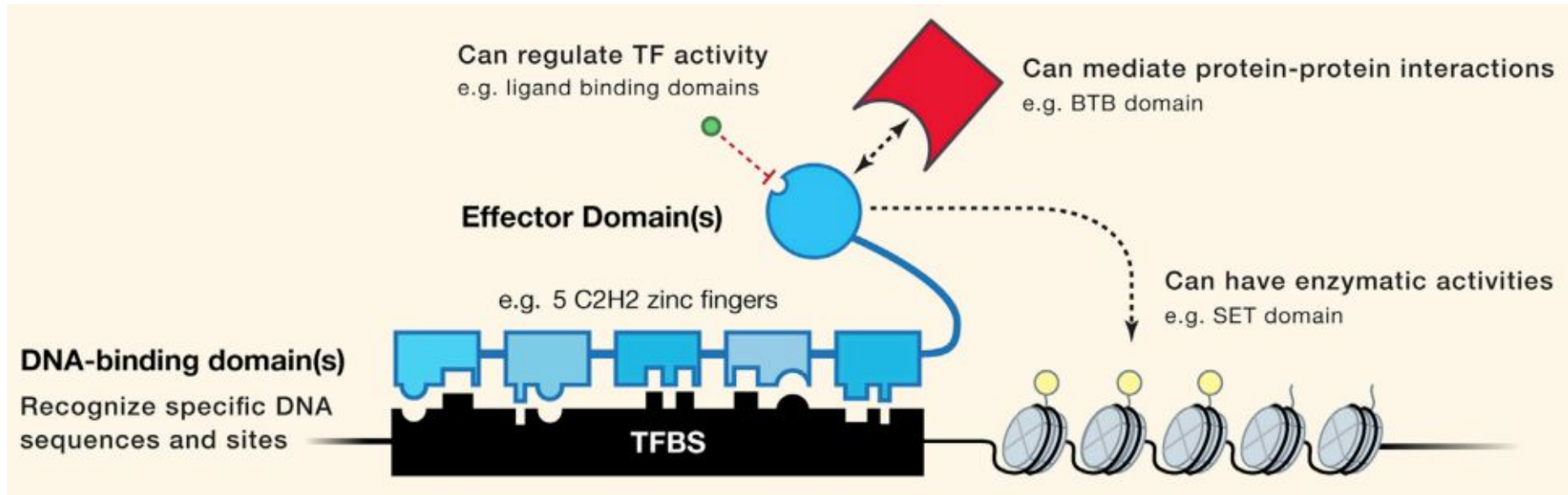


Week 06: Regulatory genomics

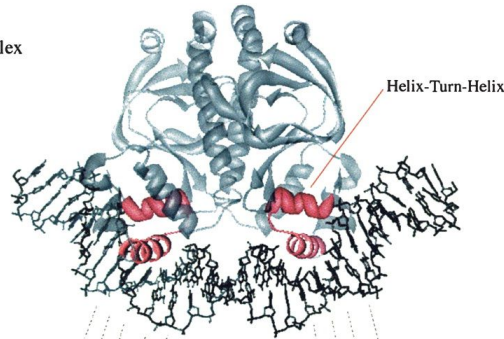
- DNA-binding sites/motifs
- Position-weight matrices
- ChIP-seq
- Motif-finding
 - Expectation-Maximization
 - Gibbs Sampling

Transcriptional regulation by TFs



Transcriptional regulation by TFs

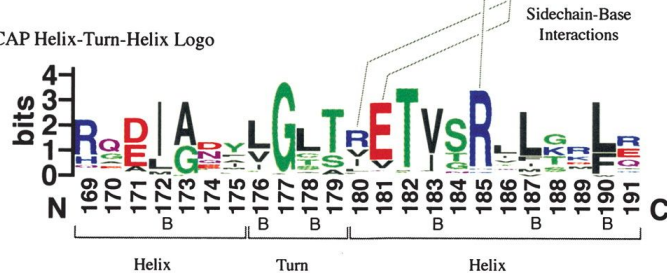
A CAP-DNA Complex



B CAP recognition site DNA Logo



C CAP Helix-Turn-Helix Logo



(A) 3D protein structure of CAP (Catabolite Activator Protein, also known as CRP), a transcriptional activator that binds at >100 sites within the *Escherichia coli* genome.

