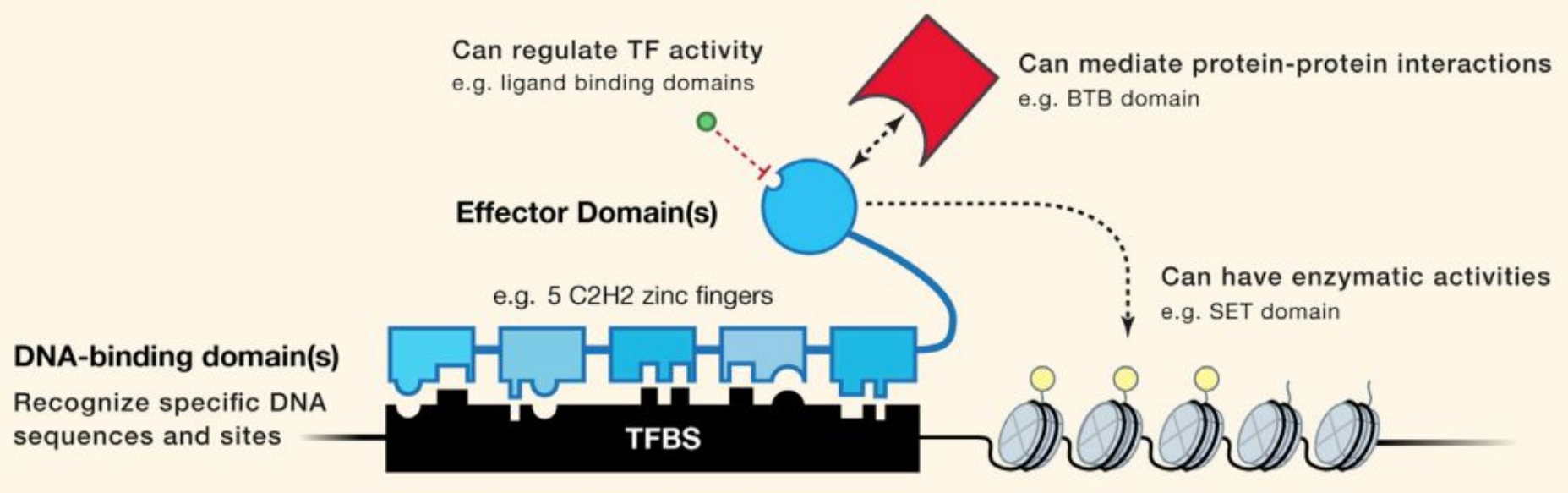
Week 06: Regulatory genomics

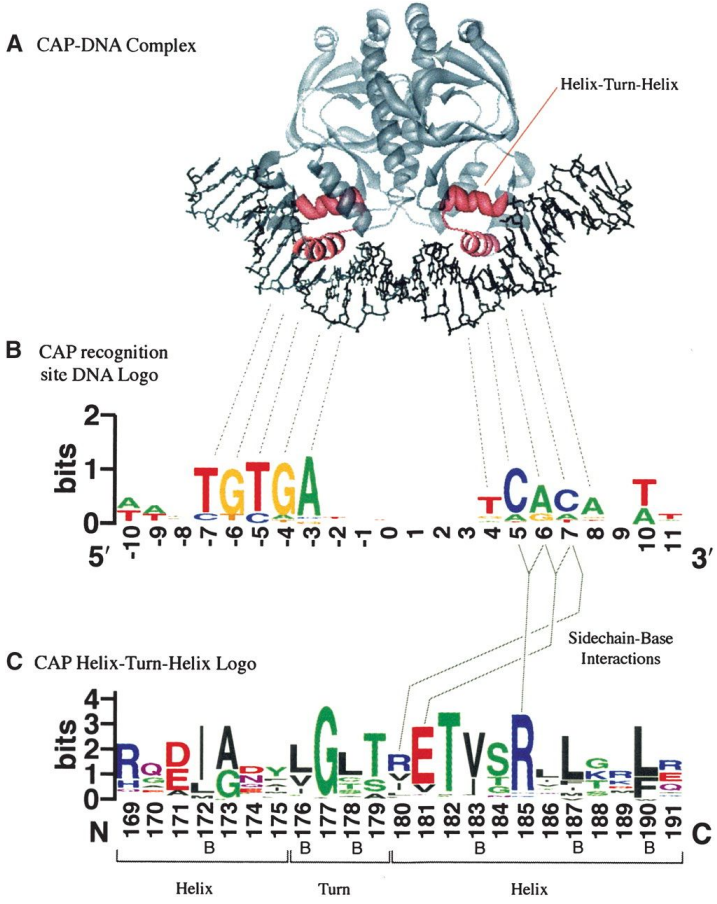
- DNA-binding sites/motifs
  - ChIP-seq
  - Position-weight matrices
  - Motif-finding
    - Expectation-Maximization
    - Gibbs Sampling

