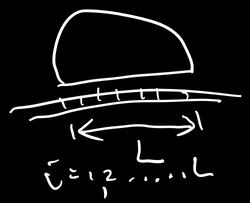
• Protein - DNA interactions

– Functional: programmable : K_d ; K_d^{ns}

– Specificity : $E = \sum_{i=1}^{L} \epsilon(i, b_i)$

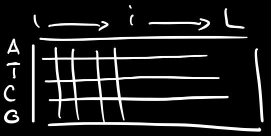


↖ (a) inferred from experimental data

(b) genome-wide binding

– Kinetics

○ Inference of $\epsilon(i,b)$



from the data

↳ collection of sequences bound by the protein

Berg & von Hippel

• if positions were independent:

$$f(i,b) = e^{-\epsilon(i,b)} / \sum_a e^{-\epsilon(i,a)}$$

↗ frequency of b @ position i

$$\Rightarrow \epsilon(i,b) = -\log f^{obs}(i,b) + \text{Const}$$

$$= + \log \frac{f^{obs}(i,\emptyset)}{f^{obs}(i,b)}$$

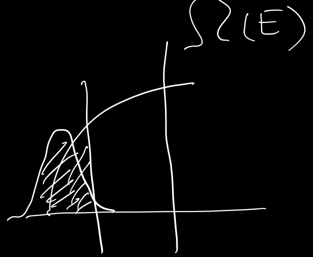
the most frequent at position i

○ non independence of contributions from positions

$$f(i,b) = e^{-\lambda\epsilon(i,b)} / \sum_a e^{-\lambda\epsilon(i,a)}$$

$$\bar{E}^{obs}[\text{selected sequences}] = \bar{E} \approx ?$$

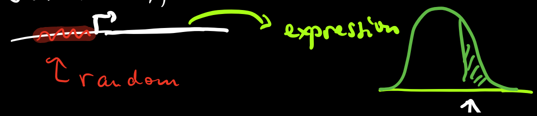
$$\underline{\lambda\epsilon(i,b)} = \log \frac{f^{obs}(i,\emptyset)}{f^{obs}(i,b)}$$



○ 2006 and later

Inference of $\underline{\epsilon(i,b)}$ to max likelihood of observed sequences

Note: Experiments :



Berg and von Hippel
Selection of DNA binding sites by regulatory proteins.
Statistical-mechanical theory and application to operators and promoters
https://pubmed.ncbi.nlm.nih.gov/3612791/

Physical constraints and functional characteristics of transcription factor–DNA interaction
https://www.pnas.org/doi/10.1073/pnas.192693599

(B) Binding in the genome-wide context

- genome is random

$$p(A)=p(T)=p(C)=p(G)=1/4$$

- M - length of genome
- Bacteria
- NO non-specific Binding
- cognate site S^* ← minimal energy $E(s^*)=0$

$$\epsilon_{ib} = \begin{array}{c}A\\T\\C\\G\end{array}\left[\begin{array}{cccccc} 0 & \varepsilon & \varepsilon & & \varepsilon & \\ \varepsilon & 0 & 0 & & & \\ \varepsilon & \varepsilon & \varepsilon & 0 & & \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & 0 & 0 \end{array}\right] \quad ; \quad \varepsilon > 0$$

$$\xleftrightarrow{L}$$

$$S^* = \{ATTCGG\}$$

→ Probability of S^* occupied

$$P(s^*) = \frac{Z_*}{Z_* + Z_{sp}}$$

$$Z_* = e^{-E(s^*)} \cdot 1 = 1$$

$$Z_{sp} = M \cdot \underbrace{\sum_{k=0}^{L} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{L-k} \binom{L}{k}}_{\text{k-position mutants}} \cdot e^{-\varepsilon \cdot k}$$

