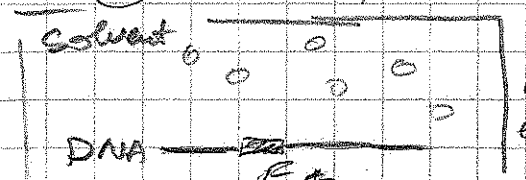
Protein-DNA interactions

① One specific site



$M = \frac{V}{v}$ ← total volume, elementary volume

E ← energy of specific binding

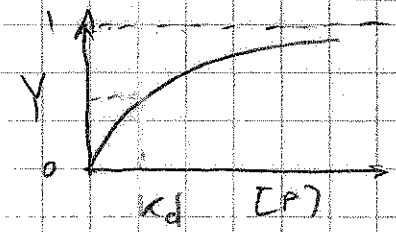
$$\binom{M}{P} = \frac{M!}{P!(M-P)!} \qquad \binom{M}{P-1} = \frac{M!}{(P-1)!(M-P+1)!}$$

$$\frac{\binom{M}{P}}{\binom{M}{P-1}} = \frac{M-P+1}{P} \simeq \frac{M}{P}$$

$$Y = \frac{e^{-E}\binom{M}{P-1}}{e^{-E}\binom{M}{P-1} + \binom{M}{P}} = \frac{e^{-E}}{e^{-E} + \binom{M}{P}/\binom{M}{P-1}} \simeq \frac{e^{-E}}{e^{-E} + \frac{M}{P}}$$

$$= \frac{P/M}{P/M + e^{E}} = \frac{P/V}{P/V + e^{E}/V} = \frac{[P]}{[P] + \underbrace{e^{E}/V}_{K_d}}$$

$$K_d \equiv \frac{e^{E}}{V}$$



② One specific sites, many non-specific sites

E - specific energy

ε - non-specific energy

D - # of non-specific sites on DNA $= \frac{\text{DNA volume}}{v}$

$$Y = \frac{e^{-E}\cdot Z(P-1)}{e^{-E}Z(P-1) + Z(P)}$$

where $Z(P)$ - partition function for P proteins to be in the solvent or on non-specific DNA.

$$Z(P) = \sum_{k=0}^{P} e^{-k\varepsilon}\binom{M-D}{P-k}\binom{D}{k}$$

← proteins on DNA; ← proteins in solvent; ↳ # of proteins on DNA

Note:

$$\binom{A}{b} = \frac{A!}{(A-b)!\,b!} \simeq \frac{A^b}{b!}$$

$b \ll A$

then $\binom{M-D}{P-k}\binom{D}{k} \simeq (M-D)^{P-k}\, D^{k}\cdot \frac{1}{k!(P-k)!} =$

$$= M^{P}\left(1-\frac{D}{M}\right)^{P-k}\left(\frac{D}{M}\right)^{k}\cdot\frac{P!}{k!(P-k)!}\cdot\frac{1}{P!}$$

$$= \left(1-\frac{D}{M}\right)^{P-k}\left(\frac{D}{M}\right)^{k}\binom{P}{k}\cdot\frac{M^{P}}{P!}$$

$$Z(P) \approx \left[\sum_{k=0}^{P} \underbrace{\binom{P}{k} (1-\pi)^{P-k} \pi^k \cdot e^{-\varepsilon k}}_{\text{Binomial prob of putting } k \text{ particles on DNA if } \pi\text{- prob to be on DNA; } \pi \equiv \frac{D}{M}} \right] \frac{M^P}{P!}$$

$$= \left(1 - \pi + \pi e^{-\varepsilon}\right)^P \cdot \frac{M^P}{P!}$$

$$\frac{Z(P)}{Z(P-1)} = \left(1 + \pi\left(e^{-\varepsilon} - 1\right)\right) \frac{M}{P} \quad ; \text{ if } \varepsilon = 0 \text{ (no affinity for non-spec DNA), then } \frac{Z(P)}{Z(P-1)} = \frac{M}{P} \text{ as above.}$$

$$= \frac{M}{P}\left(1 - \frac{D}{M} + \frac{D}{M} e^{-\varepsilon}\right)$$

$$= \frac{M-D}{P} + \frac{D}{P} e^{-\varepsilon}$$

$$Y = \frac{e^{-E}}{e^{-E} + \frac{Z(P)}{Z(P-1)}} = \frac{e^{-E}}{e^{-E} + \underbrace{\frac{M-D}{P}}_{\substack{\#\text{ of states in solvent} \\ \text{per protein}}} + \underbrace{\frac{D}{P}}_{\substack{\#\text{ of states on DNA} \\ \text{per protein}}} e^{-\varepsilon}}$$

