Last time

TF binding to target site S_0

- Probability

$$f(S_0) = \frac{1}{1 + \tilde{K}_0 / [T]_{tot}}$$

$\tilde{K}_0$ effective dissociation constant in the presence of genomic DNA

- Assuming all TFs associate to DNA and # TF molecs $\ll N$ and fast Eq.

$$\tilde{K}_0 = \sum_j K_0 / K_j = \sum_j e^{(G_0 - G_j)/kT}$$

- Two modes of binding (specific vs. non-specific)

$$e^{-G_j/kT} = e^{-G_j^{sp}/kT} + e^{-G^{ns}/kT}$$

$g_i(b_i^*) = 0$

$$G_j^{sp} = G^* + \sum_i^L g_i(b_i^{(j)})$$

$$\tilde{K}_0 = \tilde{K}^* \left(e^{\sum_i^L g_i(b_i^{(0)})/kT} \right)$$

$$\tilde{K}^* = N e^{-(G^{ns} - G^*)/kT} + \sum e^{-\sum_i^L g_i(b_i^{(j)})/kT}$$

K^* depends on protein design $(G^{ns} - G^*, g_i(b))$ and genomic sequence

Desirable Features

Specificity : $K^* \leq 1$ u M

$\tilde{K}_0$ between $1 \sim 1000$ u M

Tunability : choice of target sequence alone should determine whether TF binds
some TF may bind to different sites depending on TF conc.

Independence : no interference between target sites

⇒ design constraints on parameters ?

Consider simple model of energy matrix $g_i(b)$

$$g_i(b) = \begin{cases} 0 & \text{if } b_i = b_i^* \\ \varepsilon > 0 & \text{else} \end{cases}$$

$S^* = \{G\,C\,T \ldots\}$

	1	2	3	...
A	ε	ε	ε	
C	ε	0	ε	
G	0	ε	ε	
T	ε	ε	0	

$$\tilde{K}_0 = \tilde{K}^* \cdot e^{m(S_0)\,\varepsilon/kT}$$

$m(S_0)$ = # mismatches from S^*

For $m = 0 \sim 4$ ε needs to be $\sim 2\,kT$

so that $e^{m\varepsilon/kT} = 1 \sim 1000$

$$\tilde{K}^* = N e^{-(G^{us} - G^*)/kT} + N \xi^L(\varepsilon)$$

For $\tilde{K}^* \leq 1$ we must have

$N \xi^L(\varepsilon) \leq 1$ and $N e^{-(G^{us} - G^*)/kT} \leq 1$

For $\varepsilon > 0$ $\xi(\varepsilon) < 1$

$\xi(\varepsilon)$ smallest for $\varepsilon = \infty$ (perfect discrimination)

$$\xi(\infty) = \frac{1}{4} \rightarrow \frac{N}{4^L} \leq 1$$

condition for unique L-mer in random genome sequence of length N

Smaller ε: weakens specificity but improves tenability of binding affinity

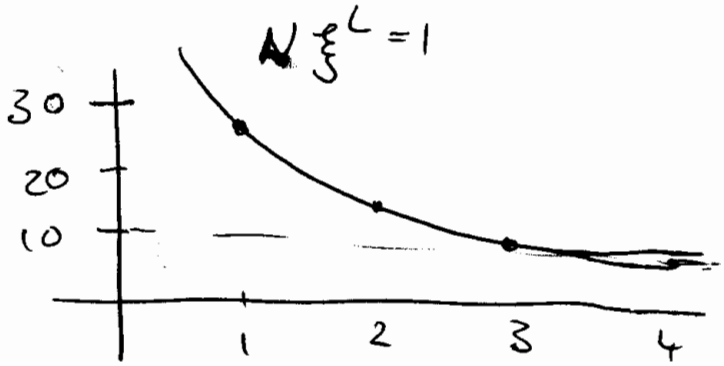
For tenability we want smallest ε with $N \xi^L(\varepsilon) \leq 1$

$$N \xi^L(\varepsilon) \sim 1$$

$$L \approx \ln\left(\frac{1}{4} + \frac{3}{4} e^{-\varepsilon/kT}\right) / \ln N$$

for $N = 5 \times 10^5$

ε	1	2	3	... ∞
L	24.5	15	12.5	11



The smaller your alphabet size, the longer words have to be

infinite ε: perfect alphabet, smaller ε: smaller effective alphabet size

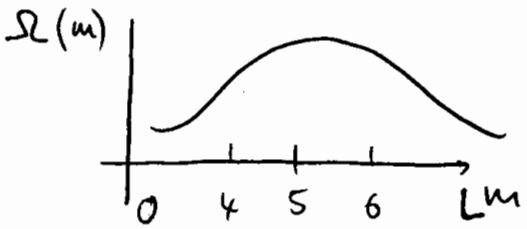
What about K^* ?

