Introduction

- research at the interface of physics and biology

Some questions:

Every piece of detail seems to matter in molecular biology, nevertheless biological organisms are very robust, capable of withstanding or adapting to large perturbations. What are the secrets?

A complex system requires multiple sub-components to be in place before the function of the whole system can be realized; how can such systems self-organize in an evolutionary process?

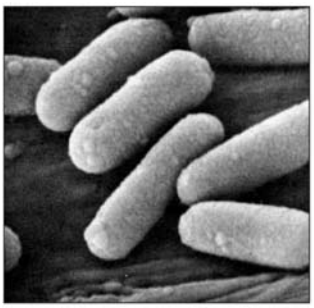
➔ ***Design principles of complex adaptive systems***

- role of theory in biology:
  - link across different scales, e.g., from components to systems -- how?
  - formulate constraints and expectation -- why?
  - make the right conceptual simplifications [cf: entropy and heat engine]

➔ *new concepts and principles from new perspectives*

This series: quantitative molecular biology of bacteria

Bacterial physiology (*E. coli*)

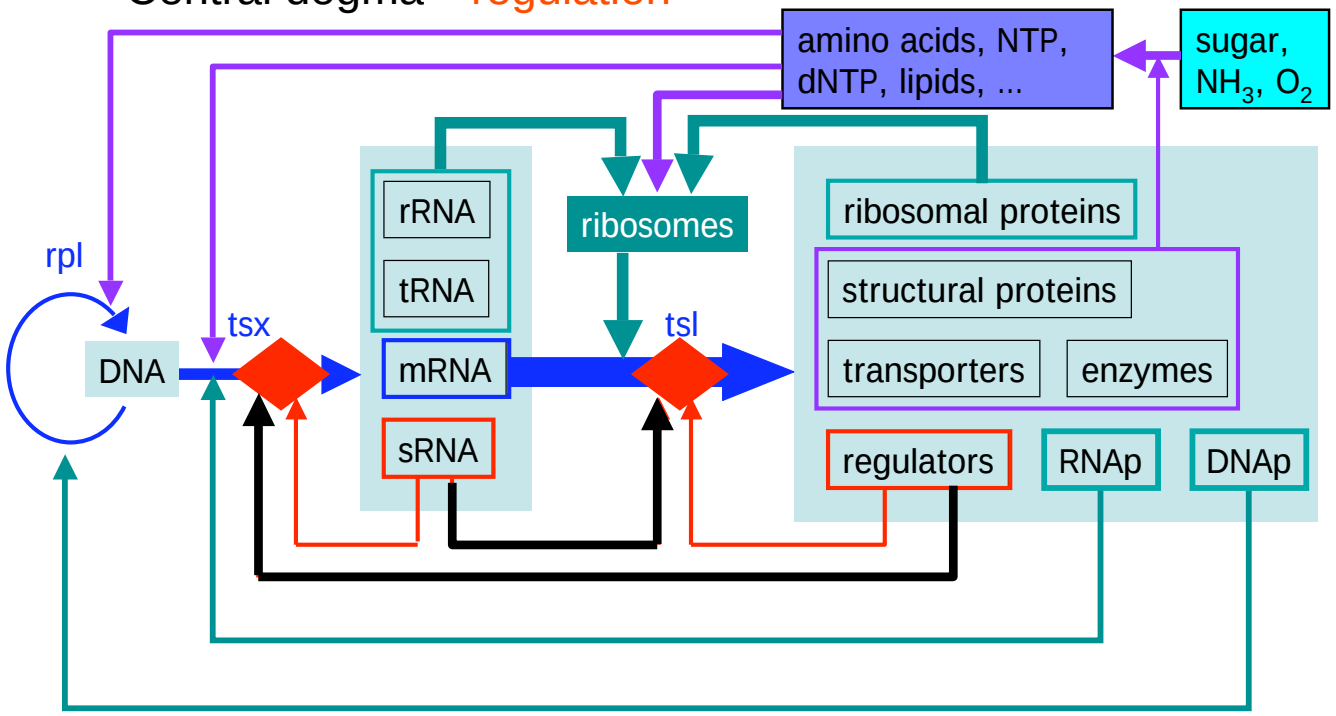


- growth glucose + NH_3 + O_2 ➔ biomass
- survival

bacteria can sense the environment and adjust "life style" according to the growth condition/medium

