Protein examples

Jun/Fos https://www.rcsb.org/3d-view/1FOS
LacI - whole complex
https://www.rcsb.org/3d-view/1EFA
DNA-binding domain bound to non-specific DNA
https://www.rcsb.org/3d-view/1OSL
DNA-binding domain bound to specific DNA
https://www.rcsb.org/3d-view/1L1M

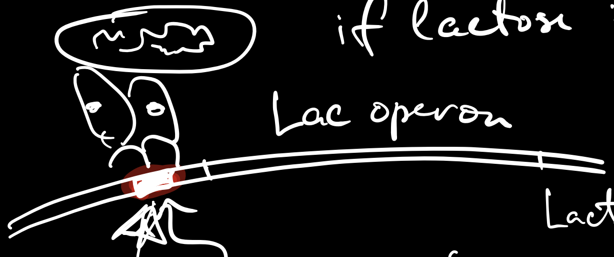
Protein-DNA interactions

1. For regulation ⇒ ...
2. Specific recognition : estimates
3. Kinetics: protein-DNA search

1. Lac repressor [LacI] regulates Lac operon

E. coli glucose - prefered

if lactose is present ⇒ Lac operon active



Binding site of LacI

Lac operon K_d

Lactose	Binding Site	Lac operon	K_d
+	not bound	+	10^{-9} M
−	bound +	−	10^{-12} M

occupancy of the binding site on DNA

$$Y = \frac{[P\cdot S]}{[P\cdot S]+[S]}$$

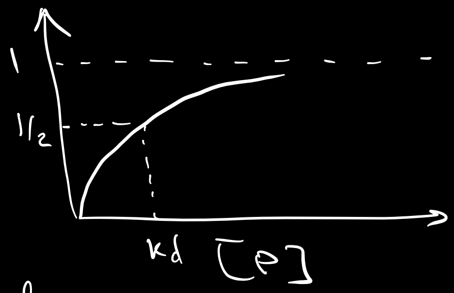
occupance

$$P + S \overset{K_d}{\rightleftharpoons} P\cdot S \;;\quad K_d = \frac{1}{v} e^{-\beta E_b} [M]$$

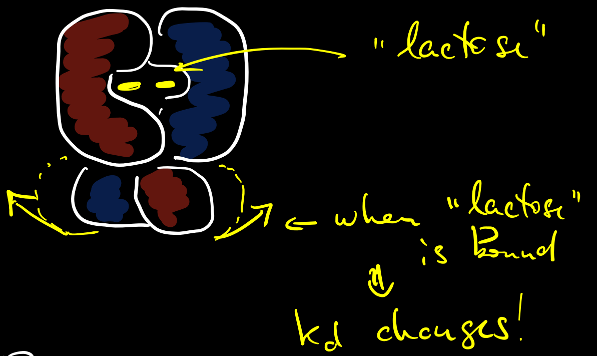
standard volume

$$K_d = \frac{[P][S]}{[P\cdot S]}$$

$$Y = \frac{[P\cdot S]/[P][S]}{[P\cdot S]/[P][S] + 1/[P]} = \frac{1/K_d}{1/K_d + 1/[P]} = \frac{[P]}{[P]+K_d}$$



Kd ⟵ depends on lactose



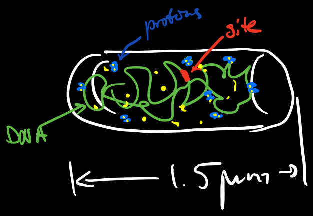
$K_d^{-lac} = 10^{-12}$ M

$K_d^{+lac} = 10^{-9}$ M

⇒ let's compute $Y^{\pm lac}$

[P] ≃ 10 molecules per Bacteria

↖ "dimers" (very stable!)



1μm $V \simeq 1.5\ \mu m^3$

Human cells
nucleus ~500 μm³

$1\ \text{Mole} = 1\ \frac{\text{molecule}}{\text{liter}}$

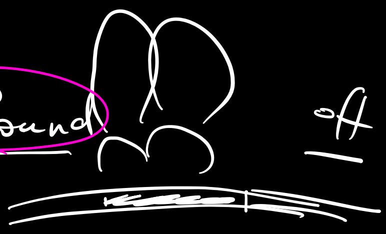
$\frac{1\ \text{molecule}}{1.5\ \mu m^3} \longrightarrow \frac{1\ \text{mol}}{\text{liter}}$

↗
$10^{15}\ \mu m^3$

$$\boxed{\frac{1\ \text{molecule}}{1.5\ \mu m^3} = 10^{15} / \underbrace{1.5\cdot 6}_{10}\cdot 10^{23} = 10^{-9}\,M}$$

$$[P] = 10\ \frac{\text{molecules}}{\text{Bacteria}} = 10^{-8}\,M$$

$$Y^{-lac} = \frac{10^{-8}}{10^{-8}+10^{-12}} \approx 1$$ ⟵ Bound



$$Y^{+lac} = \frac{10^{-8}}{10^{-8}+10^{-9}} \simeq \frac{10}{11} \simeq 0.9$$ ⟵ Bound

⟶ should be unbound

— protein .. dimer?

