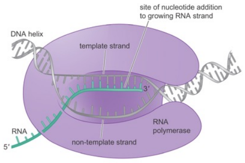
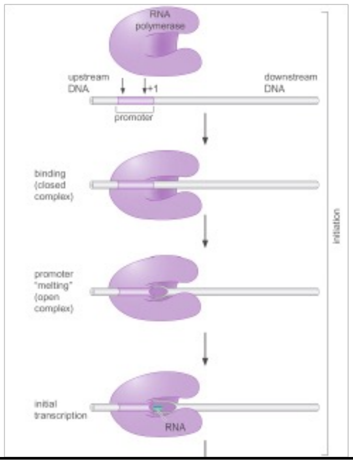
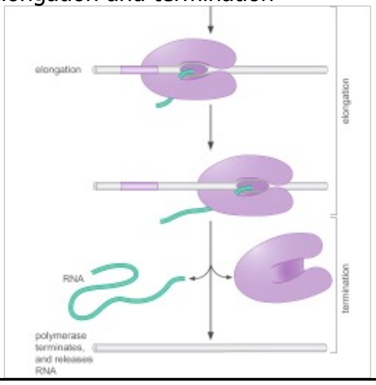
molecular biology of transcription (RNA synthesis)



transcriptional initiation

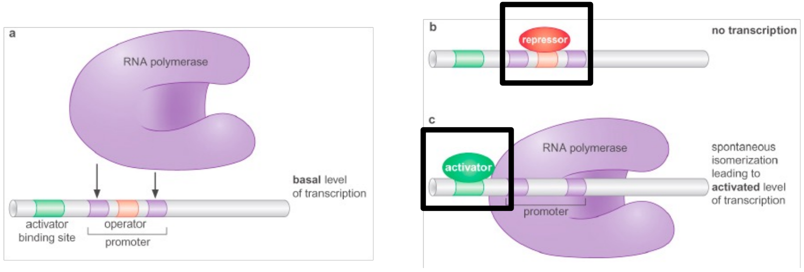


elongation and termination



transcriptional initiation control

- modulation of RNAp-promoter affinity via activators and repressors



➔ net result: rate of transcriptional initiation dependent on cellular conc of activators and repressors controlled by metabolites and signaling molecules

Topic 1: Protein-DNA Interaction

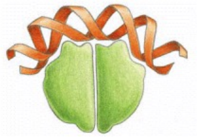
- Goals:
  - find DNA binding target seqs for each transcription factor (TF)
  - find the affinity of a TF to its DNA target as a function of its cellular concentration *in vivo*
  - find how the TF-DNA affinity depends on the target sequence
  - ➔ at what TF conc is each target sequence occupied
- Problems:
  - 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:

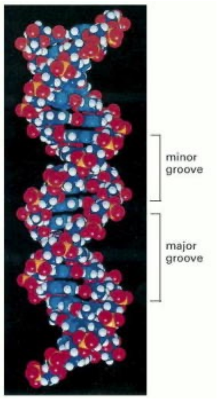
➔ ways to think quantitatively about TF-DNA interaction in the absence of detailed microscopic information

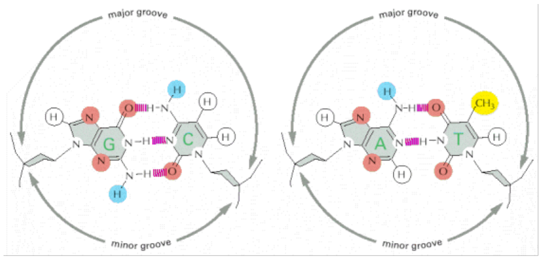
➔ link from molecule to function (an illustrative case)

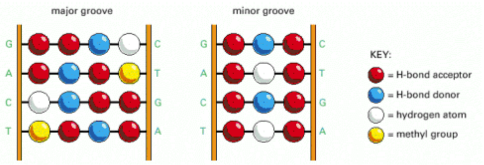
A. Empirical facts



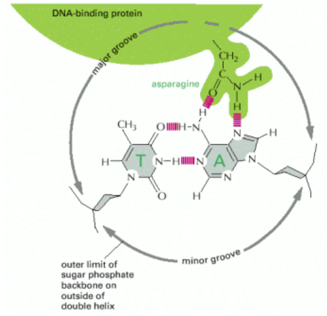
1. Transcription Factors
   - size: ~5nm (10-20 bp)
   - molecular basis of sequence recognition