- exponential growth: doubling time from 20 min to > 200 min
- coping with stressful conditions:
  - motility: flagella synthesis and chemotaxis
  - osmotic response: porin synthesis
  - heat shock response: chaperons
  - SOS response (e.g., to DNA damage)
  - quorum sensing, biofilms, bacterial community
- non-growth condition
  - stationary phase
  - dormancy
  - sporulation (e.g., *B. subtilis*)
  - competence, conjugation (exchange of genetic materials)

Central dogma + regulation

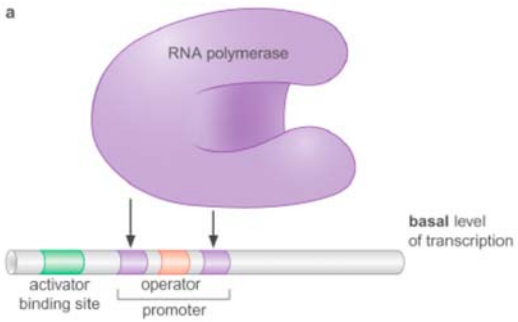


- tsx initiation control by transcription factors (TF)
- tsl initiation control by sRNA and RNA-binding proteins
- tsx termination control by sRNA and anti-terminators
- control of mRNA and protein degradation

} coupled to environmental signals; coord growth program

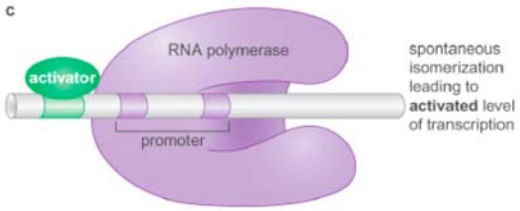
❖ transcriptional initiation control

- RNAp binding to promoter



- modulation of RNAp-promoter affinity via activators and repressors

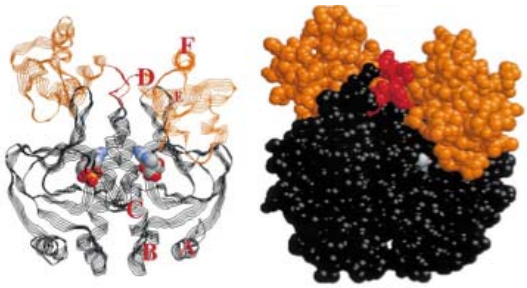
- allosteric activation/deactivation of activators and repressors by inducers or signaling molecules



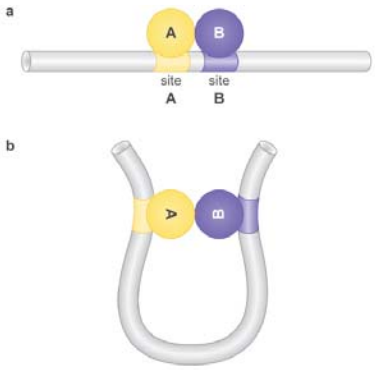
➔ net result: rate of tsx init dependent on cellular conc of activators/repressor controlled, e.g., by inducer molecules

- Molecular determinants of transcriptional initiation control
  - protein-DNA interaction

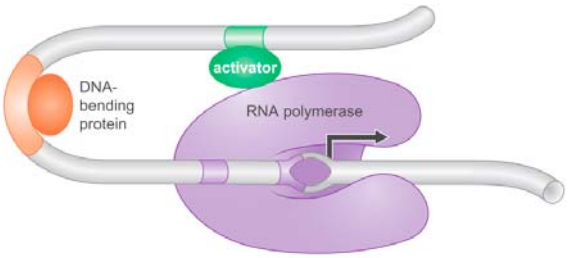
  - protein-ligand interaction



  - protein-protein interaction



  - necessity of DNA looping



Topic 1: Protein-DNA Interaction

- Goals:
  - find DNA binding target seq for each transcription factor (TF)
  - find the affinity of a TF to its DNA target as a function of its cellular concentration *in vivo*
  - find how the TF-DNA affinity depends on the target sequence
- Problems:
  - thousands of TFs each with distinct target sequence; only a few characterized in detail experimentally
  - *ab initio* molecular calculation difficult even when TF-DNA co-crystal structure available
  - need to deal with the entire genomic DNA seqs *in vivo*