(B) CAP binding-site logo (based on 59 binding sites):

- Approximately palindromic - provides two very similar recognition sites, one for each subunit of the dimer.
- The binding site lacks perfect symmetry, possibly due to the inherent asymmetry of the operon promoter region.
- The displacement of the two halves is 11 bp, or approximately one full turn of the DNA helix.
- Additional interactions occur between the protein and the first and last two bases within the DNA minor groove, where the protein cannot easily distinguish A from T, or G from C.

(C) The helix-turn-helix motif from the CAP family of homodimeric DNA binding proteins.

Consensus sequence of DNA-binding sites

EcoRI binds to the 6-mer
GAATTC (palindrome).

- Occurs once every 4^6
(= 4,096) bp in a
random DNA
sequence.

HindIII bind to GTYRAC.

- Occur once per $4^4 \times 2^2$
(= 1,024) bp.

HEM13 CCCATTGTTCTC

HEM13 TTTCTGGTTCTC

HEM13 TCAATTGTTTAG

ANB1 CTCATTGTTGTC

ANB1 TCCATTGTTCTC

ANB1 CCTATTGTTCTC

ANB1 TCCATTGTTCGT

ROX1 CCAATTGTTTTC

YCHAATTGTTCTC

Motif instance → Motif

A 0027000000010

C 464100000505

G 000001800112

T 422087088261

Position
frequency
matrix

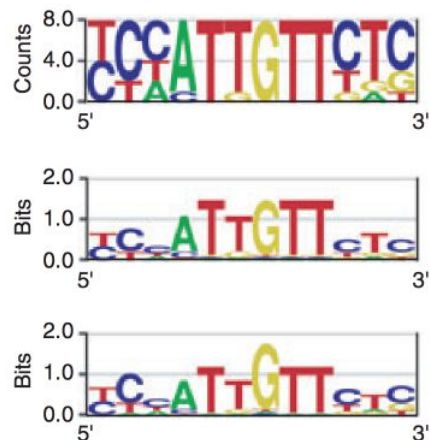


Sequence
logo



Consensus sequence of DNA-binding sites

A 002700000010
C 464100000505
G 000001800112
T 422087088261



$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

Scaling sequence logos based on 'information content' than frequency.

- $f_{b,i}$: frequency of base b at position i .
- Perfectly conserved: 2 bits of information.
- Two of the four bases occur 50% of the time each: 1 bit.
- All four bases occur equally often: no information.

HindII binds to GTYRAC.

- What is its information content?

Consensus sequence of DNA-binding sites

A 002700000010
C 464100000505
G 000001800112
T 422087088261



$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

Relative entropy (a.k.a. Kullback-Leibler distance) to correct for background nucleotide frequencies.

$$W(b,i) = \log_2 \frac{f_{b,i}}{p_b}$$

Position weight matrix (PWM).

Consensus sequence of DNA-binding sites

A 0027000000010
C 464100000505
G 000001800112
T 422087088261

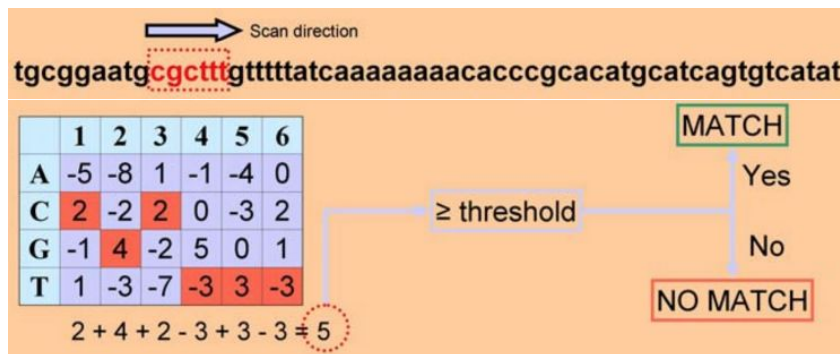


$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

Relative entropy (a.k.a. Kullback-Leibler distance) to correct for background nucleotide frequencies.

$$W(b,i) = \log_2 \frac{f_{b,i}}{p_b}$$

Position weight matrix (PWM).



