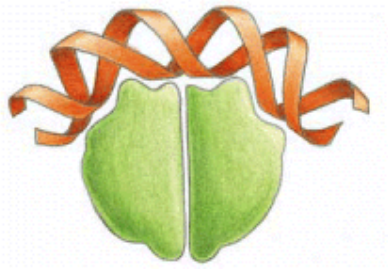
Physics 176/276
Quantitative Molecular Biology

Lecture XI: Protein-DNA Interaction

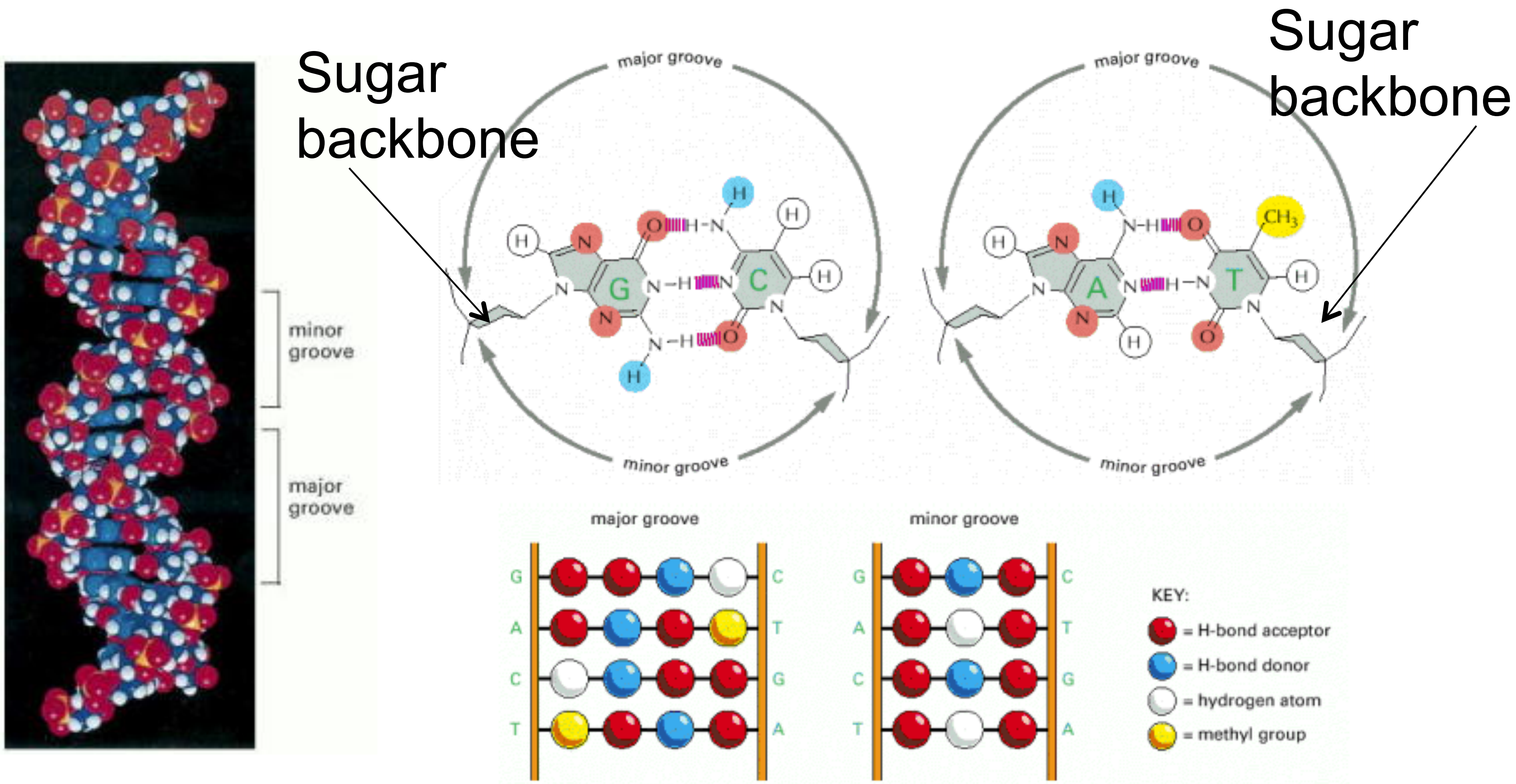
http://physics.ucsd.edu/students/courses/winter2014/physics176

A. Empirical facts

1. Transcription Factors
   - size: ~5nm (10-20 bp)



   - molecular basis of sequence recognition



Sugar backbone

Sugar backbone

KEY:

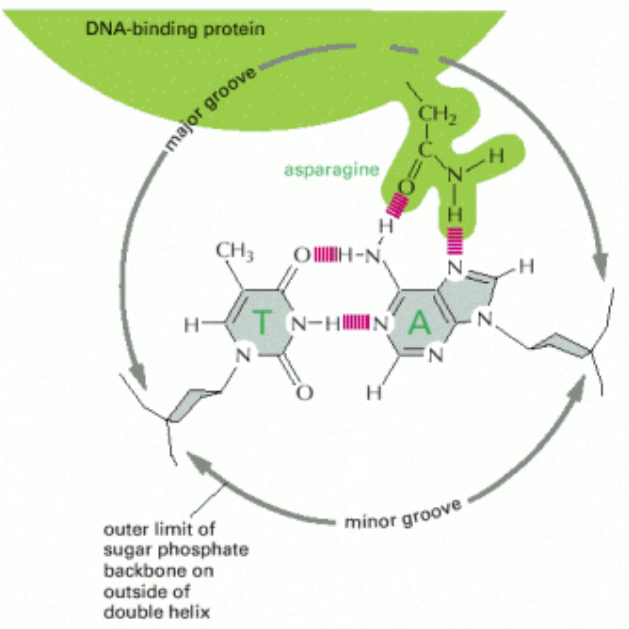
🟥 = H-bond acceptor

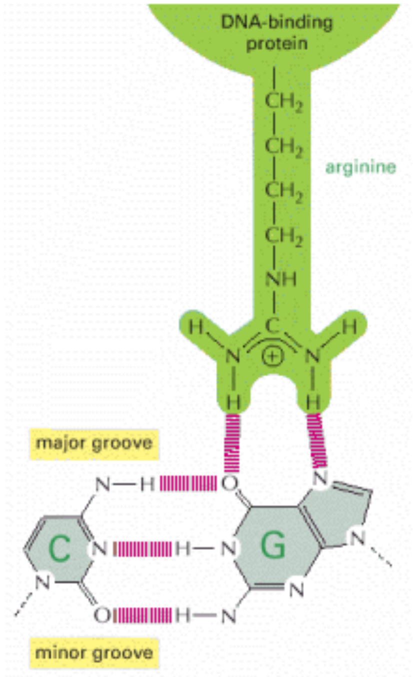
🟦 = H-bond donor

○ = hydrogen atom

🟨 = methyl group

- contact between TF and DNA

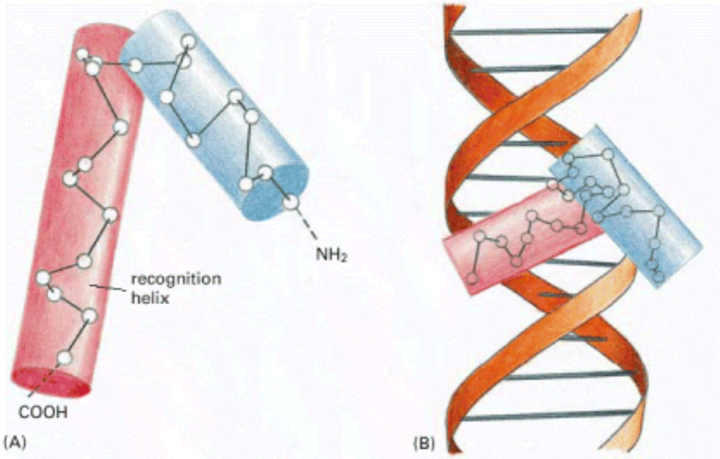

DNA-binding protein
major groove
asparagine
CH2
C
O
N
H
CH3
T
A
minor groove
outer limit of sugar phosphate backbone on outside of double helix

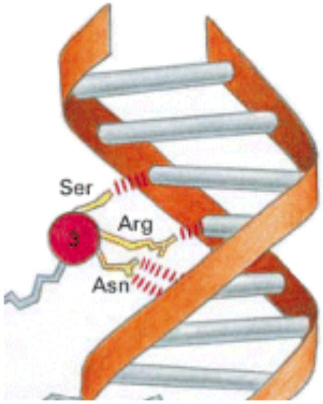


➔ structure of a TF must place the appropriate amino acids next to the base pairs they contact