Transcriptional regulation by TFs



Lambert (2018) Cell

Transcriptional regulation by TFs



(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.

Crooks (2004) Genome Res.

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.

HindII bind to GTYRAC.

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

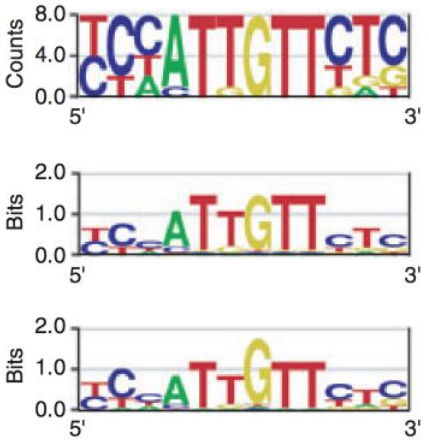
```
HEM13   CCCATTGTTCTC
HEM13   TTTCTGGTTCTC
HEM13   TCAATTGTTTAG
ANB1    CTCATTGTTGTC
ANB1    TCCATTGTTCTC
ANB1    CCTATTGTTCTC
ANB1    TCCATTGTTCGT
ROX1    CCAATTGTTTTG

        YCHATTGTTCTC
```

Motif instance ⟶ Motif

```
A 002700000010
C 464100000505
G 000001800112
T 422087088261
```

Position frequency matrix

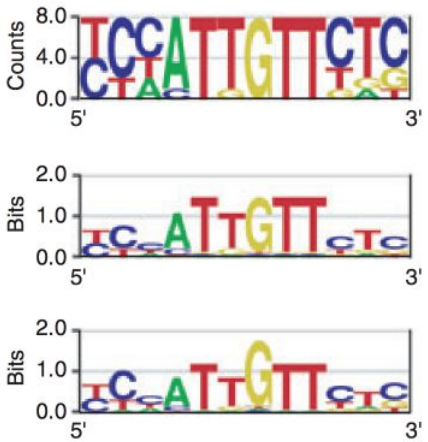


Sequence logo

D'haeseleer (2016) Nat. Biotech.

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.

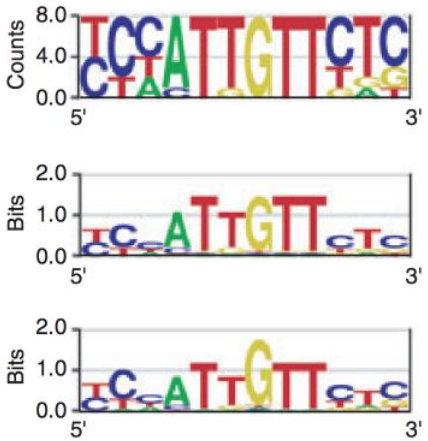
HindII bind to GTYRAC.

- What is its information content?

D'haeseleer (2016) Nat. Biotech.

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.

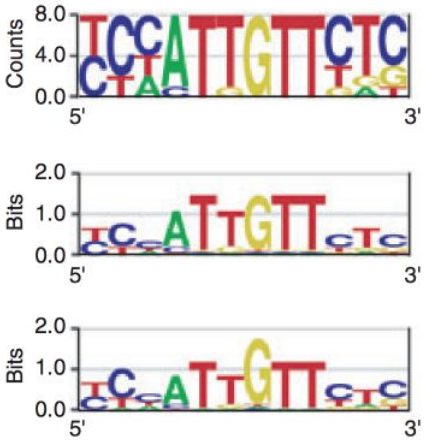
$$W(b,i) = \log_2 \frac{f_{b,i}}{p_b}$$

Position weight matrix (PWM).

D'haeseleer (2016) Nat. Biotech.

