Topic 1: Protein-DNA Interaction

- Goals:
  - find DNA binding target seqs 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 sequences; 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 seq *in vivo*

Statistical physics:

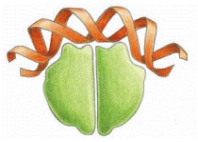
➔ ways to think quantitatively about TF-DNA interaction in the absence of detailed microscopic information

➔ link from molecule to function (an illustrative case)

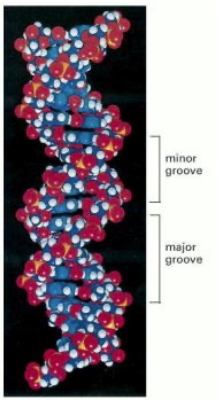
1

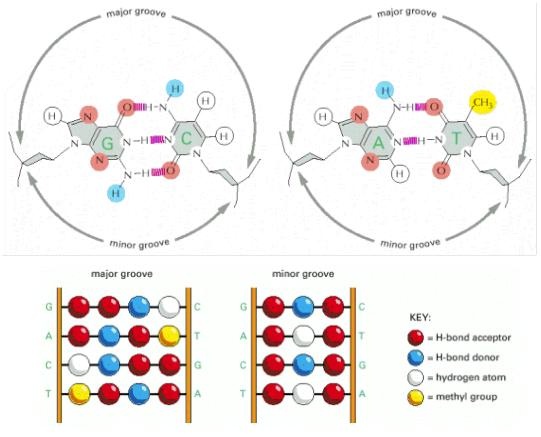
A. Empirical facts

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



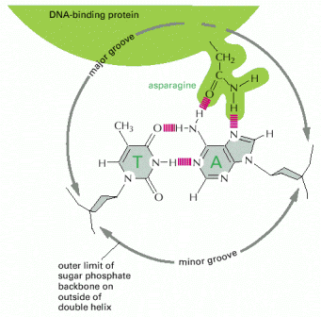
   - molecular basis of sequence recognition

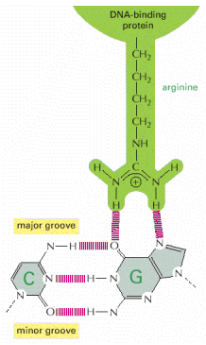




2

- contact between TF and DNA

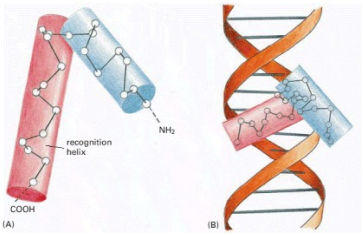


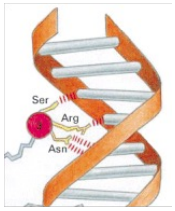


➔ structure of a TF must place the appropriate amino acids next to the base pairs they contact

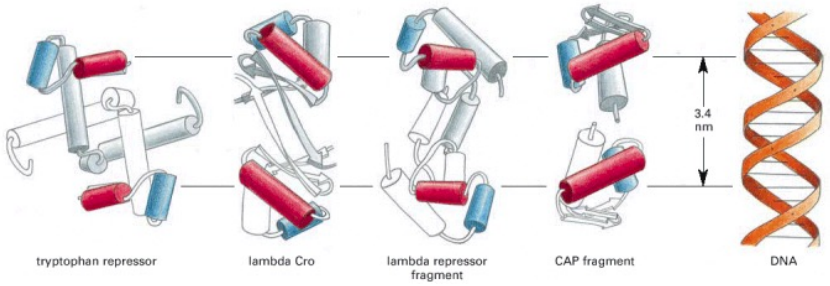
3

- various molecular strategies
  - Helix-Turn-Helix



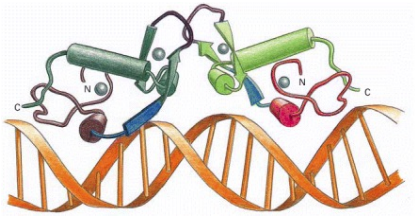


well-known examples in bacteria (note: homodimers)

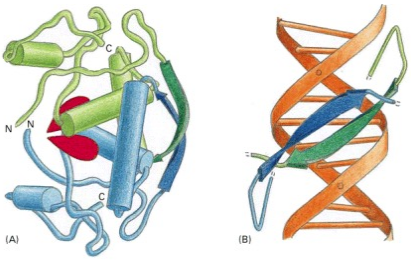


4

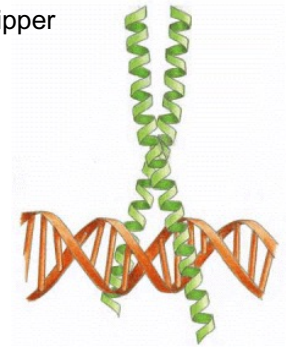
– zinc-finger domain



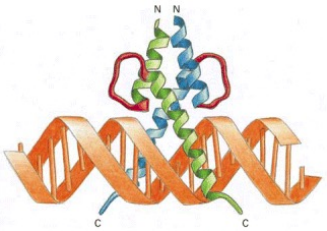
– beta-sheets



– leucine zipper



– helix-loop-helix



5

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

6

3. TF-DNA interaction

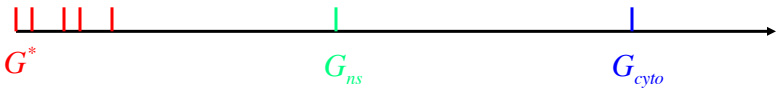
- passive (no energy consumption)
- strong electrostatic attraction <u>independent</u> of binding seq