← E.g. $k=0$

$$M \cdot \left(\frac{1}{4}\right)^L$$ ← exact matches of S^* in the random genome

$$= M\left(\frac{1}{4} + \frac{3}{4}e^{-\epsilon}\right)^L$$

under what conditions

$$Z_{sp} \le 1$$

$$M\left(\frac{1}{4} + \frac{3}{4}e^{-\varepsilon}\right)^L \le 1$$

E.g. $\varepsilon = 0 \Rightarrow P(s^*) = \frac{1}{M}$

$$Z_{sp} = M \gg 1$$

$$\varepsilon \to \infty$$

$$Z_{sp} = M \frac{1}{4^L} \qquad \underline{Z_* = 1}$$

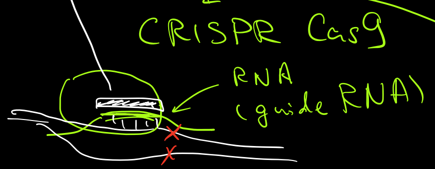
has to be longer than $\underline{17bp}$ ≈ 20bp used in practice

$$M / 4^L \le 1 \qquad 4^L \ge M$$

$$\boxed{L_{crit} \ge \frac{\log_2 M}{2}}$$

For bacteria:
$M = 5 \cdot 10^6$; $\log_2 M = 22$ Bits; $\boxed{L_{crit} = 11 \text{ bp}}$

For human
$M = 3 \cdot 10^9$; $\log_2 = 32$ Bits; $\boxed{L_{crit} = 16 \text{ bp}}$



$\epsilon \simeq 6kT$

* Estimate ε for Bacteria $L = 15-20$ bp $> L_{crit}$

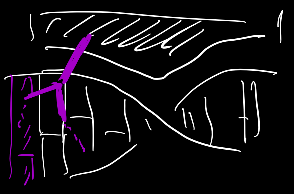
$$\frac{M}{4^L}\left(1+3e^{-\varepsilon}\right)^L \le 1 \qquad 1+3e^{-\varepsilon} \le 4M^{-1/L}$$

$$\varepsilon \geq \log\left[\frac{4M^{-1/4}-1}{3}\right]$$

$$\boxed{\varepsilon \geq \underline{2kT}}$$ $L = 15$ Bp

H-Bond $\underline{1-2}$

or small surface for hydropobic interactions



* How about non-specific Binding

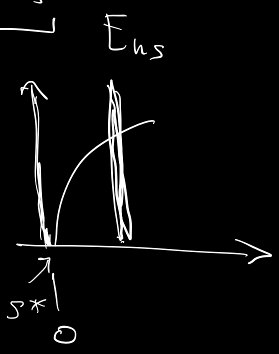
$$E_{ns} = \varepsilon_{ns}\cdot L \geq 0 \qquad \underline{P}(s^*) = \frac{Z_*}{Z_* + \underbrace{Z_{sp} + Z_{ns}}}$$

$$Z_{ns} \le 1$$

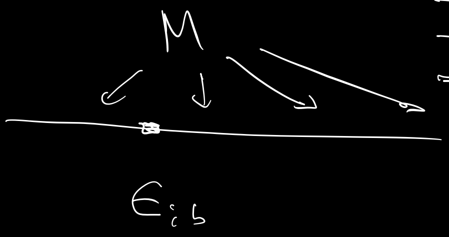
$$Z_{ns} = M\cdot e^{-E_{ns}} = M\cdot e^{-\varepsilon_{ns}\cdot L} \le 1$$

$$\varepsilon_{ns} \geq \frac{\log M}{L}$$; for Bacteria $L = 15$ Bp

$$\boxed{\varepsilon_{ns} \geq 1kT}$$



* Information theory approach



$$\underline{I_{required}} = \log_2 M \begin{cases} \to \text{bacteria: } 22 \text{ Bits} \\ \to \text{human } 32 \text{ Bits} \end{cases}$$

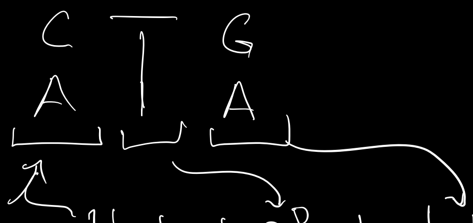
$\epsilon \to \infty$

$$\overbrace{\underbrace{A}\ \underbrace{T}\ \underbrace{G}\ \underbrace{G}}^{\text{motif}} \qquad \underline{\underline{I}} = 2\cdot L$$

2 bits 2 bits $\xrightarrow{\hspace{2em}}$ L

$$2L \geq \log_2 M \; ; \; \boxed{L \geq \frac{\log_2 M}{2}}$$

~~t = finite~~



$I = 1\,\text{bit} + 2\,\text{Bit} + 1\,\text{bit} = 4\,\text{Bit}$

$I^{\text{experimental motifs}}$ $\underline{\text{to}}$ I required

Different gene regulation strategies revealed by analysis of binding motifs
https://pubmed.ncbi.nlm.nih.gov/19815308/