Consensus sequence of DNA-binding sites

```
A 002700000010
C 464100000505
G 000001800112
T 422087088261
```

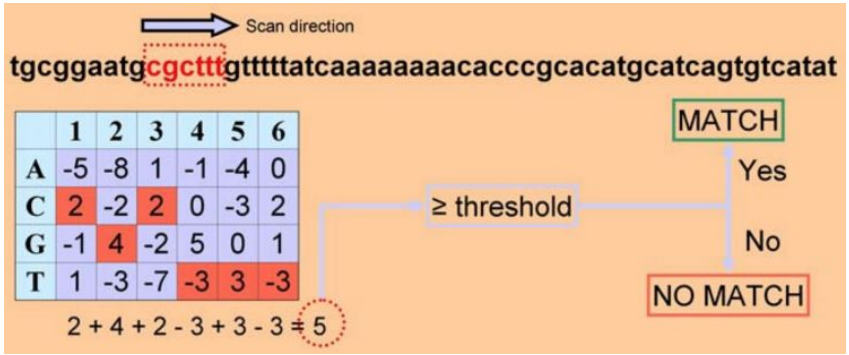


$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

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

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

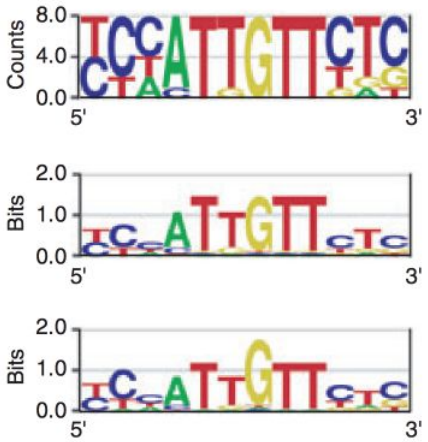
Position weight matrix (PWM).



D'haeseleer (2016) Nat. Biotech.

Consensus sequence of DNA-binding sites

```
A 002700000010
C 464100000505
G 000001800112
T 422087088261
```

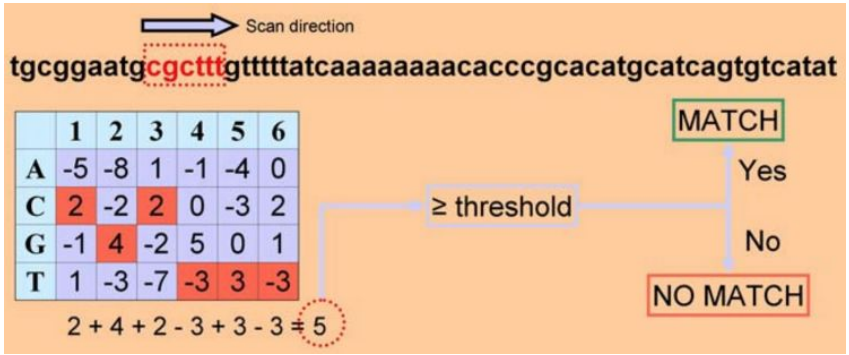


A generative model!

Assumptions:

