA
groove tracking

K. Sakata-Sogawa et al.,

Sliding vs. Spiraling
electrostatic barriers vs. hydrodynamic friction

A.G. Cherstvy and R.G. Winkler,

- 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

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 $\Delta_n = \delta_m$ -- stronger DNA-protein attraction

- Potential well near the homology region
Artificial charge periodicity in protein DNA-binding domains

• Periodicity of \( \approx 7 \text{ Å} \) and \( \approx 34 \text{ Å} \) 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

van Kampen: Mean First Passage Time for 1 cycle

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

Free energy profile: no DNA bp specificity

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

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

\[ n_p Lx^2 = Lr^2 = V \]

\[ L \text{ scales out} \]

\[ x = r / \sqrt{n_p} \]

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

\[ d = \frac{D_1}{D_3} << 1 \]

Optimal sliding length \( \lambda \)

\[ \tau = \frac{Lr}{2D_3n_p} \left( \frac{r}{\lambda n_{ads}} + \frac{\lambda}{r n_{ads}} \frac{1}{d} + \frac{2}{\sqrt{n_p yd}} \right) \left( \frac{L}{\lambda n_{ads}} \right)^{\frac{1}{\alpha} - 1} \]

Smoluchovski 3D diffusion rate to a drain of radius \( a \)

\[ \tau_s = \frac{1}{2D_3ac_p} = \frac{Lr^2}{2D_3an_p} \]

Final ratio of search times

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

\[ \frac{\tau}{\tau_s} = \frac{a}{r} \frac{2}{y \sqrt{n_p d}} \sqrt{d + 1} \]

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 $\lambda$ ineffective 1D search only
- Unbinding drift is strong, $\lambda$ is short
- Always exist proteins close to the target in solution

- Diffusion times faster than Smoluchovski
- As $n_{ads}$ grows, $\tau$ 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

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

Averaged

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

Recognition energy

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

Simple form for \( \Omega << 1 \), \( \kappa = 0 \)

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

\[ \Delta z = z_0 - z_\ast \]

\[ l_B = e_0^2 / (\varepsilon k_B T) \]
Electrostatic recognition energy $\Delta W$

- 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 $\kappa=0$ and as $e^{-\kappa R}$ with salt

$M = 11$, $R = 10\text{ Å}$, $\epsilon_c = 2$, $h = 3.4\text{ Å}$, $\delta^2 = \Delta^2$, $\Omega = 1\text{ Å}$
Protein residence time in the well

- Wells of $\sim k_B T$ 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
Funny energy barriers: Coulomb case

\[ R(\delta) = R^2 + (\Delta z + \delta)^2 \]

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

\[ \langle \Delta W_{el} \rangle_\delta \overset{Taylor \ Expansion}{\approx} \frac{e_0^2 \Delta z \langle \delta \rangle}{\varepsilon_c (R^2 + \Delta z^2)^{3/2}} + \frac{e_0^2 \langle \delta^2 \rangle (R^2 - 2\Delta z^2)}{2\varepsilon_c (R^2 + \Delta z^2)^{5/2}} \]

At \( R^2 > 2\Delta z^2 \) fluctuations of charges always reduce their attraction energy.

Computer Simulations
Electrostatic DNA-protein interactions: *lac* repressor


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

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**?

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

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

- single protein hopping randomly to left/right
- random target location
- random protein attachment point
- average over 5 runs
- $L=20000$, $\lambda = 50, 100, 200$
Interaction-induced folding and conformational adaptation

Lac repressor: $D_1 << D_3$

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

DNA loops formed by \textit{lac} repressor
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 $2\pi$ of the excess twist $Tw$: $\tau = \tau_0 - 2\pi/L$


$C=k_BT_{lw}$ -- twist modulus, $l_{tw}=750$ Å
$B=k_BT_{lp}$ -- bend modulus, $l_p=500$ Å

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

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

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

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

$\vec{r}(s) = \begin{cases} 
2A\Lambda \sin \left[ \frac{s\sqrt{1-A^2}}{\Lambda} \right] / \cosh \left[ \frac{As}{\Lambda} \right], & -2A\Lambda \cos \left[ \frac{s\sqrt{1-A^2}}{\Lambda} \right] / \cosh \left[ \frac{As}{\Lambda} \right], s - 2A\Lambda \tanh \left[ \frac{As}{\Lambda} \right] \end{cases}$
Looping charged rods: limitations of OSF theory

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

\[ \kappa = \sqrt{8\pi l_B n_0} \]

Screened interactions of charges

\[ 1/\kappa \approx 10\text{Å} \text{ in physiological solution} \]

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

\[ E_{el} \text{ of loops with Debye-Hückel interactions:} \]

OSF electrostatic rod stiffening works only for large loops \( R >> 1/\kappa \) 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

\[ \Delta E_{el} = E_{el \text{ looped}} - E_{el \text{ straight}} \]