Consensus sequence of DNA-binding sites

A 002700000010
C 464100000505
G 000001800112
T 422087088261

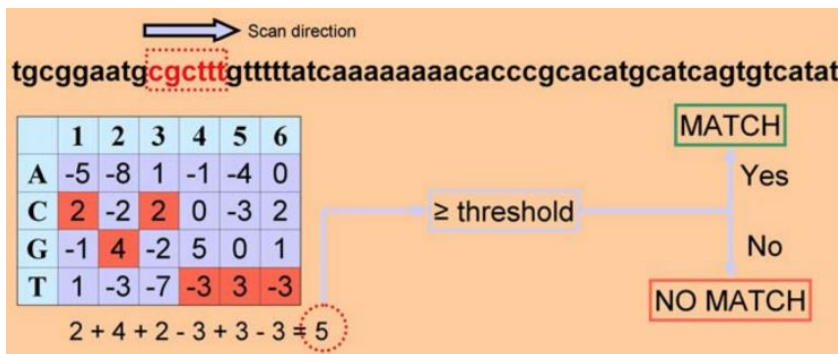
A generative model!

Assumptions:

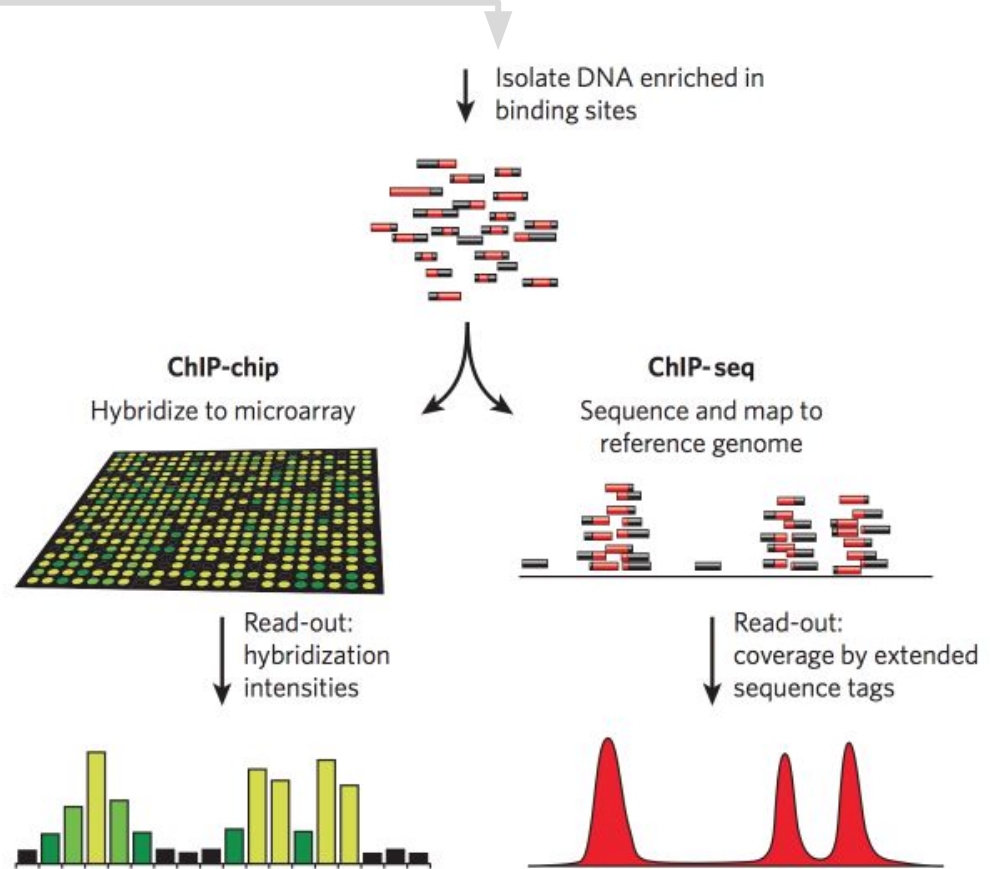
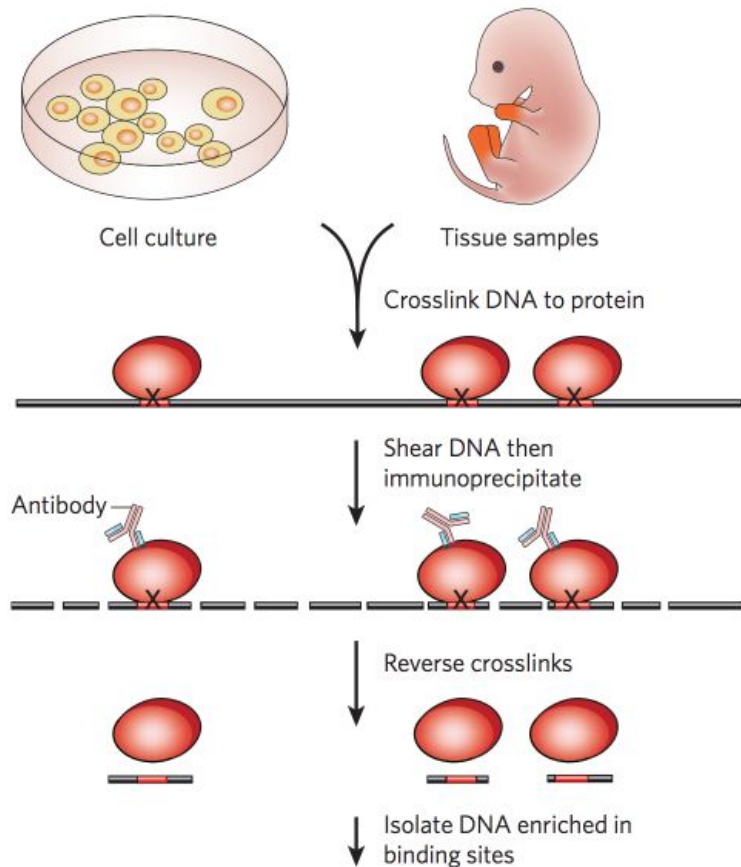
- Independence of positions
- Fixed spacing