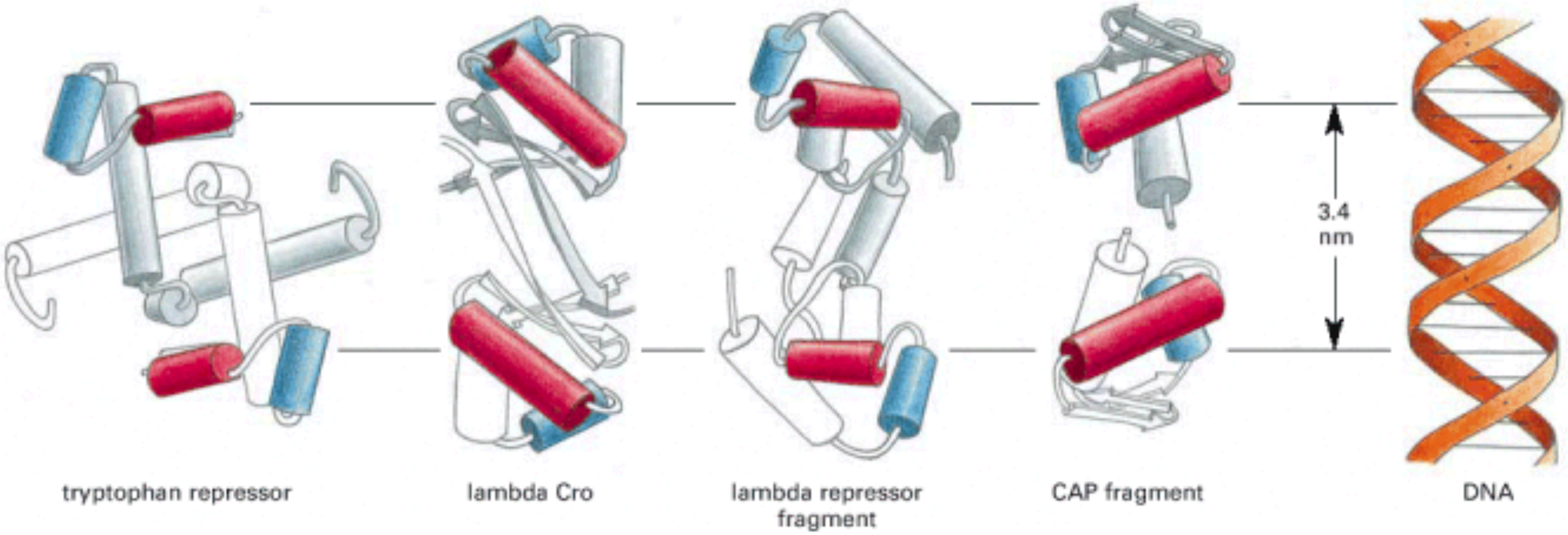
➔ Hydrogen bonds with the backbone also play a crucial role

- various molecular structure solutions
  - Helix-Turn-Helix

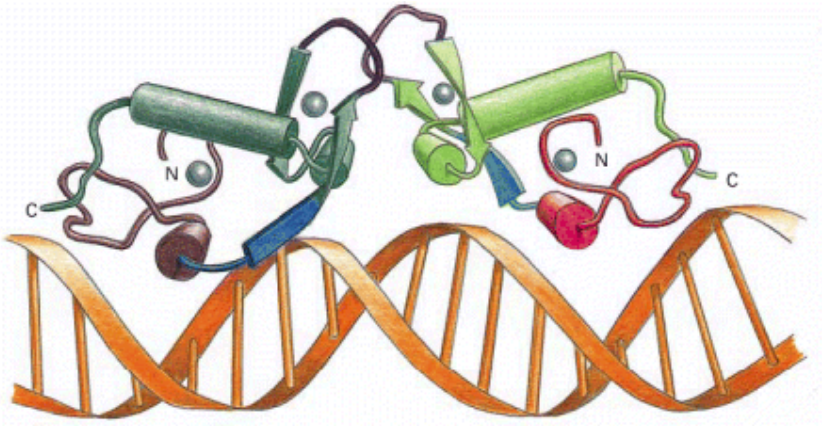




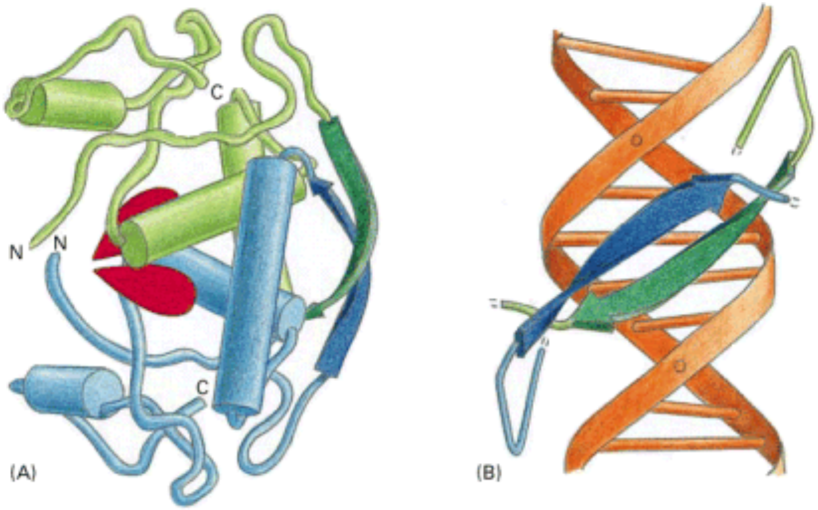
well-known examples in bacteria (note: homodimers)