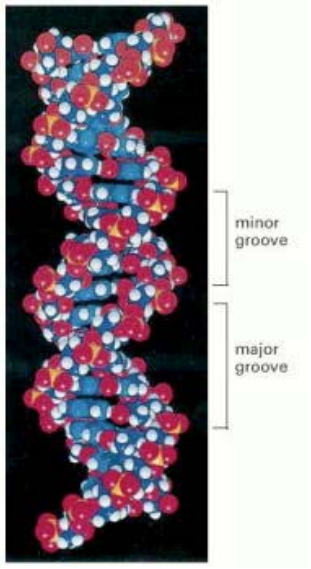
Statistical physics:

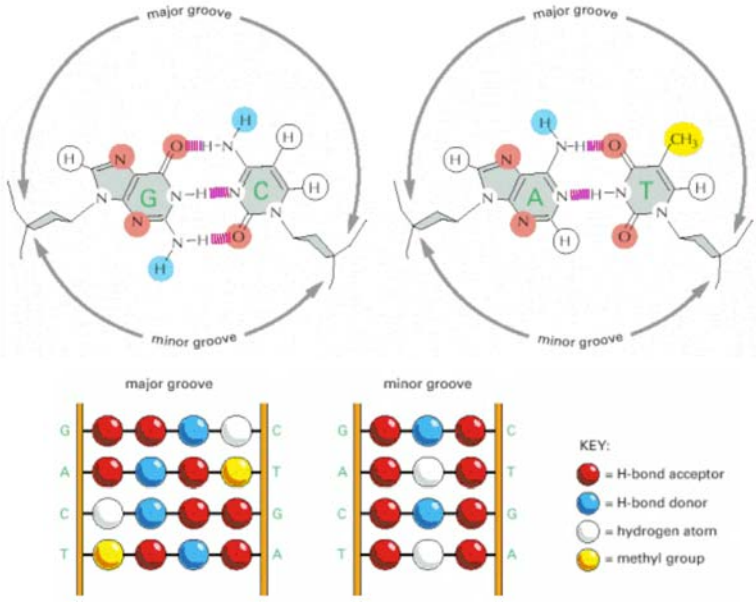
➔ ways to think quantitatively about TF-DNA interaction in the absence of detailed microscopic information

➔ link from molecule to function (an illustrative case)

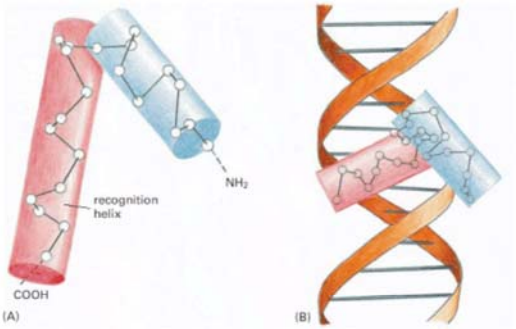
A. Empirical facts

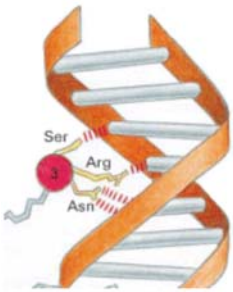
1. Transcription Factors
   - size: ~5nm (10-20 bp)
   - molecular basis of sequence recognition



