Statistical Genetics

Trascriptor factors
Transcriptional regulation

1. RNA messenger
2. DNA
3. Polymerase
4. Transcription Factor

A: Enhancer

B: Reducer
Binding sites

- Transcription factors bind to **DNA target sites**.

- Target sites have a more **specific sequence** than background DNA.

- Binding sites are short: typically segments of 10 to 15 base pairs in prokaryotes and even shorter segments in eukaryotes.

- Interaction TF – BS: electrostatic interaction between the positively charged protein and the negatively charged DNA backbone.
Biophysics of the Binding Energy

- $E$ depends on $l$ consecutive nucleotides
  
  \[ a = (a_1, ..., a_l) \]

- The single nucleotides of a binding locus give approximately independent contribution to the binding energy
  
  \[ E(a) = \sum_{i=1}^{l} \varepsilon_i(a_i) \]

- At each position there is a preferred nucleotide
  
  \[ a_i^* \text{ with } \varepsilon_i(a_i^*) = \min_a \varepsilon_i(a) \]

- Mis-match energy cost
  
  \[ \varepsilon_i(a) - \varepsilon_i(a_i^*) \approx 1 - 3 \ k_B T \]

- Energy difference bound-unbound
  
  \[ E_u - E^* \approx 15 \ k_B T \]

LAC-I repressor of E-coli
Two state approximation

- Use the match-mismatch information only

\[ \varepsilon_i(a) - \varepsilon_i(a_i^*) = \begin{cases} 
\varepsilon & \text{if } a_i \neq a_i^* \\
0 & \text{if } a_i = a_i^* 
\end{cases} \]

with \( \varepsilon \approx 2k_B T \)

- Binding energy is then related to the hamming distance

\[ E(a) = E^* + \varepsilon \cdot d(a, a^*) \]
Inference of the fitness function
Substitution dynamics in sequence space

- sequence segments of moderate length $\ell$, such that $\mu\ell N \ll 1$.

- neutral rates

$$\frac{\mu_{a \rightarrow b}}{\mu_{b \rightarrow a}} = \frac{P_0(b)}{P_0(a)},$$

- detailed balance in an arbitrary fitness landscape,

$$\frac{v_{a \rightarrow b}}{v_{b \rightarrow a}} = \frac{\mu_{a \rightarrow b}}{\mu_{b \rightarrow a}} \frac{\pi(N, s_{ab})}{\pi(N, s_{ba})} = \frac{P_0(b)}{P_0(a)} \exp[2N (F_b - F_a)] = \frac{P_{eq}(b)}{P_{eq}(a)},$$

- equilibrium distribution of fixed states is given the low-mutation limit

$$P_{eq}(a) = \frac{1}{Z} P_0(a) \exp[2NF(a)].$$
For broad-acting transcription factors, **high-affinity sites** \((E < E_b)\) are statistically **overrepresented**.
For broad-acting transcription factors, high-affinity sites \((E < E_b)\) are statistically **overrepresented**.

At stationarity, the ensembles of functional and background sites determine the **average fitness landscape** \(F(E)\) of a site:

\[
2NF(E) = \log \frac{Q(E)}{P_0(E)} + \text{const.}
\]
From Energy distribution to Fitness

- For broad-acting transcription factors, **high-affinity sites** \((E < E_b)\) are statistically **overrepresented**.

- At stationarity, the ensembles of functional and background sites determine the **average fitness landscape** \(F(E)\) of a site:

\[
2N F(E) = \log \frac{Q(E)}{P_0(E)} + \text{const.}
\]

\[
F(E) = \frac{c}{1 + \exp[\eta(E - E^*)]}
\]
For broad-acting transcription factors, high-affinity sites ($E < E_b$) are statistically overrepresented.

At stationarity, the ensembles of functional and background sites determine the average fitness landscape $F(E)$ of a site:

$$2NF(E) = \log \frac{Q(E)}{P_0(E)} + \text{const.}$$

![Graph showing the binding probability and energy distribution with equation $F(E) = \frac{c}{1 + \exp[\eta(E - E*)]}$.](image)
Inference of the binding energies
Functional sites: selection coefficients

- For functional sites
  \[ F(E) \approx \frac{c}{2} - \frac{c}{4} \eta (E - E^*) \]
  \[ = -\eta \sum_{i=1}^{l} \epsilon_i(a_i) \]
- Single site selection coefficient
  \[ s_i(a_i) \propto -\eta \epsilon_i(a_i) \]
- Frequencies at equilibrium
  \[ f_{eq}(a_i) = \frac{1}{Z} \exp[-\eta \epsilon_i(a_i)] \]
Equilibrium nucleotide frequencies: the CRP case

- Sequence Alignment
- Nucleotide frequencies

CRP binding sites

---

CRP binding sites:

```
AATAATGTGAGTTAGCTCAGTCATTTACTTAT
```

Sequence Alignment:

```
1 [ - - ] 26
ATAACCCAGGCATTTAGCTCACACTTTAT
AAAATATGTGATACCAATACCTCAGAGAAT
ATAATGCTGATCCATTACACCATATT
ATAATGGTAGGAGACCTTACACCAATATT
GATTATTTGACAAGGCTCAACTTTTG
CAACAGGGATCTACTGCAACTTTAGT
ATTCCCTGTAAGCACTGCAGAACATT
CTTATGCTGGCTAAGCACTGCAGAACATT
GACCCCGGTATAGCAGCAAATGAAACAG
AGAATTGCTAATTGTTACAGAAGAC

---

Nucleotide frequencies:

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
AAAAATGTGAGTTAGCTCAGTCATTTACTTAT
```

---

CRP:

-70 -60 -50
Binding energy per site: inference vs real data

- Nucleotide frequencies

\[
\begin{align*}
\text{Equilibrium} & \quad f_j = \frac{e^{-\eta E_j}}{\sum_{i=1}^{4} e^{-\eta E_i}} \\
E_j & \propto -\log[f_j]
\end{align*}
\]

- Experimental (-) energy measure

CRP emat. in units of \(k_B T\). Site length: 26
Comparison (i): Energy per sequence
Comparison (ii): Entropy per site

\[ S = - \sum_{i=1}^{4} f_j \log_2(f_j) \]
Comparison (ii): Entropy per site

\[ S = - \sum_{i=1}^{4} f_j \log_2(f_j) \]

Entropy = Information = Function!