— kinetics ... ← fast ...

— DNA (non-specific DNA)

$$P + DNA \rightleftharpoons P.DNA$$

$$K_d^{ns} = \frac{1}{v} e^{-\beta E_{ns}} \;; \quad K_d^{ns} \approx 10^{-5}\,M$$

$$K_d^{ns} = \frac{[P][DNA]}{[P.DNA]} \qquad \underbrace{[P]_{TOT} = [P] + [P.DNA]}_{} + \underbrace{[\cancel{P.S}]}_{0}$$

$[P]$ - ?

$$K_d^{ns}[P.DNA] = [P][DNA]$$

$$\frac{K_d^{ns}}{[DNA]}\left([P]_{TOT} - [P]\right) = [P]$$

$$[P] = [P]_{TOT} \Big/ \left(1 + \frac{[DNA]}{K_d^{ns}}\right)$$

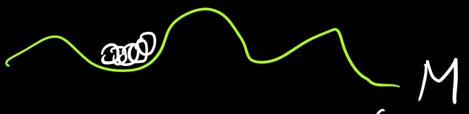
non specific

$$Y = \frac{[P]}{[P] + K_d} = \frac{[P]_{TOT}}{[P]_{TOT} + \underbrace{\boxed{K_d}\left(1 + \boxed{\frac{[DNA]}{K_d^{ns}}}\right)}_{K_d^{eff}}}$$

↗ occupancy of the site

$$\left.\begin{matrix} Y^{-lac} \\ \\ Y^{+lac} \end{matrix}\right\} \quad K_d^{ns} = 10^{-6}\,M$$

$$[DNA] = \; = 5\cdot 10^{6} \cdot 10^{-9} = 5\cdot 10^{-3}\,M$$



$M = 5\cdot 10^{6}$ Bp
↑ sites for ns Binding

$$\left(1 + \frac{[DNA]}{K_d^{ns}}\right) = 1 + \frac{5\cdot 10^{-3}}{10^{-5}} \approx 500$$

$$Y_{-\text{lac}} = \frac{10^{-8}}{10^{-8} + 10^{-12}\cdot 500} = \frac{10^{-8}}{10^{-8} + 10^{-10}\cdot 5} \simeq 1 \quad \boxed{\text{Bound!}}$$

$$Y_{+\text{lac}} = \frac{10^{-8}}{10^{-8} + 10^{-9}\cdot 500} = \frac{10}{10 + 500} = \underline{\underline{0.02}}$$

not Bound !

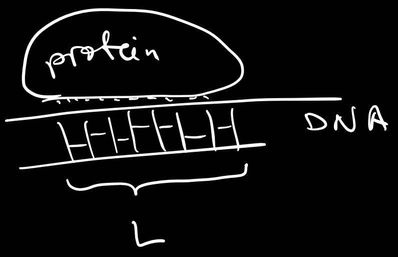
✓

- Non-specific DNA plays an important role! in making this system "programmable"
- Protein-DNA interactions ⇒ logic of gene expression

Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell
Johan Elf,Gene-Wei Li,and X. Sunney Xi
Science 2007
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2853898/

---

② Specificity



$$E = \sum_{i=1}^{L} \varepsilon(i, b_i)$$

base pair at position i

$\varepsilon(i, x)$

$L \times 4$

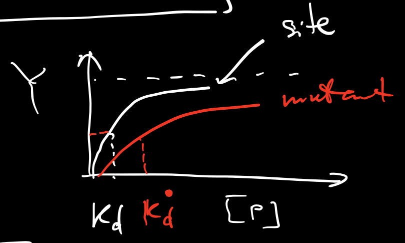
$$\begin{array}{c|cccc}
 & 1 & 2 & 3 & \\
\hline
A & -1 & 0 & \cdot & \\
T & 3 & 0 & \cdot & \\
C & 1 & -1 & \cdot & \\
G & -2 & 1 & \cdot &
\end{array}$$

1. How to learn $\varepsilon(i,x)$?
2. How does recognition work?

1. Inferring $\varepsilon(i,x)$ → measure in vitro

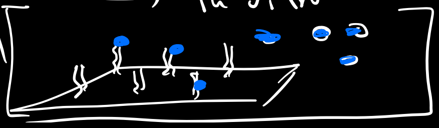
⇒ difficult

⇒ in vitro



L×4 experiments

- In vitro PBM

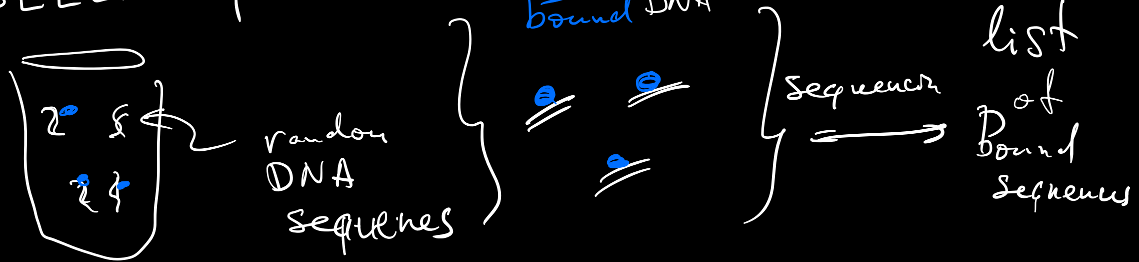


Protein abundance on sequenes of the array [M. Bulyk G. Church]