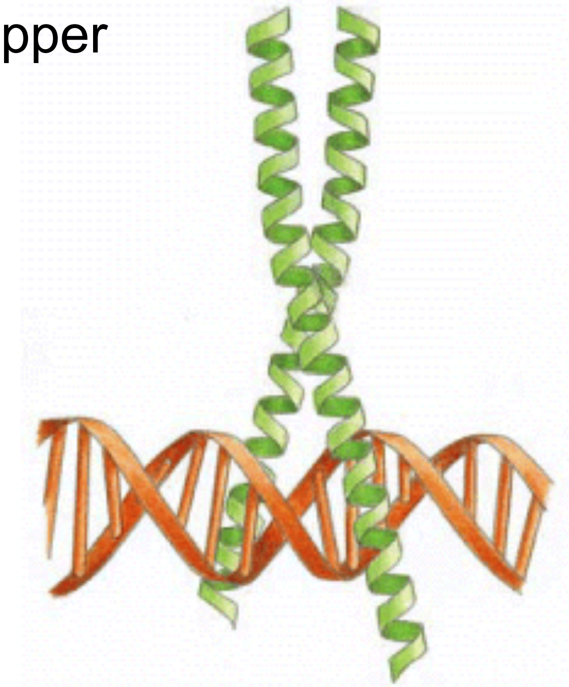
– zinc-finger domain



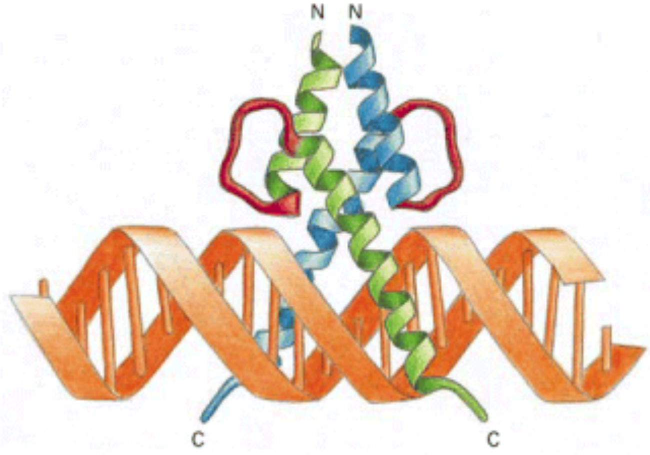
– beta-sheets



– leucine zipper



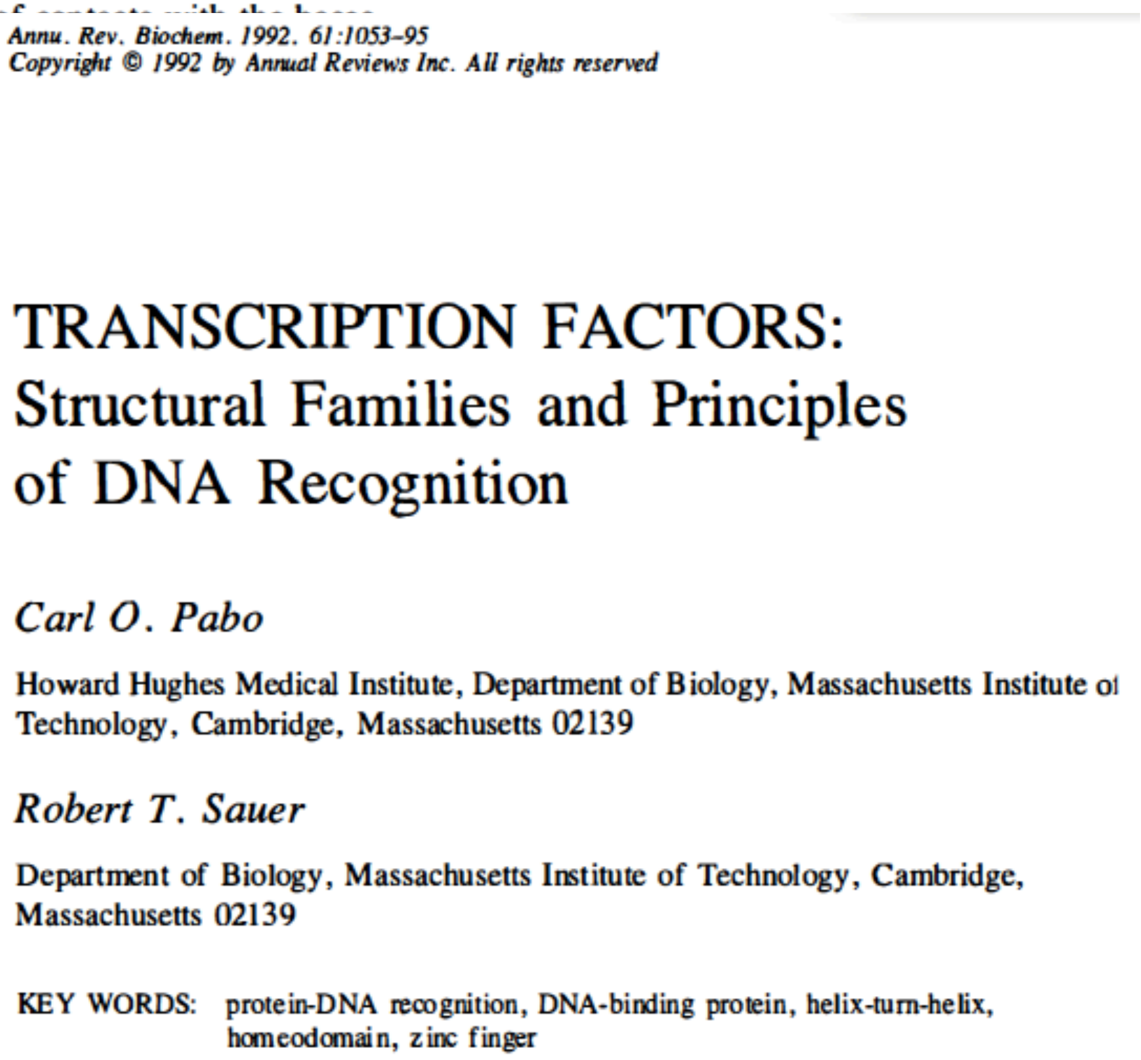
– helix-loop-helix



General Principles of Site-Specific Recognition

Although the diversity of knov
that there are no simple rules c
comparing the known complex
tions.

1. Site-specific recognition always involves a set c
   and with the DNA backbone.
2. Hydrogen bonding is critical for recognition (
   teractions also occur). A complex typically has
   hydrogen bonds at the protein/DNA interface.
3. Side chains are critical for site-specific recogniti
   which the peptide backbone makes hydrogen bo
   backbone, but side chains make most of the c
4. There is no simple one-to-one correspondence b
   bases they contact. It appears that the folding
   protein help to control the "meaning" that any p
   site-specific recognition.
5. Most of the base contacts are in the major groo
   (which are larger and offer more hydrogen-b
   groove) seem to be especially important.
6. Most of the major motifs contain an α-helical reg
   groove of B-form DNA. There are examples of
   regions of polypeptide chain that play critical ro
   base contacts from these regions appear to be
7. Contacts with the DNA backbone usually invol
   salt bridges with the phosphodiester oxygens.
8. Multiple DNA-binding domains usually are
   recognition. The same motif may be used more
   the active binding species is a homodimer or het
   polypeptide contains tandem recognition motifs
   tended arm and a HTH unit; a homeodomain and POU-specific domain,
   etc) may also be used in the same complex.
9. Recognition is a detailed structural process. Hydration can play a critical role in recognition; sequence-dependent aspects of the DNA structure may also be important.



Annu. Rev. Biochem. 1992. 61:1053–95

TRANSCRIPTION FACTORS: Structural Families and Principles of DNA Recognition

Carl O. Pabo

Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Robert T. Sauer

Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

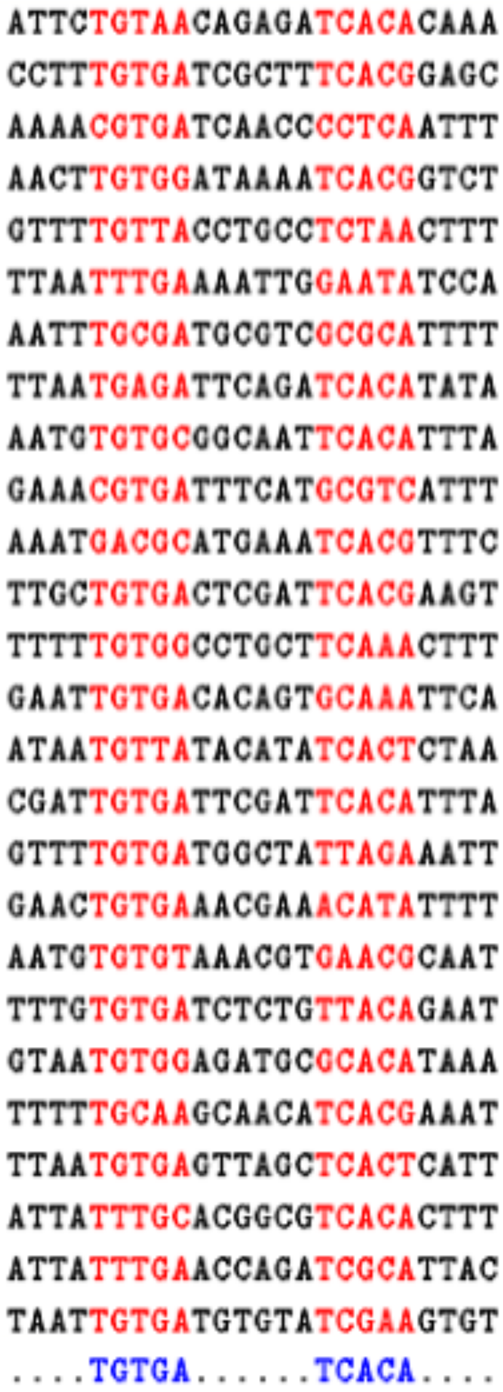
KEY WORDS: protein-DNA recognition, DNA-binding protein, helix-turn-helix, homeodomain, zinc finger

2. DNA binding sequences

- typically 10-20 bp in bacteria

protein	target sequence
lac repressor	5’ AATTGTGAGCGGATAACAATT 3’ TTAACACTCGCCTATTGTTAA
CRP	TGTGAGTTAGCTCACA ACACTCAATCGAGTGT
λ repressor	TATCACCGCCAGAGGTA ATAGTGGCGGTCTCCAT

- lots of sequence variants
- consensus sequence often <u>palindromic</u>
- common to have 2~3 mismatches from the core consensus sequence
-- “fuzzy” binding motif



```
ATTCTGTAACAGAGATCACACAAA
CCTTTGTGATCGCTTTCACGGAGC
AAAACGTGATCAACCCCTCAATTT
AACTTGTGGATAAAATCACGGTCT
GTTTTGTTACCTGCCTCTAACTTT
TTAATTTGAAAATTGGAATATCCA
AATTTGCGATGCGTCGCGCATTTT
TTAATGAGATTCAGATCACATATA
AATGTGTGCGGCAATTCACATTTA
GAAACGTGATTTCATGCGTCATTT
AAATGACGCATGAAATCACGTTTC
TTGCTGTGACTCGATTCACGAAGT
TTTTTGTGGCCTGCTTCAAACTTT
GAATTGTGACACAGTGCAAATTCA
ATAATGTTATACATATCACTCTAA
CGATTGTGATTCGATTCACATTTA
GTTTTGTGATGGCTATTAGAAATT
GAACTGTGAAACGAAACATATTTT
AATGTGTGTAAACGTGAACGCAAT
TTTGTGTGATCTCTGTTACAGAAT
GTAATGTGGAGATGCGCACATAAA
TTTTTGCAAGCAACATCACGAAAT
TTAATGTGAGTTAGCTCACTCATT
ATTATTTGCACGGCGTCACACTTT
ATTATTTGAACCAGATCGCATTAC
TAATTGTGATGTGTATCGAAGTGT
....TGTGA......TCACA....
```

3. TF-DNA interaction

- passive (no energy consumption)
- strong electrostatic attraction independent of binding seq

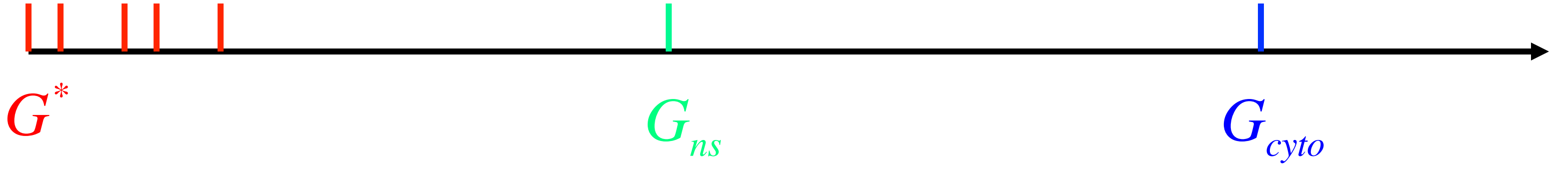
  e.g., $[TF-DNA] > 10 \times [TF]_{free}$ for LacI in 0.1M salt

  ➔ non-specific binding: $G_{ns} - G_{cyto} \simeq -15kT$

  (kT ≈ 0.62 kcal/mole at 37°C; ≈ 2.5 kJ/mole)