$$= \frac{P/(M-D)}{P/(M-D) + e^{E}\left(1 + \frac{D}{M-D} \cdot e^{-\varepsilon}\right)}$$

recall

$$M = \frac{V}{v} \; ; \; D = \frac{\text{DNA vol}}{v}$$

$$\Rightarrow M - D = \frac{\text{Solvent Volume}}{v}$$

$$= \frac{[P]v}{[P]v + e^{E}\left(1 + [DNA] \cdot v \cdot e^{-\varepsilon}\right)}$$

where $[P] \equiv \dfrac{P}{\text{solvent volume}}$; $[DNA] = \dfrac{D}{\text{solvent volume}}$

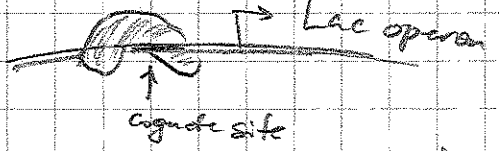
$$= \frac{[P]}{[P] + e^{E}/v\left(1 + \frac{[DNA]}{e^{\varepsilon}/v}\right)} = \frac{[P]}{[P] + \underbrace{K_d\left(1 + \frac{[DNA]}{K_d^{ns}}\right)}_{K_d^{eff}}}$$

K_d^{eff} - higher than K_d due to non-spec. Binding

$$K_d \equiv \frac{e^{E}}{v} \; , \; K_d^{ns} \equiv \frac{e^{\varepsilon}}{v}$$

Example: Lac repressor

Function:
repress Lac operon



Lactose	Lac repressor	Lac operon	K_d
+	unbound	expressed	10^{-9} M
−	bound	repressed	10^{-12} M

[P] = 10 molecules per E.coli cell $\simeq 10 \cdot 10^{-9}$ M

① Naive approach, i.e. disregarding non-specific binding

$$Y^{-lac} = \frac{[P]}{[P]+k_d} = \frac{10\cdot10^{-9}}{10\cdot10^{-9}+10^{-12}} = \frac{1}{1+10^{-3}} = 0.999\ldots \quad \underline{\text{bound!}}$$

$$2\,k_d^{-lac} = 10^{-12}\,M$$

$$Y^{+lac} = \frac{10\cdot10^{-9}}{10\cdot10^{-9}+10^{-3}} = \frac{10}{11} = 0.9 \quad \underline{\text{also bound, but shouldn't}}$$

② Take into account non-specific Binding to DNA

$K_d^{ns} \simeq 10^{-6}$ M irrespective of lactose

[DNA] = $5\cdot10^6$ bp / cell / ~10 bp footprint

$$\left(1+\frac{[DNA]}{K_d^{ns}}\right) = 1+\frac{5\cdot10^6\cdot10^{-9}}{10^{-6}\cdot10} = 1+500 \simeq 500$$

$$Y^{-lac} = \frac{10^{-8}}{10^{-8}+10^{-12}\cdot500} = 0.95 \quad \underline{\text{bound}}$$

$$Y^{+lac} = \frac{10^{-8}}{10^{-8}+10^{-9}\cdot563} \simeq 10^{-2} \quad \underline{\text{non-bound}}$$

⊛ Model of binding : PWM : position weight matrix

(1) Assume energy is sum of contributions of bps.

$$E_{\alpha\ell} = \sum_{i=1}^{L} \epsilon_{i\alpha_i} \quad ; \text{ where } \alpha_i = A, T, G, C, 2, 3, \ldots \quad i = 1 \ldots L$$

base-pair sequence of DNA in the site of length L

$\epsilon_{i\alpha}$ - $L \times 4$ matrix of energy (aka PWM)

Then

$$Z = \prod_{i=1}^{L} \sum_{\alpha=1}^{4} \left[e^{-\beta \epsilon_{i\alpha}} \cdot p_0(\alpha) \right]$$

$p_0(\alpha)$ ← background prob. of bp α

Σ for individual positions

$$p_\alpha(i) = \frac{e^{-\beta \epsilon_{i\alpha}} \cdot p_0(\alpha)}{\sum_{\alpha'} e^{-\beta \epsilon_{i\alpha'}} \cdot p_0(\alpha')}$$