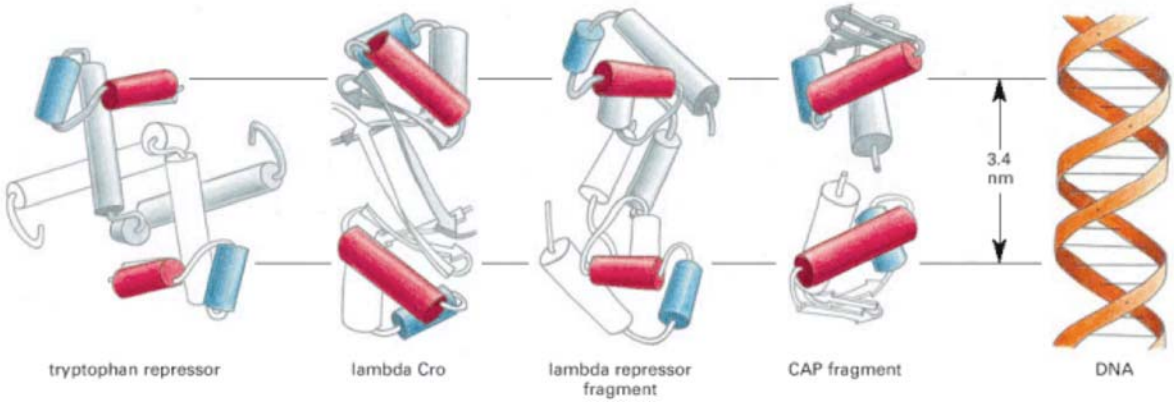


- various molecular strategies
  - Helix-Turn-Helix



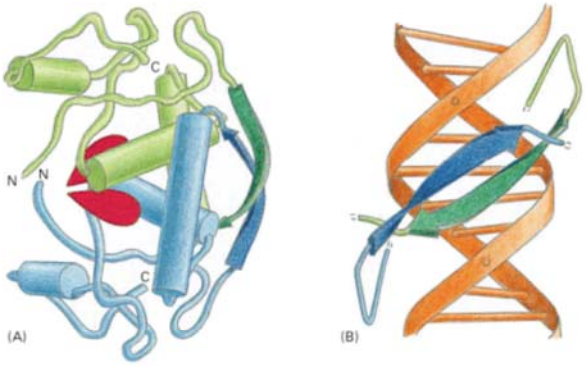


well-known examples in bacteria (note: homodimers)



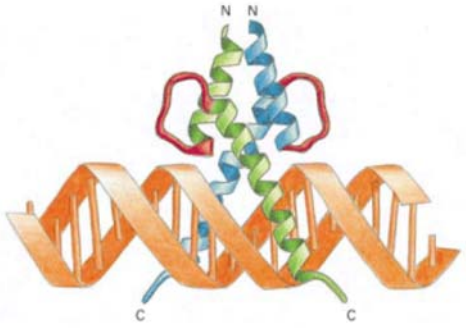
– zinc-finger domain

– beta-sheets



– leucine zipper

– helix-loop-helix



2. DNA binding sequences

- typically 10-20 bp in bacteria

protein	target sequence
lac repressor	5' AATTGTGAGCGGATAACAATT 3' TTAACACTCGCCTATTGTTAA
CRP	TGTGAGTTAGCTCACT ACACTCAATCGAGTGA
λ repressor	TATCACCGCCAGAGGTA ATAGTGGCGGTCTCCAT

- lots of sequence variants
- consensus sequence often palindromic
- 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

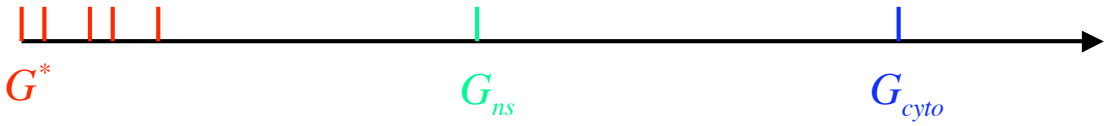
- passive (no energy consumption)
- strong electrostatic attraction indept of binding seq

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

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

  ($RT \approx 0.62$ kcal/mole at 37C)
- additional energy gained from hydrogen bonds to preferred sequences

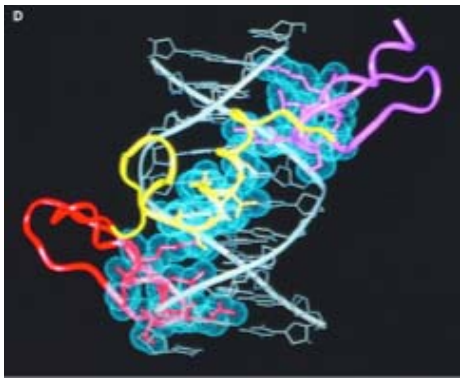
  strongest binder: $G^* - G_{ns} \simeq -15RT$



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

---

- relative binding affinity for Mnt



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?