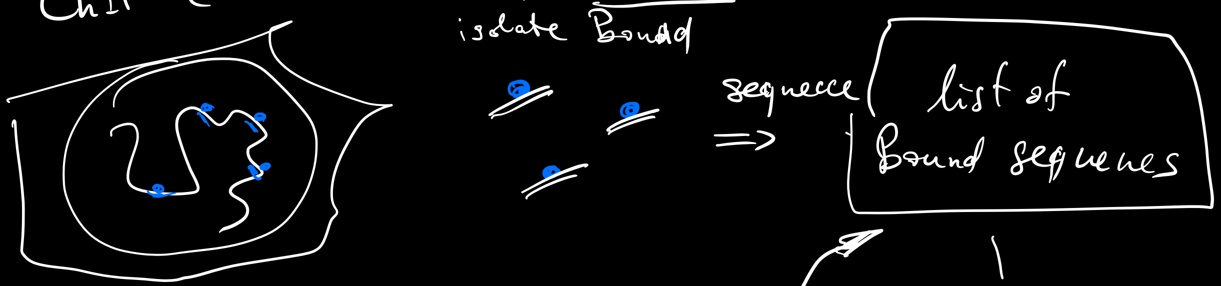
⇒ list of Bound sequences

In vitro

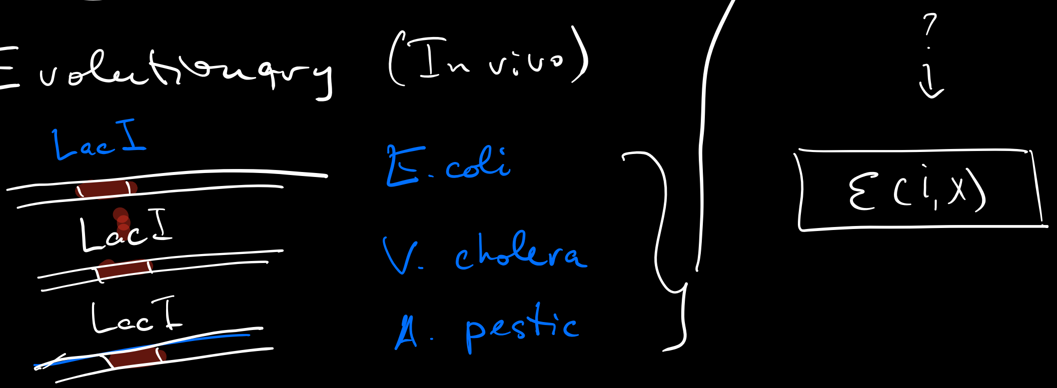
- SELEX seq



- ChIP (Cut-and-Run) In vivo



- Evolutionary (In vivo)



1 Infer $\varepsilon(i,x)$ from know Bound sequences

$f(i,x) \leftarrow$ frequency of x at i

ATTCGGC
ATCGGCC
TTCGCCC

↑

$f(1,A) = 2/3$

Boltzman equilibrium in the evolution of sites

$$P_i(x) \approx e^{-\beta\varepsilon(i,x)} \cdot \underset{\sim}{P_0(x)} / Z$$ background in the genome

$$\beta \mathcal{E}(i,x) = -\log \frac{P_i(x)}{P_0(x)} + \text{const}$$

$$\boxed{\beta \mathcal{E}(i,x) = -\log \frac{f(i,x)}{P_0(x)}} \quad \cancel{\text{const}}$$

observed freq in the genom

1989
Berg & von Hippel

2. (2006 ...) H. Busemaker (Columbia Univ)
Justin Kinney (Princeton)

$$L(\text{seq}, \mathcal{E}(i,x)) \xrightarrow[\boxed{\{\mathcal{E}(i,x)\}}]{} \max \qquad \text{MCMC}$$

observed sequences

$$\Longrightarrow \mathcal{E}(i,x) \text{ matrices for lots of proteins}$$

for Bacteria / Yeast

in vitro ≃ in vivo

→ to find the site
out of $10^6 - 10^9$ alternatives
→ kinetics

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

Inference methods:
1. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE
https://pubmed.ncbi.nlm.nih.gov/16873464/
2. Precise physical models of protein-DNA interaction from high-throughput data
https://pubmed.ncbi.nlm.nih.gov/17197415/