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

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

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

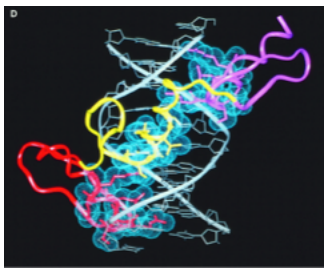
  strongest binder: $G^* - G_{ns} \simeq -15kT$



- <u>graded increase</u> in binding energy for sequences with partial match to the preferred sequence

7

• 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?

8

- Issues to be 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

9

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

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

cell vol: few um^3
1/V_{cell} ~ 1 nM

- dissociation constant (*in vitro*)

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

- fraction of sequence bound:

$$f(S) \equiv \frac{[P\cdot S]}{[S]+[P\cdot S]} = \frac{[P]}{[P]+K(S)} \approx \frac{[P]_{tot}}{[P]_{tot}+K(S)} \quad \text{if } [S]_{tot} \ll [P]_{tot}$$

- approx. <u>additive</u> 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)

10

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

$S_n = \{b_1^{(n)}, b_2^{(n)}, ..., b_L^{(n)}\}$ (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: $G_n \equiv G(S_n) = G^* + \Delta G_n$, where $\Delta G_n \equiv \sum_{i=1}^{L} \mathcal{G}_i\left(b_i^{(n)}\right)$

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

$$\Rightarrow \quad f_j = \frac{[P\cdot S_j]}{\sum_{n=1}^{N}[P\cdot S_n]} = \frac{K_j^{-1}}{\sum_{n=1}^{N}K_n^{-1}} = \frac{1}{1+\sum_{n\neq j}K_j/K_n} = \frac{1}{1+\sum_{n\neq j}e^{(\Delta G_j-\Delta G_n)/kT}}$$

- multiple (N_P) TFs [grand canonical ens]

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

- cf: *in vitro* binding

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

11

- effective *in vivo* binding constant

$$\Rightarrow \quad f_j \approx \frac{1}{1+\underbrace{\left(\sum_{n\neq j}^{N}e^{(\Delta G_j-\Delta G_n)/kT}\right)}_{\widetilde{K}_j}/N_P} \qquad \Rightarrow \quad K(S) = \widetilde{K}_j/V_{cell} = \widetilde{K}_j \text{ in nM}$$

- cf: *in vitro* binding

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

  - 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

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

- since typical $N_P = 1 \sim 1000$ molecules/cell (nM),

expect functional demand for $\widetilde{K}_j = 1 \sim 1000$ nM

$$\widetilde{K}_j = e^{\frac{\Delta G_j}{kT}}\cdot\underbrace{\sum_{\{n=1(\neq j)\}}^{N} e^{-\frac{\Delta G_n}{KT}}}_{\equiv Z\approx 1} \approx \begin{cases} 1 & \text{consensus seq} \\ e^{1\sim 3} = 3\sim 10 & \text{each mismatch} \end{cases}$$

(Mnt matrix applied to *E. coli* genome or *randomly scrambled* genomes)

➔ effect of the rest of genome: comparable to <u>one</u> good 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

12

How to "set" $Z \approx 1$? "annealed approx" (valid for large $\ln N$) [cf: Derrida's REM]

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

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

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

➔ $Z \approx 1$ from the design of TF-DNA interaction $(\mathcal{G}_i(b), L)$

➔ use simpler model to gain insight

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

- 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

13

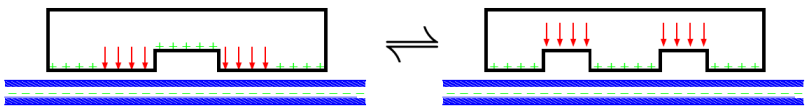
C. Kinetics of target search

- consider simple additive model of binding energy:

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

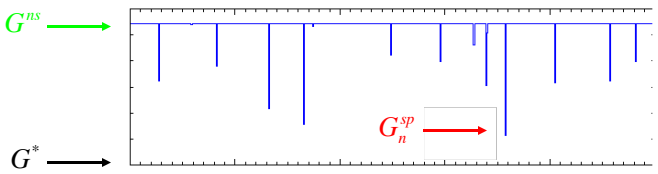
if valid for all $0 \le m \le L$, then the kinetics of target search would be slow since G_n - G_{n-1} is typically of the order $\text{std}(G) \approx \sqrt{L} \cdot \varepsilon \gg kT$