Specific + Non-Specific contribution

$$\sum_j g \; e^{-\sum g_i(b_i^{(j)})/kT} = \underbrace{\sum_{j=1}^{N} e^{-m_j \varepsilon/kT}}_{Z}$$

m_j # mismatches between S_j and S^*

Z depends on genomic sequence



Histogram : # segments with m mismatches

$$Z = \sum_m \Omega(m) \, e^{-m\varepsilon/kT}$$

Assume $\Omega(m)$ is binomially distributed (each binding seq S_j is random)

L rolls of 4-sided die

$$B(m) = \frac{L!}{m!(L-m)!} \left(\frac{3}{4}\right)^m \left(\frac{1}{4}\right)^{L-m}$$

For sufficiently long genome seq. we can approximate

$$\Omega(m) \cong N \, B(m) \quad \text{(annealed appr.)}$$

$$Z \cong \sum_{m=0}^{L} N \frac{L!}{m!(L-m)!} \left(\frac{3}{4}\right)^m \left(\frac{1}{4}\right)^{L-m} e^{-m\varepsilon/kT}$$

$$= N \left[\frac{1}{4} + \frac{3}{4} e^{-\varepsilon/kT}\right]^L = N \, \xi^L(\varepsilon)$$

$$\xi(\varepsilon) = \sum_b f_b \, e^{-g(b)/kT}$$

$$= \frac{1}{4} + \frac{3}{4} e^{-\varepsilon/kT}$$

For typical discrimination energy $\varepsilon \sim 2kT$
(hydrogen bonds →)

one gets $L = 10 \sim 20$ bp (size of typical TF binding site in bacteria)

Bacteria seem to respect

$$N \xi^L(\varepsilon) \sim 1$$

Non specific binding

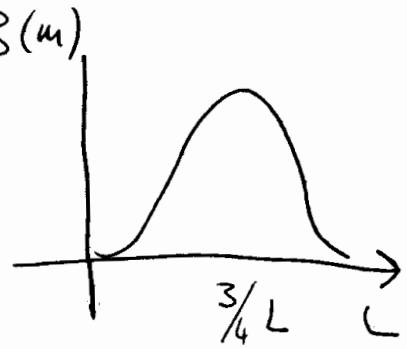
$$N e^{-(G^{us} - G^*)/kT} \leq 1$$

What might prevent $G^{us} - G^*$ from being arbitrarily large?

Recall $e^{-G_j/kT} = e^{-G^{us}/kT} + e^{-G_j^{sp}/kT}$

If G^{us} is very large

then $G_j \cong G_j^{sp} = m_j \varepsilon$

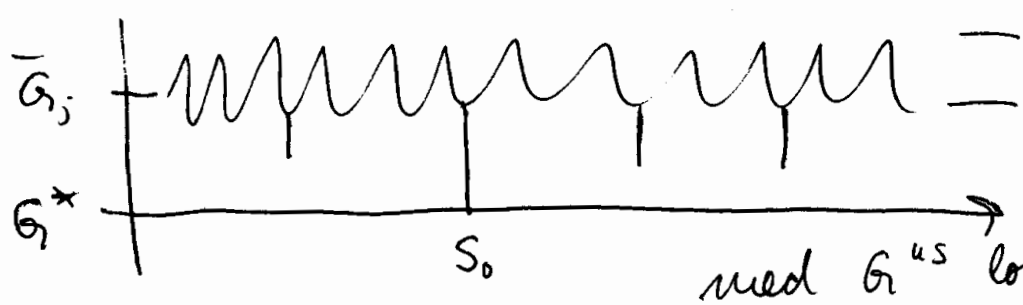


Only variation through m

$$\text{mean}(m) = \frac{3}{4} L$$

$$\text{var}(m) = \frac{1}{4} \cdot \frac{3}{4} L$$

For $L = 10$, $\varepsilon = 2$ $\quad G_1 = 15 \pm 3kT$



Problem: spurious traps

need G^{us} low enough for good search kinetics