B. Thermodynamics of DNA target recognition

- binding sequence (L nucleotides):

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

- binding constant (*in vitro*)

$$K(S) \equiv [P] \cdot [S] / [P \cdot S] \propto e^{G(S)/kT}$$

- fraction of sequence bound:

$$p(S) = \frac{[P]}{[P] + K(S)}$$

- 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)

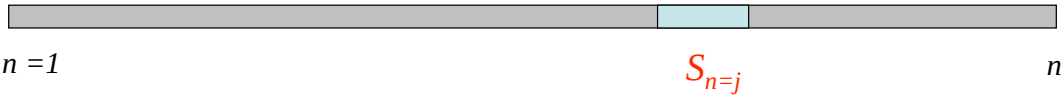
binding free energy of "consensus" seq $S^* = \{b_1^*, b_2^*, ..., b_L^*\}$

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 Genomic background

Q: occupation probab p_j of a "target site" S_j in genomic DNA?



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

$$S_n = \{b_1^{(n)}, b_2^{(n)}, ..., b_L^{(n)}\}$$ (with $N \sim 10^7$ for *E. coli*)

in vitro binding constant: $K_n \equiv K(S_n) = [P] \cdot [S_n] / [P \cdot S_n] \propto e^{G_n/kT}$

binding energy: $G_n \equiv G(S_n) = G^* + \sum_{i=1}^{L} \mathcal{G}_i\left(b_i^{(n)}\right)$

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

$$\Rightarrow \quad p_j = \frac{e^{-G_j/kT}}{e^{-G_j/kT} + \sum_{n \neq j} e^{-G_n/kT}} = \frac{1}{1 + \sum_{n \neq j} e^{(G_j - G_n)/kT}}$$

- multiple (N_P) TFs

$$\Rightarrow \quad p_j \approx \frac{1}{1 + \left(\sum_{n \neq j} e^{(G_j - G_n)/kT}\right) / N_P}$$

- cf: *in vitro* binding

$$p(S) = \frac{[P]}{[P] + K(S)} = \frac{1}{1 + K(S)/[P]}$$

• effective *in vivo* binding constant

$$p_j \approx \frac{1}{1+\underbrace{\left(\sum_{n\neq j}^{N} e^{(G_j-G_n)/kT}\right)}_{\widetilde{K}_j}\Big/ N_P}$$

• cf: *in vitro* binding

$$p(S) = \frac{1}{1+K(S)/[P]}$$

– depends on competiton from the rest of the genome
– even for "strong" target ($G_j \ll G_n$),
large number of genomic sites (N) can make effective binding very weak

• since typical $N_P = 1 \sim 1000$ molecules/cell (nM),
expect functional demand for $\widetilde{K}_j = 1 \sim 1000$ nM

$$\widetilde{K}_j = e^{\sum_{i=1}^{L} \mathcal{G}_i\left(b_i^{(j)}\right)/kT} \cdot \underbrace{\sum_{n=1(\neq j)}^{N} e^{-\sum_{i=1}^{L} \mathcal{G}_i\left(b_i^{(n)}\right)/kT}}_{\equiv Z \approx 1} \approx \begin{cases} 1 & \text{consensus seq} \\ e^{1\sim 3} = 3 \sim 10 & \text{one mismatch} \end{cases}$$

(Mnt matrix applied to *E. coli* genome)

➔ effect of the rest of genome: equivalent to a single site S^*
➔ $\widetilde{K}_j$ **tunable** in the desired range by "adjusting" no. mismatches

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

How to "set" $Z \approx 1$?

$$Z = \sum_{n=1(\neq j)}^{N} e^{-\sum_{i=1}^{L} \mathcal{G}_i\left(b_i^{(n)}\right)/kT} \approx N \cdot \mathbf{E}\left[\prod_{i=1}^{L} e^{-\mathcal{G}_i(b)/kT}\right]$$

"annealed approx" (valid for large $\ln N$) [cf: Derrida's REM]

$$= N \cdot \left[\sum_{b\in\{A,C,G,T\}} f_b \cdot e^{-\mathcal{G}_i(b)/kT}\right]^L \approx 1$$

iid sequence with nt frequency f_b ; Mnt matrix with f_b of *E. coli*

➔ $Z \approx 1$ from the <u>design</u> of TF-DNA interaction ($\mathcal{G}_i(b)$, L)
➔ use simpler model to gain insight

$$\mathcal{G}_i(b) = \begin{cases} 0 & \text{if } b = b_i^* \\ \varepsilon & \text{if } b \neq b_i^* \end{cases} \quad \Rightarrow \quad Z \approx N \cdot \left[\frac{1}{4} + \frac{3}{4} e^{-\varepsilon/kT}\right]^L$$