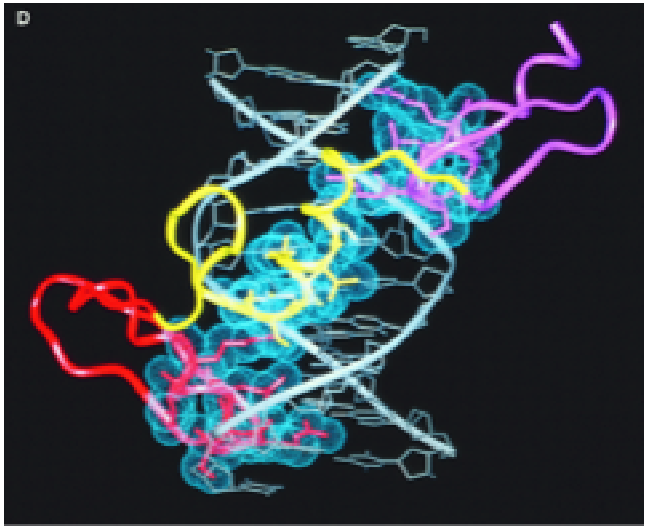
- additional energy gained from hydrogen bonds to preferred sequences

  strongest binder: $G^{*} - G_{ns} \simeq -15kT$



- graded increase in binding energy for sequences with partial match to the preferred sequence

- relative binding affinity for Mnt (repressor of phage P22)



binding energy matrix

(in unit of kT ≈ 0.6 kcal/mole)

pos.	10	11	12	13	14	15	16	17
A	1.8	2.4	1.6	1.0	0	2.1	0.8	1.1
C	2.4	1.9	4.2	2.1	0.3	0	0	0
G	0	1.6	0	0	1.2	3.2	1.0	1.2
T	3.0	0	2.2	2.2	0.6	2.2	0.7	0.3

(D.S. Fields, Y. He, A. Al-Uzri & G. Stormo, 1997)

(from competitive binding expts)

➔ weak energetic preference -- weak specificity

➔ similar results for other TFs studied (e.g., LacI, λ-CI, λ-Cro)

- double mutation: binding energy approx additive

➔ Can we say something generic about
the design of TF-DNA interaction from these facts/data?

- Issues addressed here:
  - range of TF-DNA affinity *in vivo*
  - dependence of this affinity on variation in target sequence
  - why weak specificity of TF-DNA interaction?
    [“design rule” for TF]
  - why fuzzy motifs
    [choice of DNA targets]
- Issues not addressed:
  - what is the target sequence of a given TF
    [can be probed experimentally]
  - fluctuations in TF-DNA binding

B. Thermodynamics of DNA target recognition

- binding sequence (L nt):

$$S = \{b_1, b_2, ..., b_L\}, \quad b_i \in \{A, C, G, T\}$$

- TF: N_P/cell

$$[P]_{tot} = N_P / V_{cell}$$

- dissociation constant (*in vitro*)

$$K(S) \equiv [P]\cdot[S]/[P\cdot S]$$

$$\propto e^{G(S)/kT}$$

- fraction of sequence bound:

$$f(S) \equiv \frac{[P\cdot S]}{[S]+[P\cdot S]} = \frac{[P]}{[P]+K(S)}$$

$$\approx \frac{[P]_{tot}}{[P]_{tot}+K(S)} \qquad \text{if } [S]_{tot} \ll [P]_{tot}$$

- approx. additive binding free energy

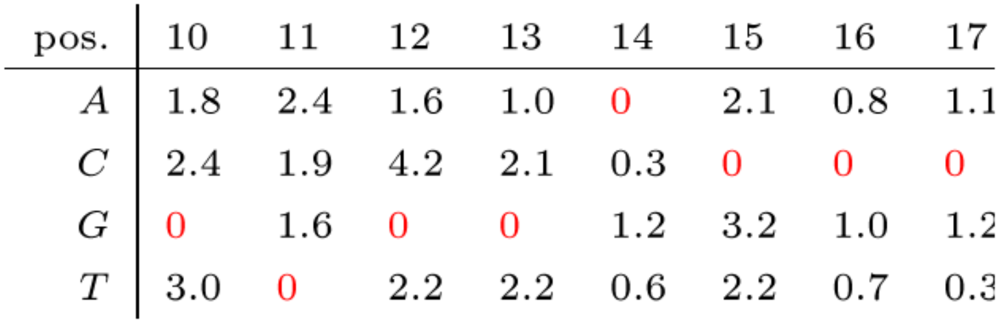
$$G(S) \approx G^* + \sum_{i=1}^{L} \mathcal{G}_i(b_i)$$

binding energy matrix

(in unit of kT ≈ 0.6 kcal/mole)

binding free energy of "consensus" seq

$$S^* = \{b_1^*, b_2^*, ..., b_L^*\}$$

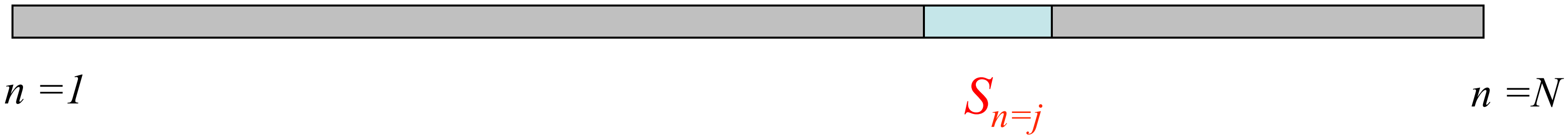


pos.	10	11	12	13	14	15	16	17
A	1.8	2.4	1.6	1.0	0	2.1	0.8	1.1
C	2.4	1.9	4.2	2.1	0.3	0	0	0
G	0	1.6	0	0	1.2	3.2	1.0	1.2
T	3.0	0	2.2	2.2	0.6	2.2	0.7	0.3

(D.S. Fields, Y. He, A. Al-Uzri & G. Stormo, 1997)

in vivo binding: Effect of the genomic background

Q: occupation freq f_j of a “target site” S_j in genomic DNA?



model genomic DNA as a collection of N “sites” of L nt each

$$S_n = \{b_1^{(n)}, b_2^{(n)}, \ldots, b_L^{(n)}\}$$

(with $N \sim 10^7$ for *E. coli*)

$$G_n \equiv G(S_n) = G^* + \Delta G_n$$

where

$$\Delta G_n = \sum_{i=1}^{L} g_{n,i}$$

- single TF in bacterium cell (assume TF confined to DNA)

$$f_j = \frac{1}{1 + \sum_{n \neq j} e^{(\Delta G_j - G_{bkg})/kT}}$$

$$e^{-\beta G_{bkg}} \equiv Z_{bkg} = \sum_{k \neq j} e^{-\beta \Delta G_k} + N e^{-\beta \Delta G_{ns}}$$

- effective *in vivo* binding

⇨ $$f_j \approx \frac{1}{1+\underbrace{\sum_{n\neq j}^{N} e^{(\Delta G_j - G_{bkg})/kT}}_{\widetilde{K}_j}} \qquad \widetilde{K}_j = e^{\beta\sum_{i=1}^{L} g_{j,i}}\, Z_{bkg}$$

To convert in concentration remember 1 molecule in E. coli volume≈1nM

- **binding depends on competition from the rest of the genome**
- even for "strong" target ($G_j \ll G_n$), large *N* can make effective binding weak

$$\text{e.g., if } \Delta G_j = 0,\ G_{ns} - G^* \approx 15kT, \text{ then } \widetilde{K}_j = N \cdot e^{-15} \approx 3 \text{ nM}$$

Note: for the Lac repressor, $K_{\mathrm{O1}} \approx 1 \text{ pM}$ *in vitro* while $\widetilde{K}_{O1} \approx 3\ n\mathrm{M}$

Typical cost of a mismatch: 1-3 kT ➔ $e^{\beta\Delta G} \approx 3-10$

➔Effect of the rest of genome at least equivalent to a single good site

Re-derivation by the grand canonical ensemble

$$f(S) = \frac{e^{\beta\mu}}{e^{\beta\mu} + e^{\beta G(S)}}$$

$$\beta\mu \propto \log(concentration)$$

$$f(S) \equiv \frac{[P \cdot S]}{[S] + [P \cdot S]} = \frac{[P]}{[P] + K(S)}$$

$$\approx \frac{[P]_{tot}}{[P]_{tot} + K(S)} \qquad \text{if } [S]_{tot} \ll [P]_{tot}$$

$$K(S) \equiv [P] \cdot [S] / [P \cdot S]$$

$$\propto e^{G(S)/kT}$$

Let’s use it to derive at the board the expression when multiple copies N_p of the TF are present.

$$f_j = \frac{1}{1 + \sum_{n \neq j}^{N} e^{\beta(\Delta G_j - \Delta G_{bkg})} \Big/ N_p} = \frac{1}{1 + e^{\beta \Delta G_j} Z_{bkg} \Big/ N_p}$$