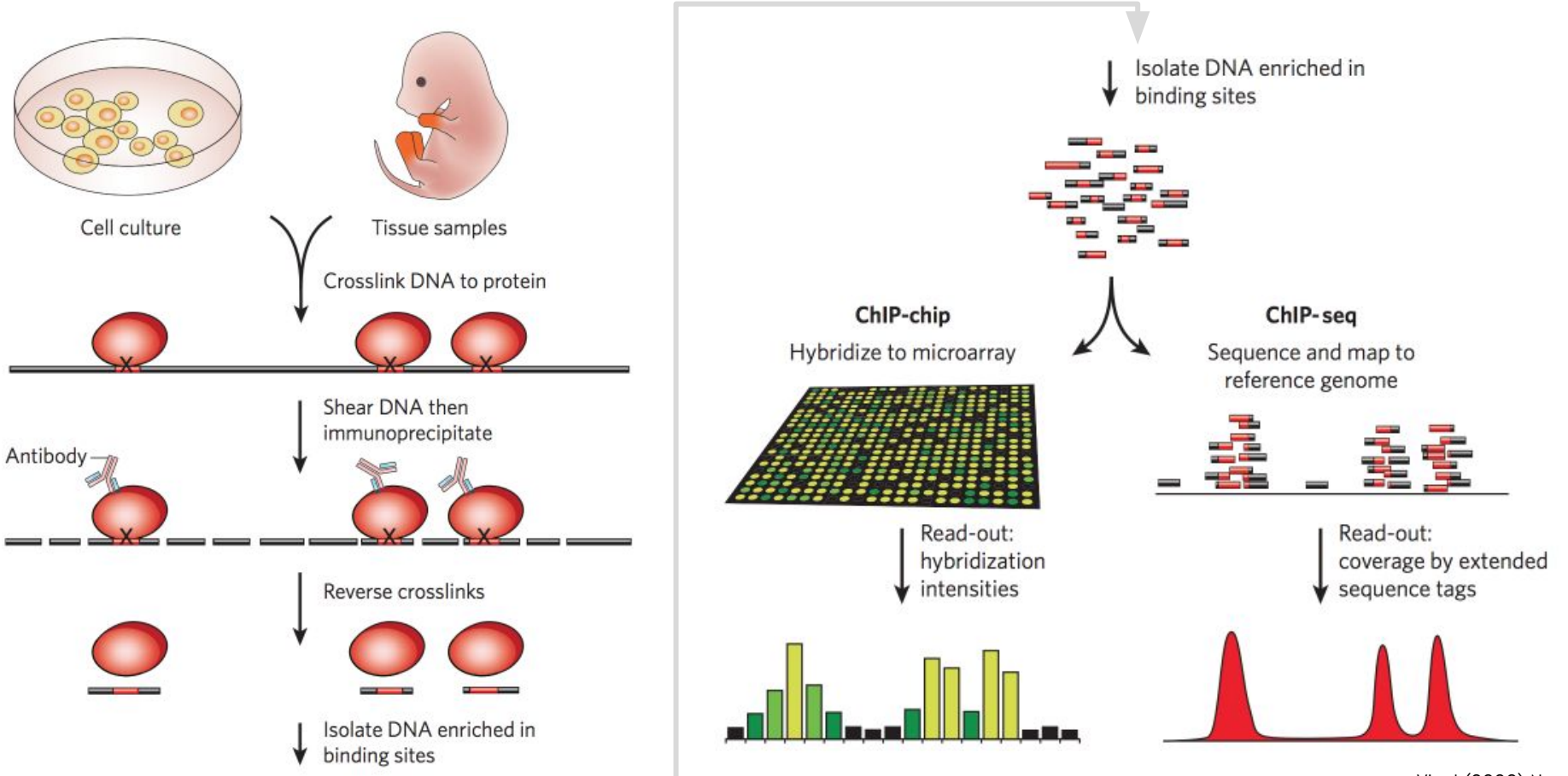
- Independence of positions
- Fixed spacing

Position weight matrix (PWM).



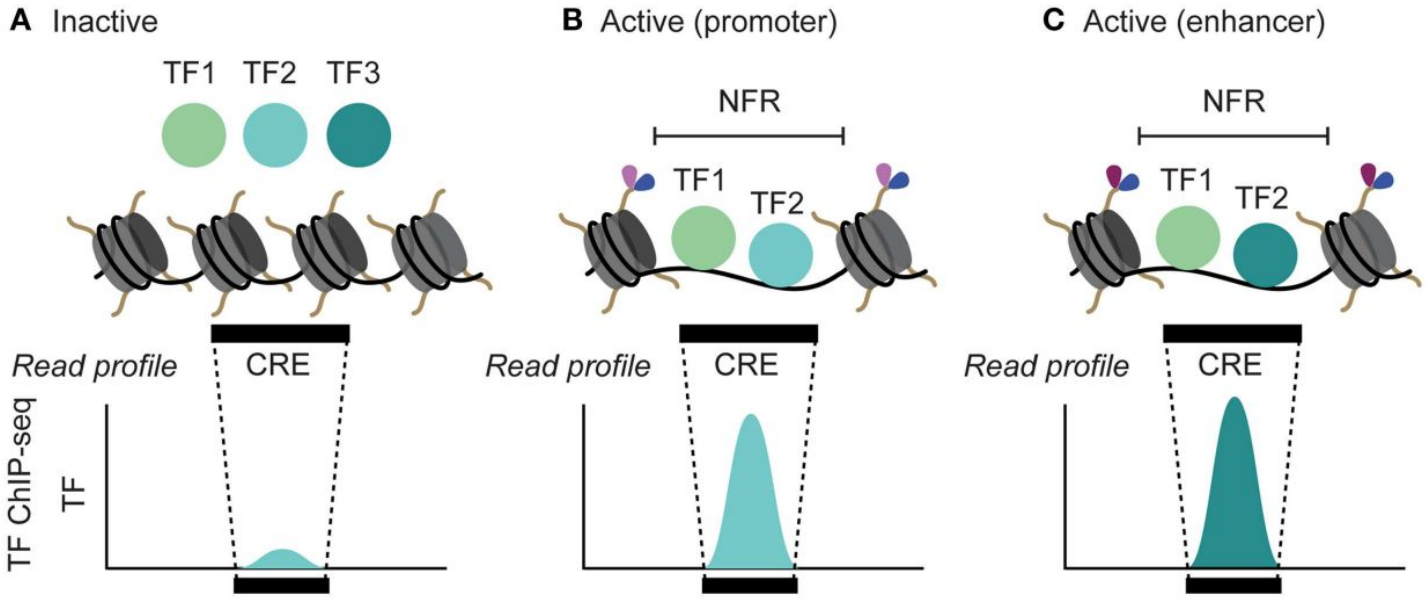
D'haeseleer (2016) Nat. Biotech.