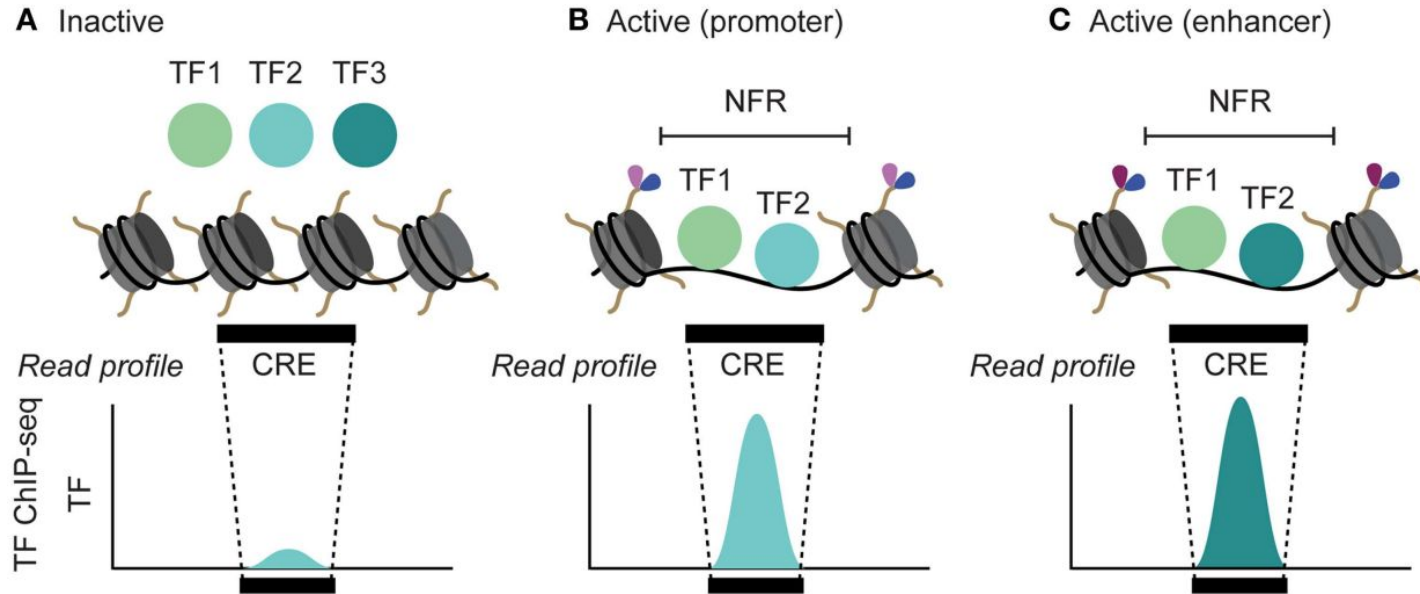
Position weight matrix (PWM).