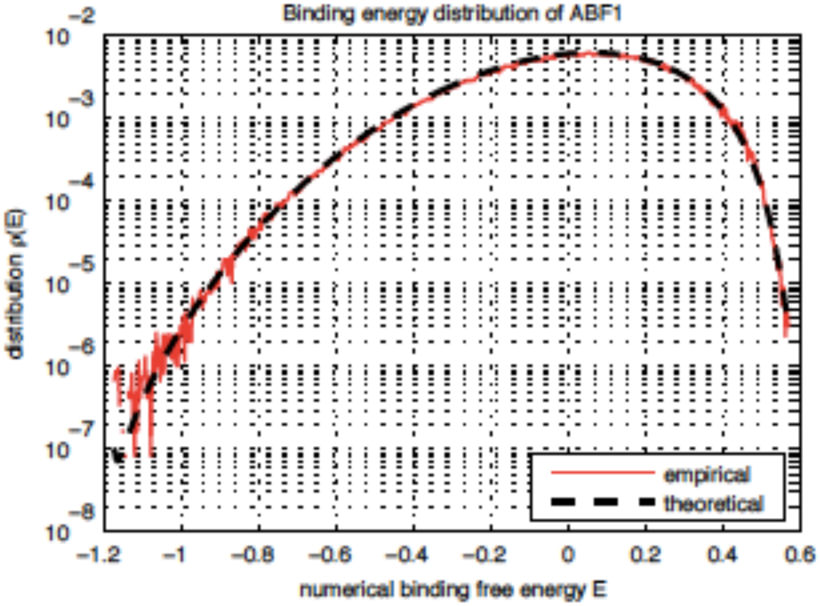
How to "set" $Z_{bkg} \approx \mathrm{N}^{*}_{\mathrm{p}}$? (a desired copy number where binding of consensus starts to be effective and not affected by binding at other sites)

"annealed approx" [cf: upcoming REM]

$$Z_{bkg} - Ne^{-\beta\Delta G_{ns}} = \sum_{n=1(\neq j)}^{N} e^{-\Delta G_n/kT} \approx N\cdot\mathbf{avg}\left[\!\left[e^{-\Delta G/kT}\right]\!\right] = N\cdot\mathbf{avg}\left[\!\left[\prod_{i=1}^{L} e^{-g_i(b)/kT}\right]\!\right]$$

$$= N\cdot\prod_{i=1}^{L}\left\{\mathbf{avg}\left[\!\left[e^{-g_i(b)/kT}\right]\!\right]\right\} = N\cdot\prod_{i=1}^{L}\left\{\sum_{b\in\{A,C,G,T\}} f_b\cdot e^{-g_i(b)/kT}\right\}$$

iid sequence with nt frequency f_b



➔ $Z_{bkg} \approx \mathrm{N}^{*}_{\mathrm{p}}$ from the <u>design</u> of TF-DNA interaction

Simple model to gain insight

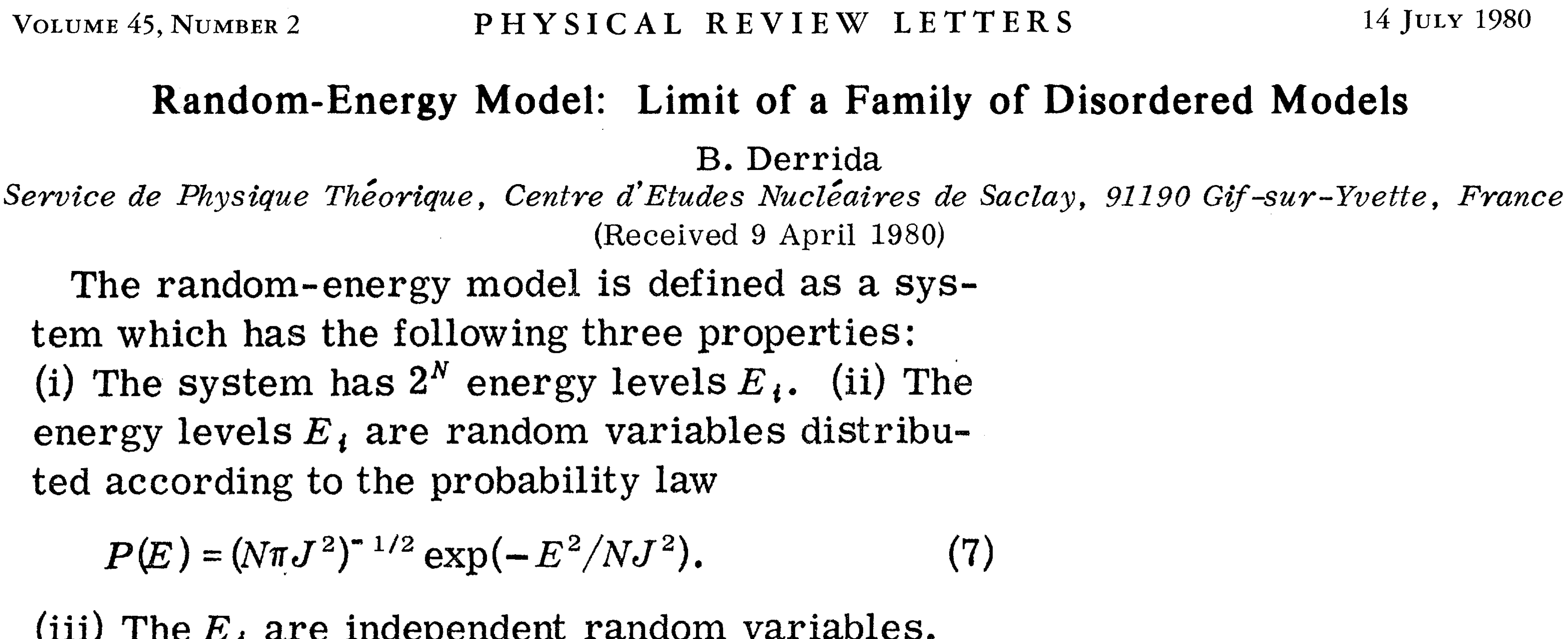
$$g_i(b) = \begin{cases} 0 & \text{if } \; b = b_i^* \\ \varepsilon & \text{if } \; b \neq b_i^* \end{cases} \Rightarrow \boxed{Z_{bkg} - Ne^{-\beta\Delta G_{ns}} = Z_{sp} \approx N \cdot \left[\tfrac{1}{4} + \tfrac{3}{4} e^{-\varepsilon/kT}\right]^L}$$

e.g. to have $Z_{sp} = 1$ for $N = 10^7$

ε/kT	1	2	3	4
L	25	15	13	12

$$Ne^{-\beta\Delta G_{ns}} \approx 1 \Rightarrow \Delta G_{ns} \approx 16kT$$

- physiological range: $\varepsilon \sim 2\ kT$
- biochem of TF-DNA interaction allows for flexible tuning of $\mathrm{Z_{bkg}}$



Volume 45, Number 2 PHYSICAL REVIEW LETTERS 14 July 1980

Random-Energy Model: Limit of a Family of Disordered Models

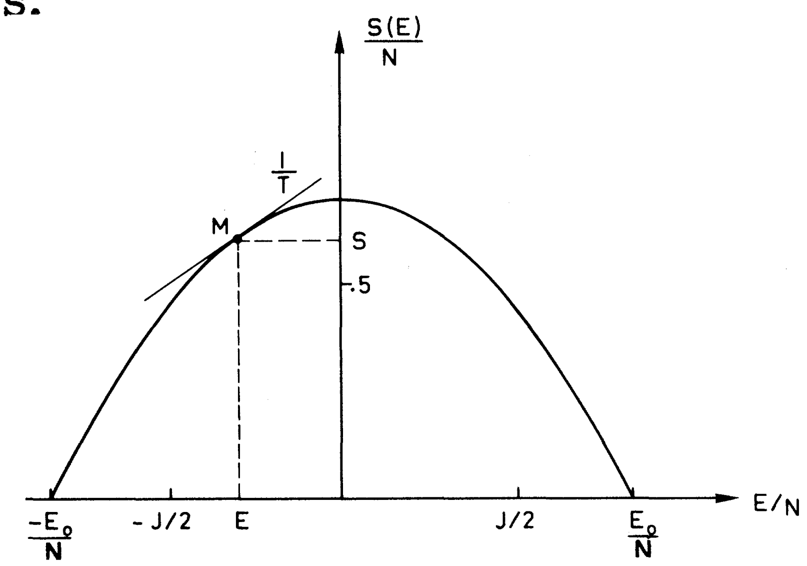
B. Derrida
Service de Physique Théorique, Centre d'Etudes Nucléaires de Saclay, 91190 Gif-sur-Yvette, France

(Received 9 April 1980)

The random-energy model is defined as a system which has the following three properties: (i) The system has 2^N energy levels E_i. (ii) The energy levels E_i are random variables distributed according to the probability law

$$P(E) = (N\pi J^2)^{-1/2} \exp(-E^2/NJ^2). \tag{7}$$

(iii) The E_i are independent random variables.



<log Z> is given by log<Z> as long as T>T$_c$, i.e. the entropy is positive and contributing states are >>1.