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



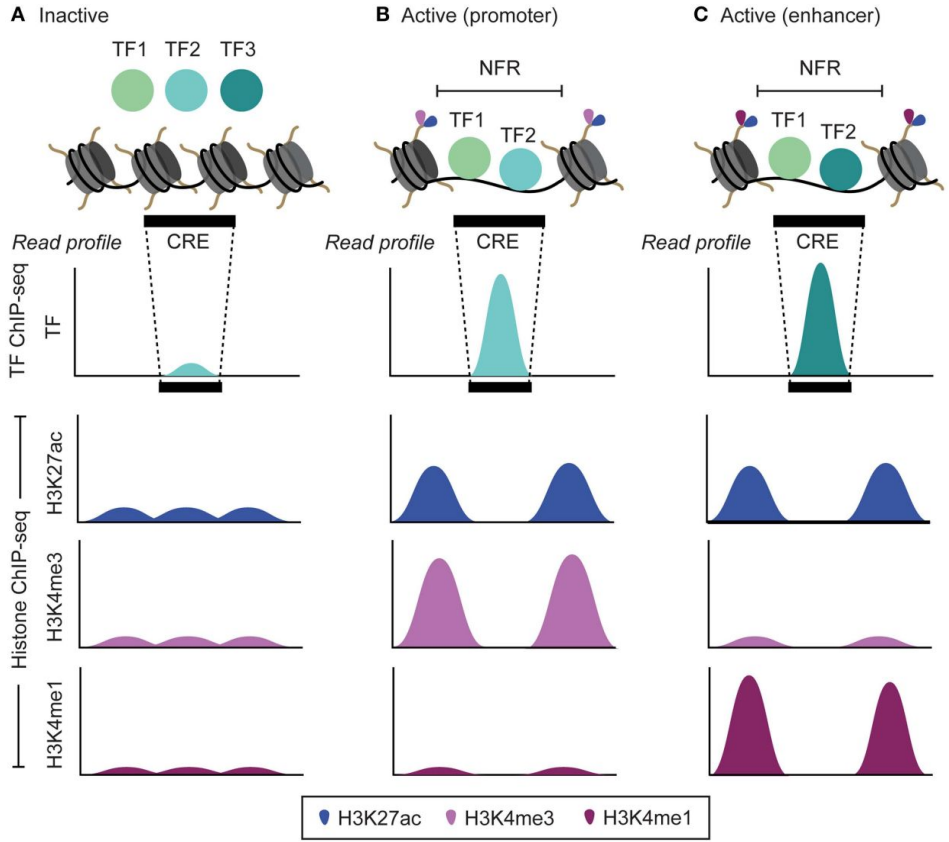
Visel (2009) Nature

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



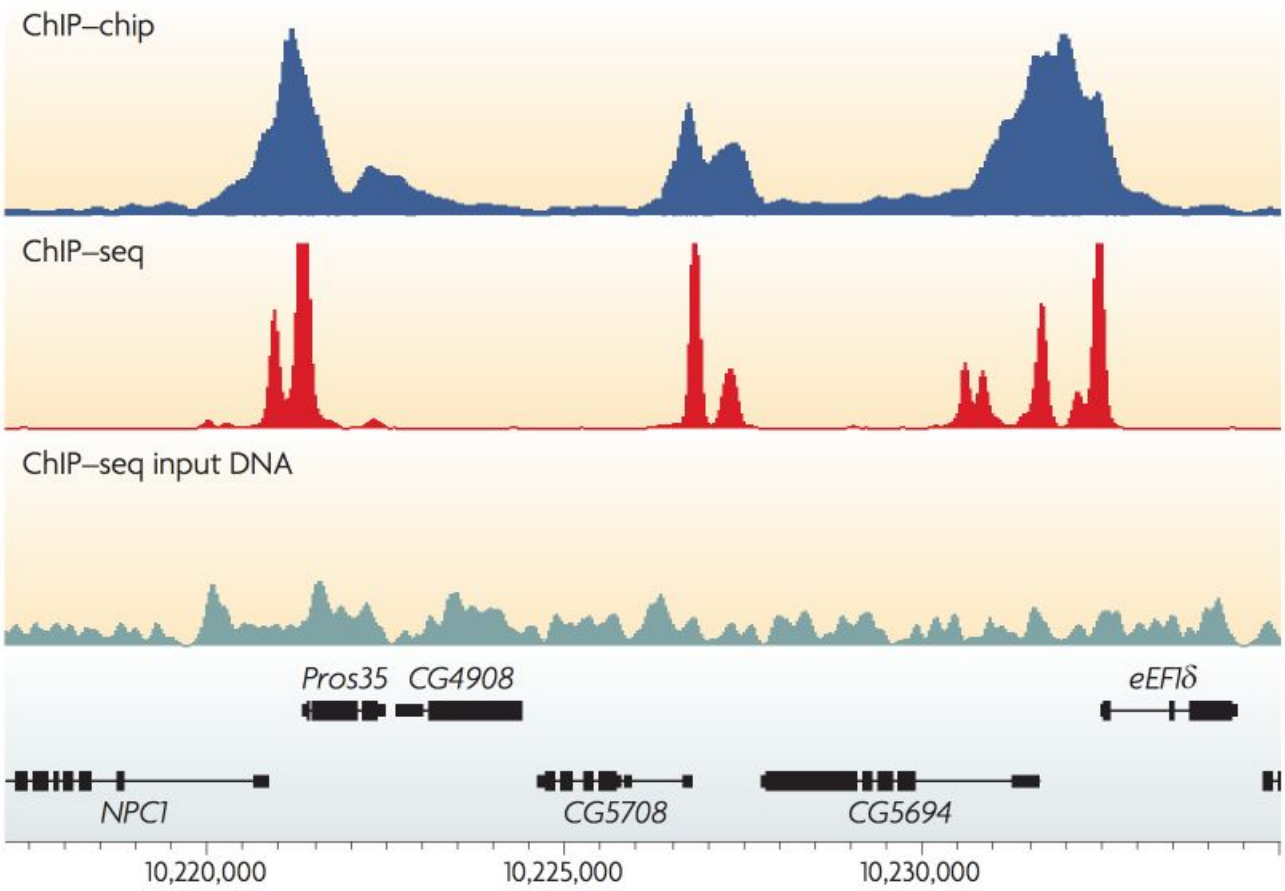
Pundhir (2015) Front. Genet.

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



Pundhir (2015) Front. Genet.

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

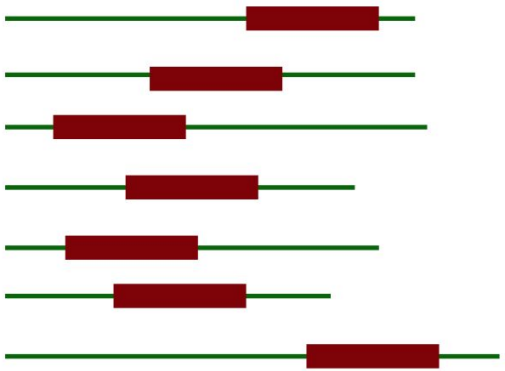


Park (2009) Nat. Rev. Genet.