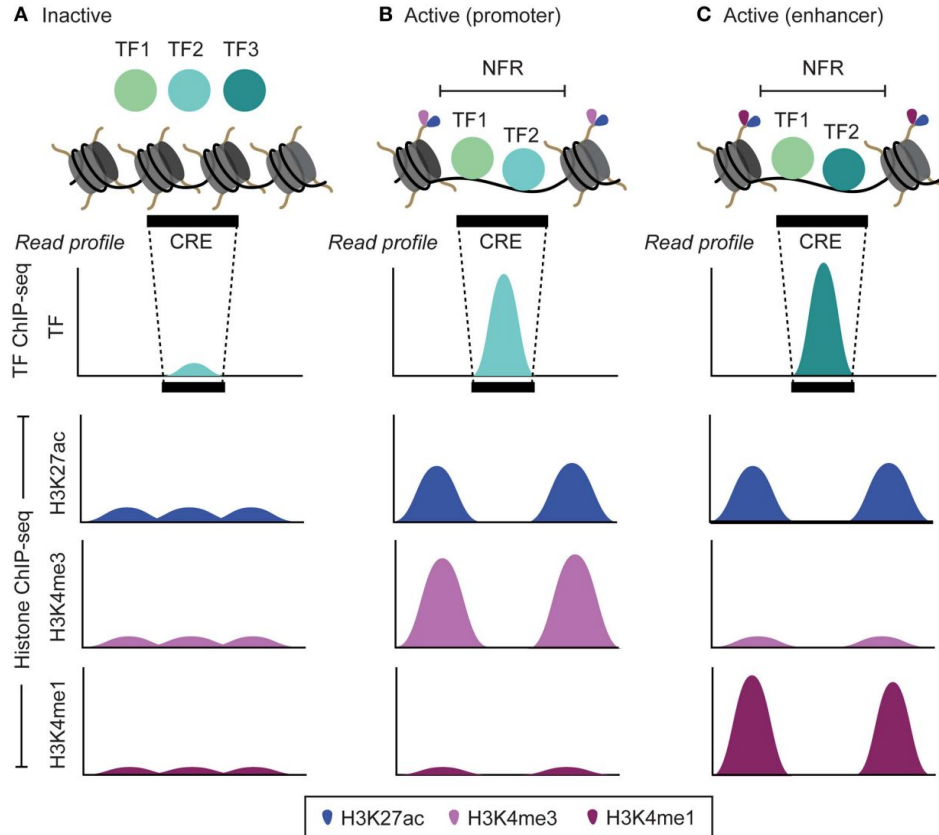
Mapping of regulatory elements using ChIP-chip and ChIP-seq



Mapping of regulatory elements using ChIP-chip and ChIP-seq



Mapping of regulatory elements using ChIP-chip and ChIP-seq



Mapping of regulatory elements using ChIP-chip and ChIP-seq