← freq. in SELEX experiment

Const

$$\beta \tilde{\epsilon}_{i\alpha} = -\log \frac{p_i(\alpha)}{p_0(\alpha)} + \text{const} ;$$

Interestingly mean energy : $\langle E \rangle = -\sum_{i,\alpha} p_\alpha(i) \log \frac{p_i(\alpha)}{p_0(\alpha)}$

$= -I$ information content of the motif

(if $p_0(\alpha) = 1/4$)

(2) Simple model of specific binding

$$\epsilon_{i\alpha} = \begin{cases} 0 & \text{for } \alpha_i^* \text{ in native site} \\ \epsilon > 0 & \text{otherwise} \end{cases}$$

then we can ask what value of ϵ is sufficient for specific recognition of the native site of length L

$$Z = N \cdot \sum_{m=0}^{L} \binom{L}{m} \left(\frac{1}{4}\right)^{L-m} \left(\frac{3}{4}\right)^{m} e^{-\beta m \epsilon}$$

Σ m mismatches

$$\beta F_{genom} = -\log Z \qquad F_{native} = E - TS = 0 = 0$$

$\Sigma = 0$

Need $\beta F_{genome} \geq 0 \Rightarrow Z \leq 1$

$$Z_1 = N\left(\frac{1}{4} + \frac{3}{4}e^{-\beta\epsilon}\right)^L \leq 1$$

$$\log_2(1 + 3e^{-\beta\epsilon}) \leq \frac{\log_2 N}{L} + 2$$

Bacteria $N = 10^7$; $L = 12$ $\quad \epsilon \geq 1.6\ kT$

mammals $N = 10^9$ $\quad L = 6$ $\quad \epsilon \geq 4\ kT$! impossible!

⊛ Comparing to non-specific DNA

$$Z_{sp} = N\left(\frac{1}{4} + \frac{3}{4}e^{-\beta\epsilon}\right)^L \leq Z_{ns} = N \cdot e^{-\beta E_{ns}} \leq 1$$

$$\frac{E_{ns}}{L} = 1\ kT$$

⊛ Information-theoretic argument

Consider "motif" given by the frequencies $p_i(\alpha)$

The information content of each base-pair i is

$$I_i = -\sum_\alpha p_i(\alpha) \log_2 \frac{p_i(\alpha)}{p_0(\alpha)} \quad ; \text{ (in Bits)}$$

Example: 1) $p_0(\alpha) = 1/4$ for $\alpha = A, T, G, C$; $p_0(A) = 1$, $p_i(G) = p_i(C) = p_i(T) = 0$

then $I_i = 2$ bits

2) if $p_i(A) = p_i(G) = 1/2$, then $I_i = 1$ bit of information

Total information content of motif

$$I = \sum_{i=1}^{L} I_i = \sum_{i=1}^{L} \sum_{\alpha=1}^{4} p_i(\alpha) \log_2 \frac{p_i(\alpha)}{p_0(\alpha)} = \langle E \rangle \text{ (see above)}$$

computed

$$\epsilon_{i\alpha} = -\log_2 \frac{p_i(\alpha)}{p_0(\alpha)}$$

The meaning of the information content:

To make a choice out of two alternatives you need 1 bit of information;
out of 4 alternatives: 2 bits,
out of N you need $\log_2 N$ bits.

For a protein to find it site out of N alternatives in the genome of N bp, the protein needs a "motif" with $I_{min} = \log_2 N$

Bacteria: $N \approx 5 \cdot 10^6 - 10^7$ $I_{min} \approx 23$ Bits

Observed for bacterial proteins $I = 23$ bits!
(Wunderlich & Mirny, 2009)

Human: $N = 3 \cdot 10^9$ $I_{min} \approx 31$ bit

$I_{observed} \simeq 12$ bits $\ll I_{min}$!

Even when only accessible 1% of DNA is considered
$N = 3 \cdot 10^7$ $I_{min} = 25$ bits $\gg 12$ bits!

Sufficient information is provided by 2-3 sites next to each other

2 sites of the same or different proteins.

=> "combinatorial regulation"
a single binding of protein to DNA is not sufficient as it happens ~ every $2^I \approx 4000$ pb
($I = 12$ bits)