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

Sequences are not aligned, we don't know motif positions.

We also don't know what the motif looks like.



The motif model learning task:

- Given: a set of sequences that are thought to contain occurrences of an unknown motif of interest
- Do:
  - infer a model (PWM) of the motif, and
  - predict the locations of the motif occurrences in the given sequences.

Expectation-Maximization: Iteratively refine positions / motif profile

Gibbs sampling: Iteratively sample positions / motif profile

Expectation-Maximization algorithm (EM)

$\hat{\theta}_A$= ?

$\hat{\theta}_B$= ?

A coin-flipping experiment

- θ_A & θ_B are the biases of two coins A & B.
- **Goal**: Estimate $\theta = (\theta_A, \theta_B)$ by repeating the following procedure five times:
  - Randomly choose one of the two coins (with equal probability)
  - Perform ten independent coin tosses with the selected coin.

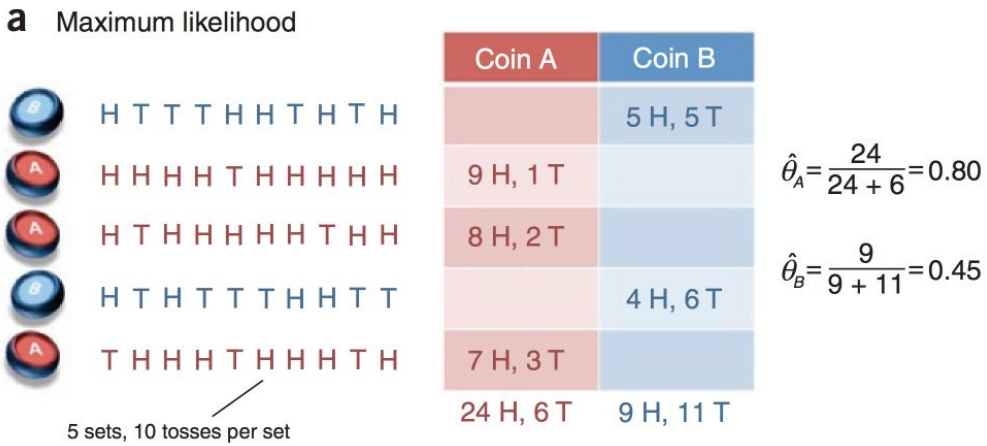
$x = (x_1, x_2, \ldots, x_5) \mid x_i \in \{0,1,\ldots,10\}$ is the no. of heads observed during the ith set of tosses.

$z = (z_1, z_2, \ldots, z_5) \mid z_i \in \{A,B\}$ is the identity of the coin used during the ith set of tosses.

Maximum likelihood estimation: statistical model that has the highest probability of generating the observed data – θ that maximizes $\log P(x,z;\theta)$.

Do & Batzoglou (2008) Nat. Biotech.

Expectation-Maximization algorithm (EM)



A coin-flipping experiment

- θ_A & θ_B are the biases of two coins A & B.
- **Goal**: Estimate $\theta = (\theta_A, \theta_B)$ by repeating the following procedure five times:
  - Randomly choose one of the two coins (with equal probability)
  - Perform ten independent coin tosses with the selected coin.

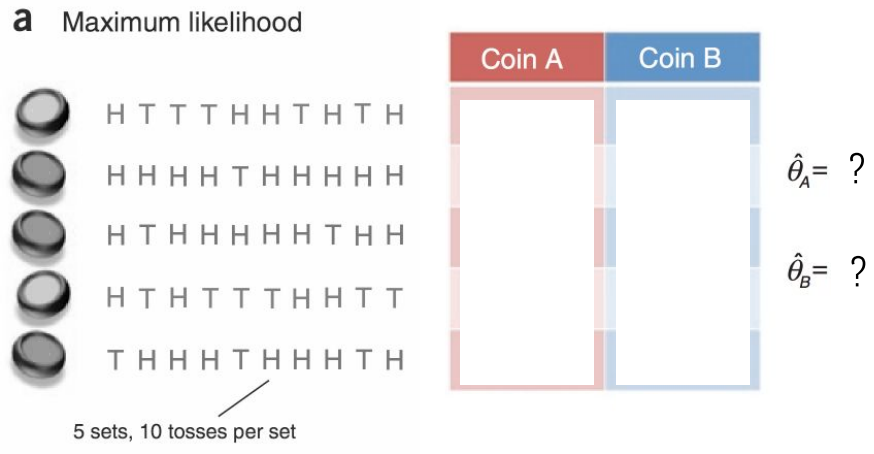
$x = (x_1, x_2, \ldots, x_5) \mid x_i \in \{0,1,\ldots,10\}$ is the no. of heads observed during the ith set of tosses.