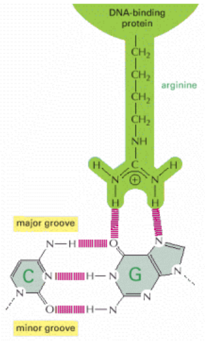






- contact between TF and DNA

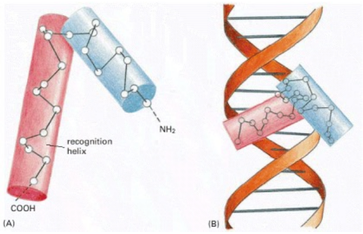


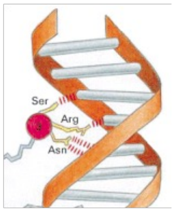


➔ structure of a TF must place the appropriate amino acids next to the base pairs they contact

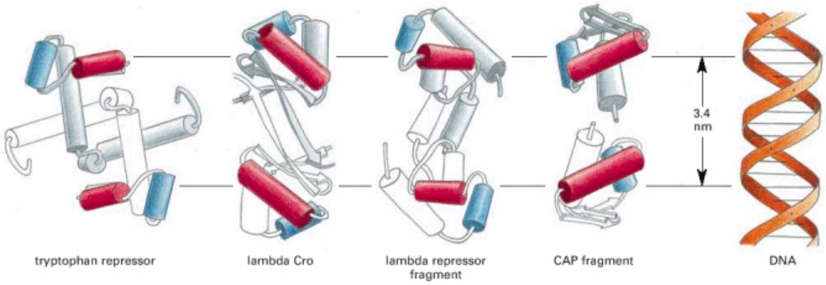
5

- various molecular strategies
  - Helix-Turn-Helix



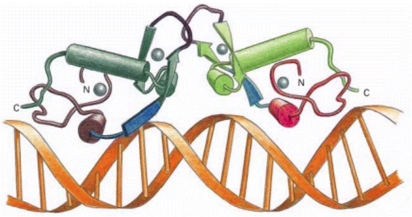


well-known examples in bacteria (note: homodimers)

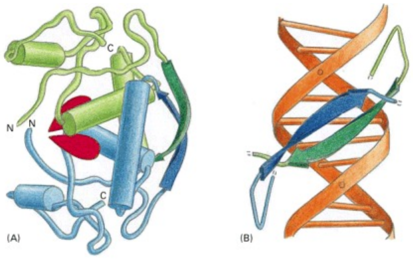


6

– zinc-finger domain

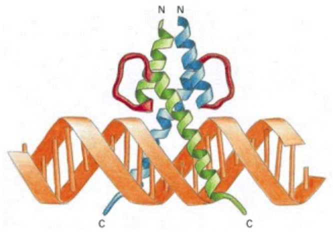


– beta-sheets



– leucine zipper

– helix-loop-helix

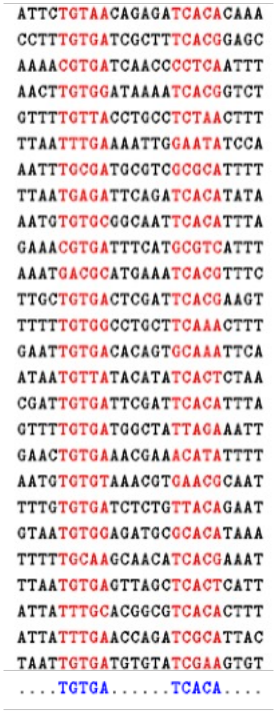


2. DNA binding sequences

- 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 independent of binding seq

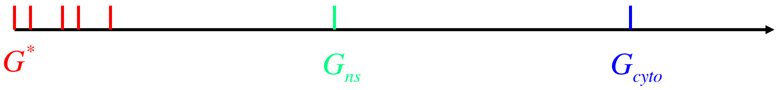
  e.g.$[TF-DNA] > 10 \times [TF]_{free}$ for LacI in 0.1M salt