• physiological range: $\varepsilon \sim 2\ kT$
• $\widetilde{K} \approx e^{(\#\text{mm})\cdot\varepsilon/kT}$ (5-10x per mismatch)
• biochem of TF-DNA interaction allows for flexible tuning of $\widetilde{K}$

to have $Z = 1$ for $N = 10^7$

ε/kT	1	2	3	4
L	25	15	12	11

C. Kinetics of target search

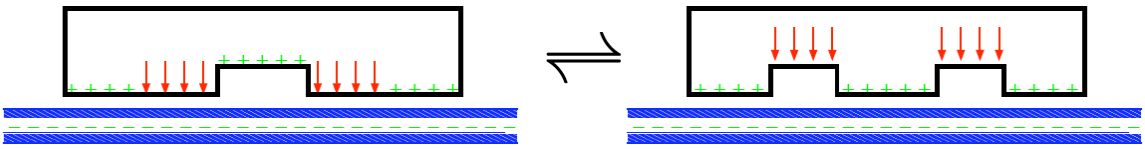
• consider simple additive model of binding energy:

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

if valid for all $0 \le m \le L$, then the kinetics of target search would be slow

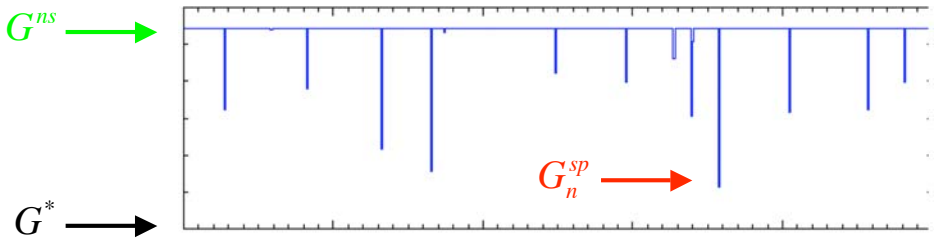
since $\mathrm{var}(G) \approx \sqrt{L}\cdot\varepsilon \gg kT$

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



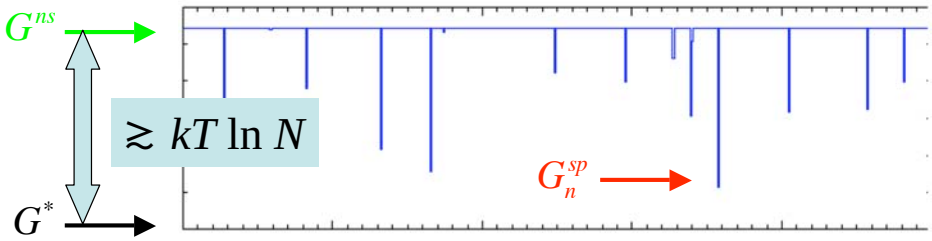
specific binding: $G_n^{sp} = G^* + m(n)\cdot\varepsilon$ non-specific binding: G^{ns}

Boltzmann weight: $e^{-G_n/kT} \to e^{-G_n^{sp}/kT} + e^{-G^{ns}/kT}$



kinetic barriers reduced as $G^{ns} \to G^*$

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



kinetic barriers reduced as $G^{ns} \to G^*$

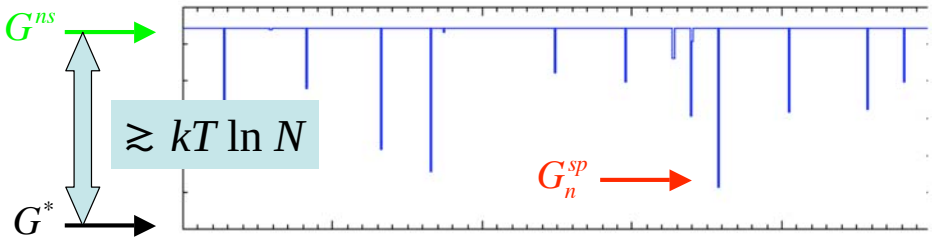
statistical mechanics of the two-state model:

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

➔ for $Z \approx 1$, need to have $Z^{sp} \approx 1$ and $Z^{ns} \le 1$

➔ $G^{ns} - G^* \gtrsim kT \ln N \approx 16\, kT$

• effect of kinetic slow down ?



kinetic barriers reduced as $G^{ns} \to G^*$

-- 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: $\bar{\tau} = \tau_0 \cdot \int dG \left[1 + e^{\left(G^{ns} - G\right)/kT}\right] \cdot \Omega(G) \Big/ N$

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

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

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

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

Global search dynamics (smooth landscape)

• 1D diffusion along the genome:

$$\left.\begin{aligned} N &= 5 \times 10^6 \text{ bp} \approx 1 \text{ mm} \\ D_1 &\approx 1\ \mu\text{m}^2 / \text{sec} \end{aligned}\right\} \quad T_{1D} \sim \frac{N^2}{D_1} \sim 10^6 \text{ sec}$$

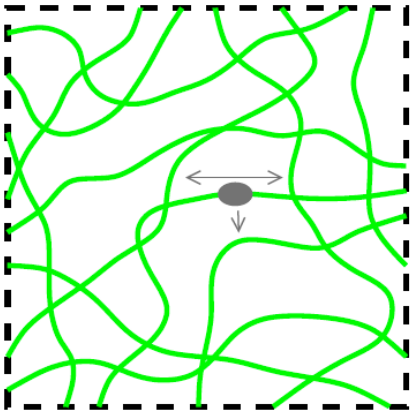
• 3D diffusion direct from cytoplasm:

$$\left.\begin{aligned} V_{cell} &\approx 3\ \mu\text{m}^2 \\ \ell_{TF} &\approx 15 \text{ bp} = 5 \text{ nm} \\ D_{cyto} &\approx 10\ \mu\text{m}^2 / \text{sec} \end{aligned}\right\} \quad T_{3D} \sim \frac{1}{4\pi} \frac{V_{cell}}{\ell_{TF} \cdot D_{cyto}} \sim 10 \text{ sec}$$

– faster mainly due to the reduced redundancy of 3D random walk
– but TFs typically associate strongly to DNA (subcompartmentalization)
[e.g., for the Lac repressors, $G^{cyto} - G^{ns} \approx 15\ kT$]

- combined 1D/3D search:
  - assume random DNA packing
  - hopping dist: $N_x \sim 300$ bp
  - hopping time: $T_\times \sim \frac{N_\times^2}{D_1} \sim 10$ msec

$$T_{1D/3D} \sim \frac{1}{4\pi} \frac{V_{cell}}{N_\times \cdot (N_\times^2 / T_\times)} \sim 10\,\text{sec}$$



- 3D diffusion direct from cytoplasm:

$$\left.\begin{aligned} V_{cell} &\approx 3\ \mu\text{m}^2 \\ \ell_{TF} &\approx 15\ \text{bp} = 5\ \text{nm} \\ D_{cyto} &\approx 10\ \mu\text{m}^2/\text{sec} \end{aligned}\right\} \quad T_{3D} \sim \frac{1}{4\pi} \frac{V_{cell}}{\ell_{TF} \cdot D_{cyto}} \sim 10\,\text{sec}$$

  - faster mainly due to the reduced redundancy of 3D random walk
  - but TFs typically associate strongly to DNA (subcompartmentalization)
    [e.g., for the Lac repressors, $G^{cyto} - G^{ns} \approx 15\ kT$]

Summary:

- specificity of target recognition: $Z^{sp} \approx 1$
  - ➔ $\varepsilon \approx 2$ kT, $L \approx 15$ bp, leading to $\widetilde{K}_j \approx e^{m_j \varepsilon / kT}$
  - ➔ affinity of target sites become “programmable”
- kinetic accessability of target: $G^{ns} - G^* \approx 15\ kT$
- combined 1D/3D search

➔ to what extent is “programmable” interactions used ?

➔ search process for multimer?

➔ eukaryotes?

many differences, e.g., $N_p = 10^2 \sim 10^4$ in budding yeast
(need another von Hippel!)