Derivation at the board

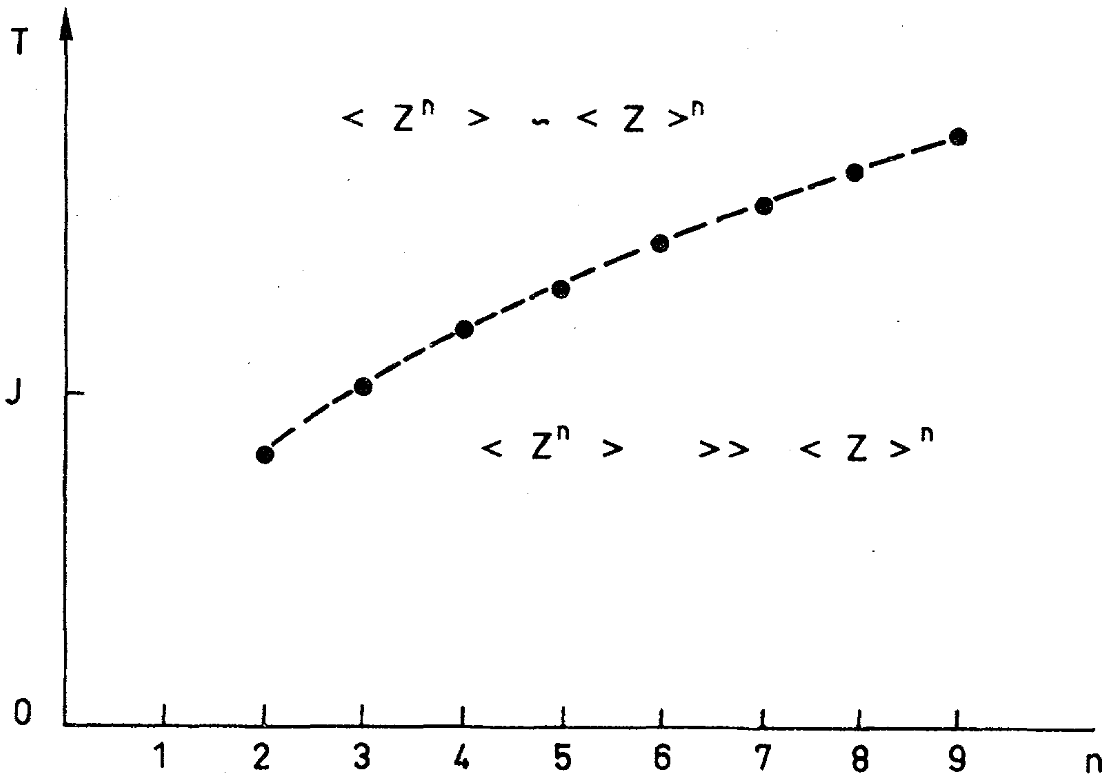
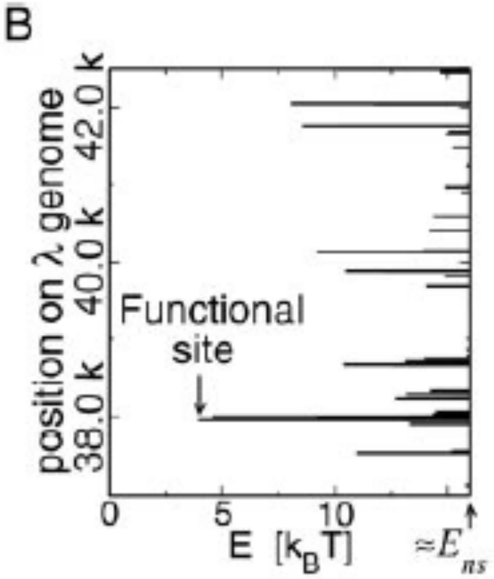


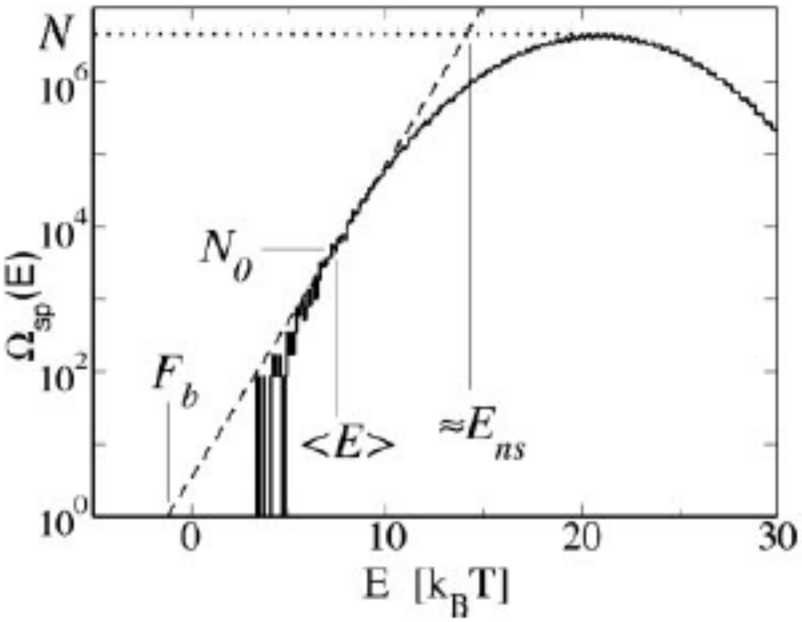
FIG. 1. The critical temperatures $T_n = \sqrt{n}\, T_c$ of the moments $\langle Z^n \rangle$ of the partition function. In the high-temperature region $T > T_n$, $\langle Z^n \rangle \sim \langle Z \rangle^n$. In the low-temperature region $T < T_n$, $\langle Z^n \rangle$ is much larger than $\langle Z \rangle^n$.

This is quite generic: all moments have their own critical temperature, where they start being dominated by fluctuations

Derivation at the board

Experimental data for Cro





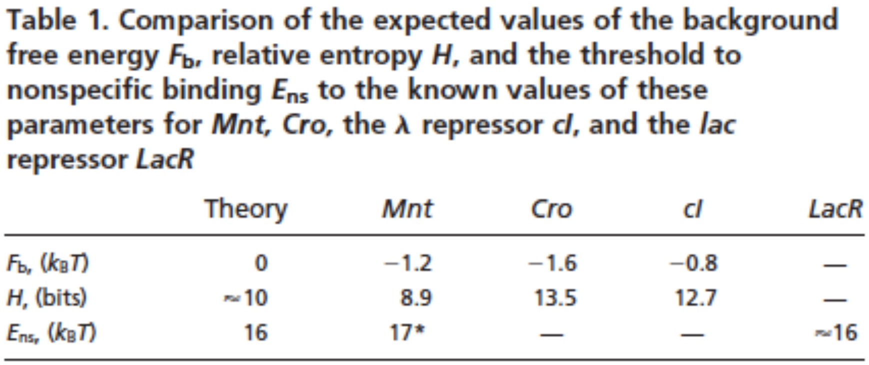
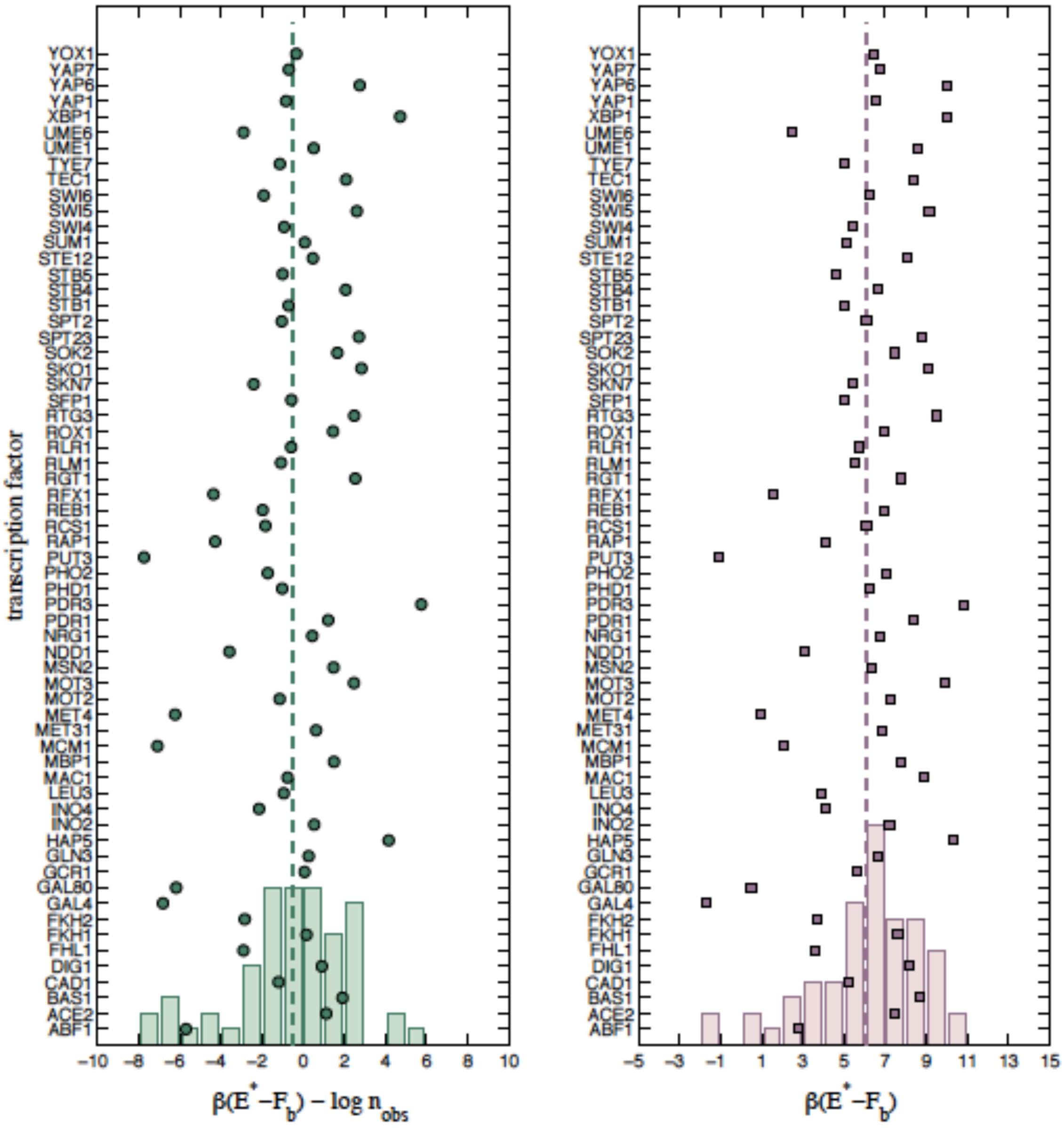