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

  (kT ≈ 0.62 kcal/mole at 37˚C)

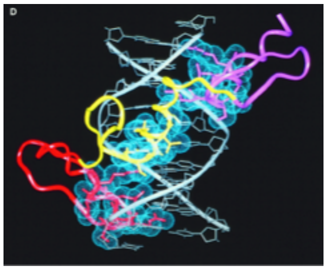
- additional energy gained from hydrogen bonds to preferred sequences

  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



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 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

11

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

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

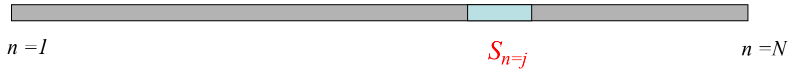
pos.	10	11	12	13	14	15	16	17
A	1.8	2.4	1.6	1.0	0	2.1	0.8	1.1
C	2.4	1.9	4.2	2.1	0.3	0	0	0
G	0	1.6	0	0	1.2	3.2	1.0	1.2
T	3.0	0	2.2	2.2	0.6	2.2	0.7	0.3

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

12

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*)

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

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

- 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]

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

- cf: *in vitro* binding

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

13

- effective *in vivo* binding constant

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

- cf: *in vitro* binding

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

$$\Rightarrow\quad K(S) = \widetilde{K}_j/V_{cell} = \widetilde{K}_j \text{ in nM}$$

– 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

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

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

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

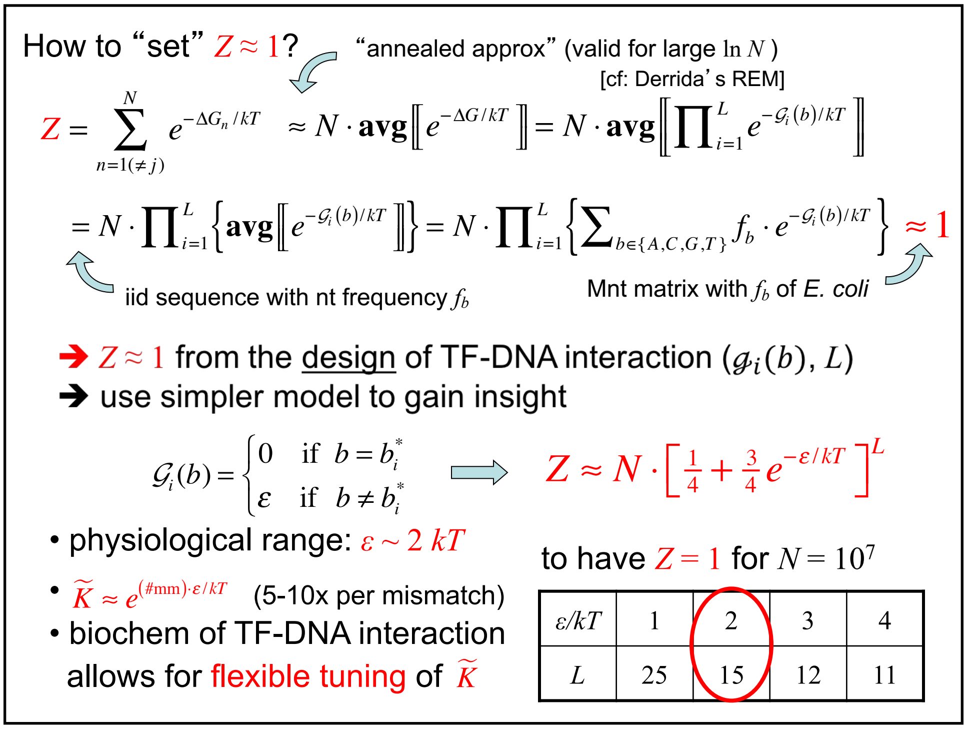
(Mnt matrix applied to *E. coli* genome or *randomly scrambled* genomes)

➔ 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, $K_{O1} \approx 1\text{ pM}$ *in vitro* while $\widetilde{K}_{O1} \approx 3\ n\text{M}$

14



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

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

- physiological range: $\varepsilon \sim 2\ kT$
- $\widetilde{K} \approx e^{(\#\mathrm{mm})\cdot\varepsilon/kT}$ (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

15