- **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}

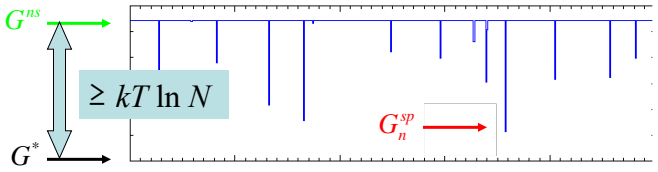
Boltzmann weight: $e^{-G_n/kT} \to e^{-G_n^{sp}/kT} + e^{-G^{ns}/kT}$



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

14

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



kinetic barriers reduced as

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

statistical mechanics of the two-state model:

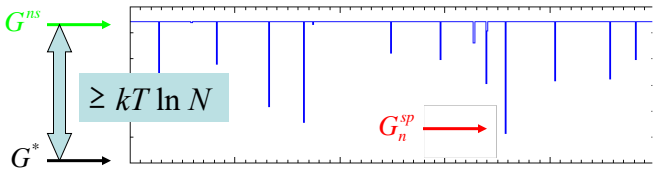
$$Z \equiv \sum_{n=1}^{N} e^{-\left(G_n - G^{*}\right)/kT} \rightarrow \underbrace{\sum_{n=1}^{N} e^{-\left(G_n^{sp} - G^{*}\right)/kT}}_{Z^{sp}} + \underbrace{\sum_{n=1}^{N} e^{-\left(G^{ns} - G^{*}\right)/kT}}_{Z^{ns}}$$

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

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

15

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

density of states

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

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

➔ 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 and Arc repressors, G^{ns} - $G^{*} \approx 15\, kT$]

16

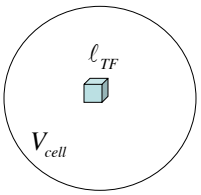
D. Global search dynamics (smooth landscape)

- 1D diffusion along the chromosome: [Elf & Xie, 2007]

$$\left.\begin{array}{l} N = 5\times 10^6 \text{ bp} \approx 1 \text{ mm} \\ D_1 \approx 0.1\ \mu\text{m}^2/\text{sec} \end{array}\right\} \quad T_{1D} \sim \frac{N^2}{D_1} \sim 10^7 \text{ sec}$$

-- may be overcome by increasing # of TF; for parallel search, $T_{1D} \sim 1/(\#TF)^2$
-- cost: covers the chromosome with lots of "useless" TFs

- 3D diffusion directly from the cytoplasm:



$$T_{3d} \sim \frac{1}{4\pi}\frac{V_{cell}}{\ell_{TF}\cdot D_{cyto}} = \frac{1}{4\pi}\underbrace{\left(V_{cell}/\ell_{TF}^3\right)}_{\text{search volume}}\cdot\underbrace{\left(\ell_{TF}^2/D\right)}_{\text{search time per volume}} \sim 10 \text{ sec}$$

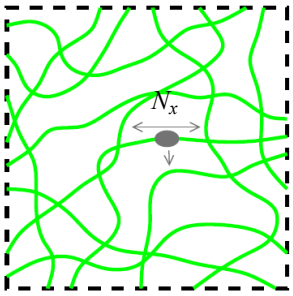
$V_{cell} \approx 3\mu m^3$
$\ell_{TF} \approx 15bp = 5nm$
$D_{cyto} \approx 10\mu m^2/\text{sec}$

– 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} - $G^{ns} \approx 15\ kT$]

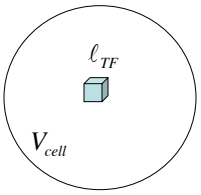
17

- combined 1D/3D search:
  - dense DNA packing in cell
  - short-time: slide on DNA (over scale N_x)
  - long-time: random walk on 3D network
  - slide dist: $N_x \sim 300$ bp
  - slide time: $T_x \sim N_x^2/D_1 \sim 0.1$ sec

$$T_{1D/3D} \sim \frac{1}{4\pi}\frac{V_{cell}}{N_x\cdot(N_x^2/T_x)} \sim 100 \text{ sec}$$



- 3D diffusion directly from the cytoplasm:



$$T_{3d} \sim \frac{1}{4\pi}\frac{V_{cell}}{\ell_{TF}\cdot D_{cyto}} = \frac{1}{4\pi}\underbrace{\left(V_{cell}/\ell_{TF}^3\right)}_{\text{search volume}}\cdot\underbrace{\left(\ell_{TF}^2/D\right)}_{\text{search time per volume}} \sim 10 \text{ sec}$$

$V_{cell} \approx 3\mu m^3$
$\ell_{TF} \approx 15bp = 5nm$
$D_{cyto} \approx 10\mu m^2/\text{sec}$

– 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} - $G^{ns} \approx 15\ kT$]

18

Summary:

- specificity of target recognition: $Z^{sp} \approx 1$
  - ➔ $\varepsilon \approx 2$ kT, $L \approx 15$ bp, gives $\widetilde{K}_j \equiv \sum_{n \neq j}^{N} e^{(G_j - G_n)/kT} \approx e^{m_j \varepsilon / kT}$
  - ➔ affinity of target sites become "programmable"
- kinetic accessability of target predicts 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!)

19