Table 1. Comparison of the expected values of the background free energy F_b, relative entropy H, and the threshold to nonspecific binding E_{ns} to the known values of these parameters for *Mnt, Cro,* the λ repressor *cI*, and the *lac* repressor *LacR*

	Theory	*Mnt*	*Cro*	*cI*	*LacR*
F_b, (k_BT)	0	−1.2	−1.6	−0.8	—
H, (bits)	≈10	8.9	13.5	12.7	—
E_{ns}, (k_BT)	16	17*	—	—	≈16

Experimental data for *S. cerevisiae*

The typical expression level of TFs is marginally sufficient for the binding of the strongest sites. The chemical potential is again largely independent of individual binding and dominated by many terms.



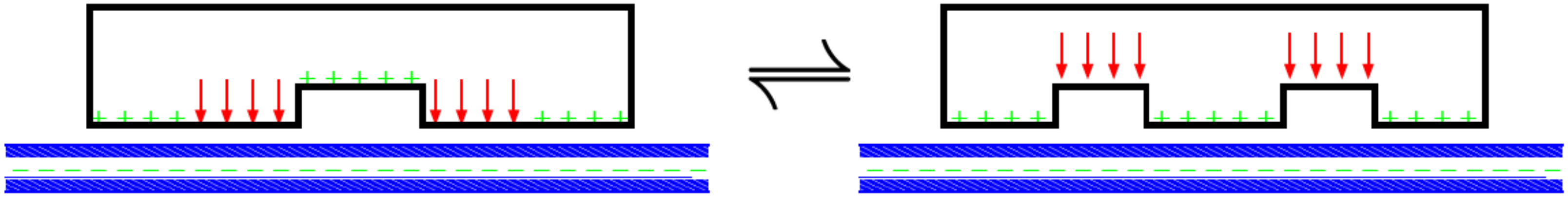
C. Kinetics of target search

- consider simple additive model of binding energy:

$$G_n = G^* + m(n) \cdot \varepsilon \qquad \text{where} \qquad m(n) = \left\| S_n - S^* \right\|$$

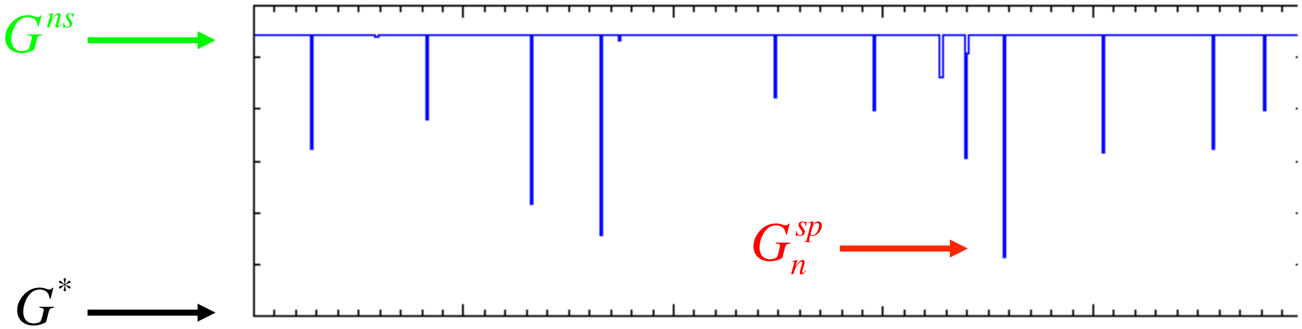
if valid for all $0 \le m \le L$, then the kinetics of target search would be slow since the environment is rugged with traps $\gg kT$

- **two-state model** of TF-DNA binding [Winter, Berg, von Hippel, 81]



specific binding: $G_n^{sp} = G^* + m(n) \cdot \varepsilon$

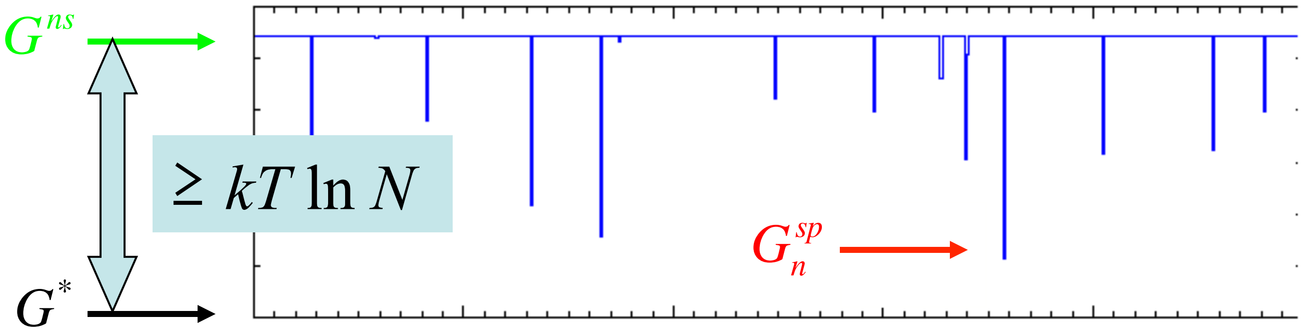
non-specific binding: G^{ns}



kinetic barriers reduced as

$$G^{ns} \to G^*$$

- if G^{ns} is too low, thermodynamic specificity will be lost



kinetic barriers reduced as

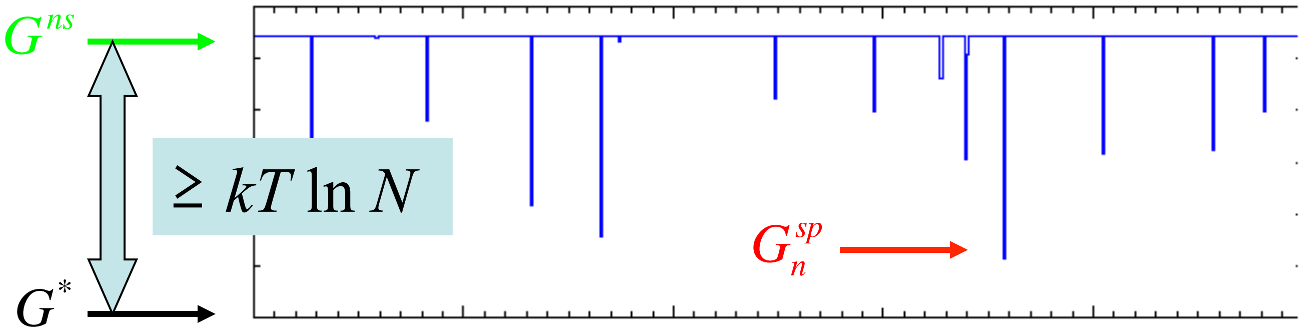
$$G^{ns} \to G^{*}$$

statistical mechanics of the two-state model:

$$Z \equiv \sum_{n=1}^{N} e^{-\left(G_n - G^{*}\right)/kT} \to \underbrace{\sum_{n=1}^{N} e^{-\left(G_n^{sp} - G^{*}\right)/kT}}_{Z^{sp}} + \underbrace{\sum_{n=1}^{N} e^{-\left(G^{ns} - G^{*}\right)/kT}}_{Z^{ns}}$$

➔ G^{ns} - $G^{*} \geq kT \ln N \approx 16\ kT$ ensures Z_{ns} small

- effect of kinetic slow down ?



kinetic barriers reduced as

$$G^{ns} \rightarrow G^*$$

-- for each trap with binding energy $G^{sp}{}_n < G^{ns}$

escape time: $\tau_n = \tau_0 \cdot e^{\left(G^{ns} - G_n^{sp}\right)/kT}$

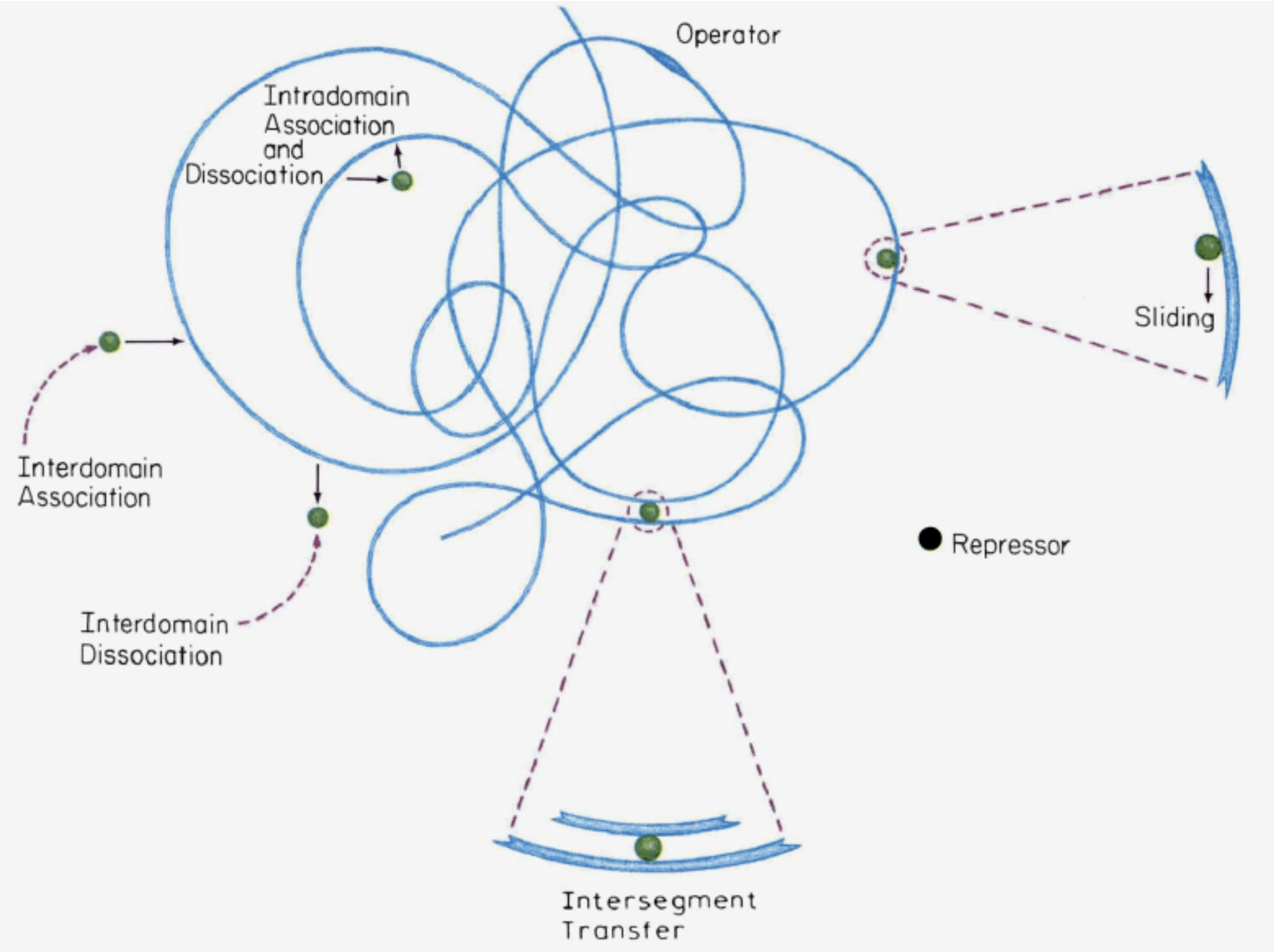
-- average escape time: $$\bar{\tau} = \tau_0 \cdot \sum_G \left[1 + e^{\left(G^{ns} - G\right)/kT}\right] \cdot \overbrace{\Omega(G)}^{\text{density of states}} \Big/ N$$

$$= \tau_0 \cdot \left[1 + e^{\left(G^{ns} - G^*\right)/kT} \cdot Z^{sp} / N\right]$$

➔ for $Z^{sp} \approx 1$, kinetic slowdown insignificant if G^{ns} - G^* ≤ $kT \ln N$

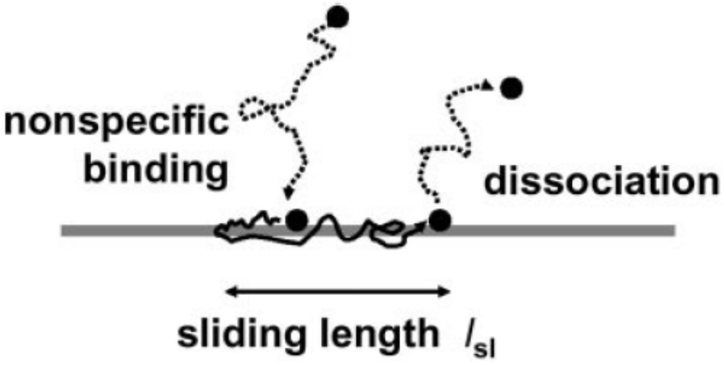
➔ both thermodynamics and kinetics okay if G^{ns} - G^* ≈ $kT \ln N$

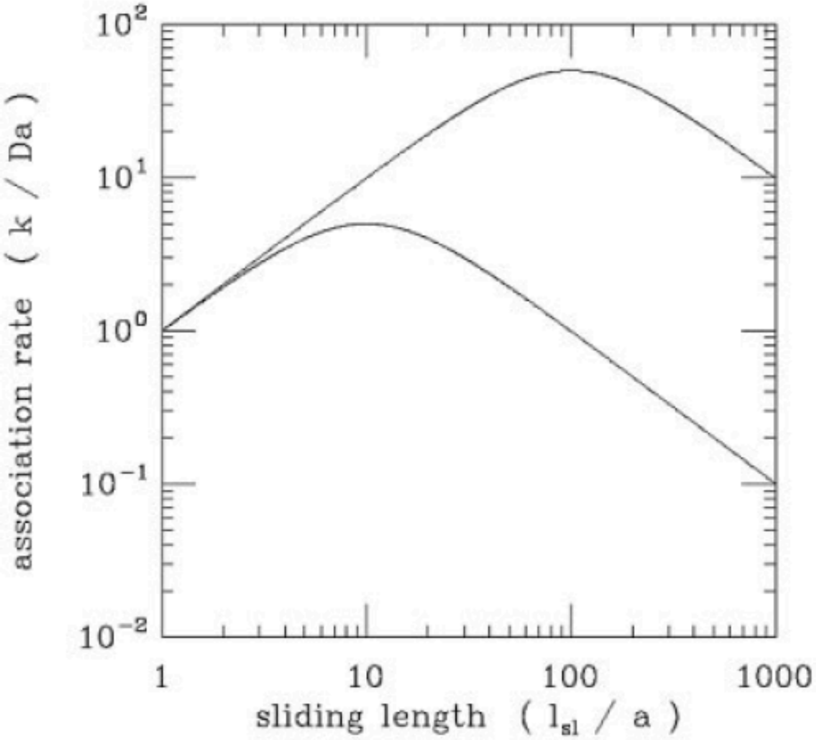
[Note: for the Lac and Arc repressors, G^{ns} - $G^* \approx 15\ kT$]



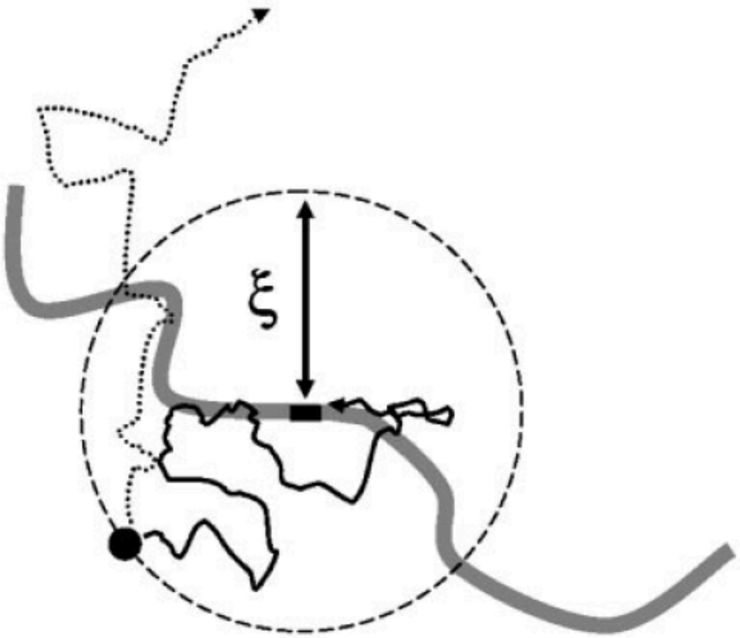
Various mechanisms for facilitated diffusion considered by Winter et al. in their series of papers (references at the end)

Sliding





Dependency of speed-up on the sliding length



Targeting radius

Refs where the material presented is discussed

Von Hippel PH, Berg OG, PNAS, 83, 1608, (1986).

Berg OG, von Hippel PH, J. Mol. Biol., 193, 723, (1987).

Gerland et al., PNAS, 99, 12015, (2002).

Aurell et al., Phys. Biol., 4, 134, (2007).

Derrida B., Phys. Rev. Lett., 45, 79, (1980).

Derrida B., Phys. Rev. B, 24, 2613, (1981).

Winter et al., Biochemistry, vol. 20(21) (1981) series of three papers

Halford SE & JF Marko, Nucl. Acids Res., 32, 3040, (2004)

Elf et al., Science, 316, 1191, (2007).

Hammar et al., Science, 336, 1595, (2012)