$z = (z_1, z_2, \ldots, z_5) \mid z_i \in \{A,B\}$ is the identity of the coin used during the ith set of tosses.

Maximum likelihood estimation: statistical model that has the highest probability of generating the observed data – θ that maximizes $\log P(x,z;\theta)$.

Do & Batzoglou (2008) Nat. Biotech.

Expectation-Maximization algorithm (EM)



$x = (x_1, x_2, \ldots, x_5) \mid x_i \in \{0,1,\ldots,10\}$ is the no. of heads observed during the ith set of tosses.

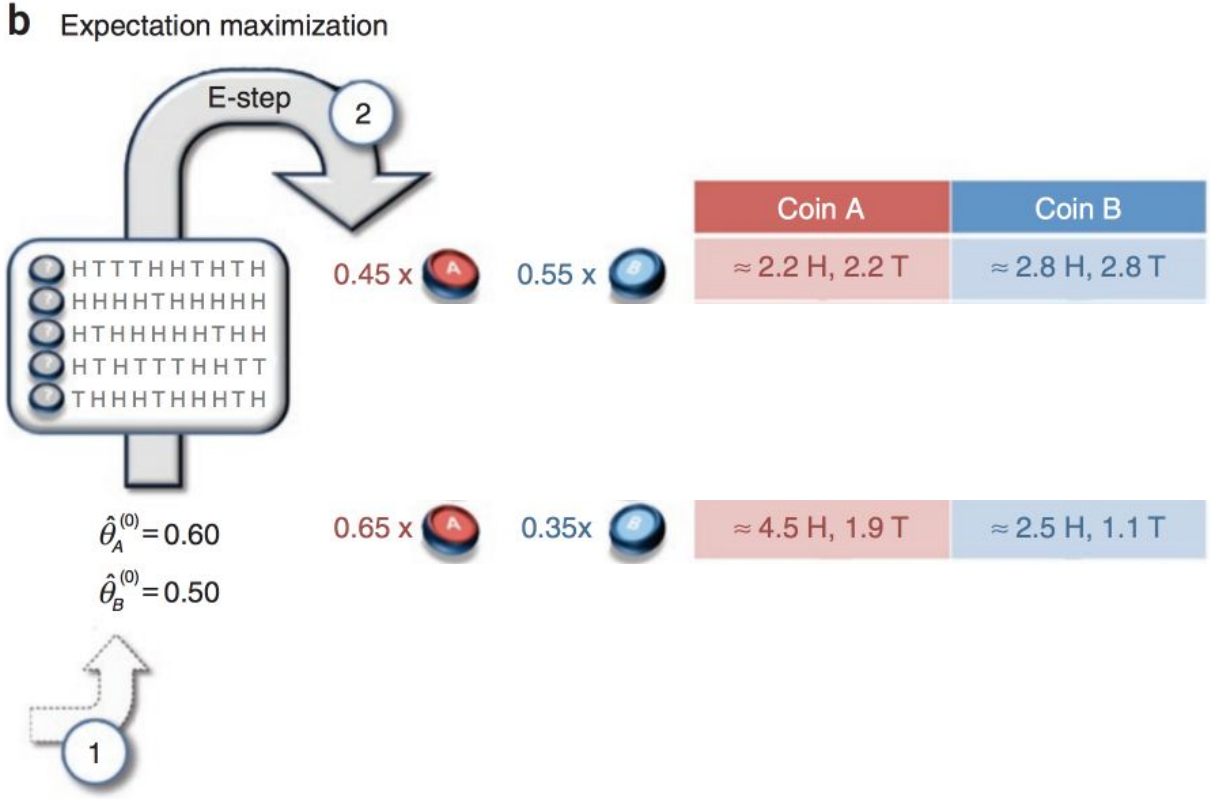
$z = (z_1, z_2, \ldots, z_5) \mid z_i \in \{A,B\}$ is the identity of the coin used during the ith set of tosses. [**Hidden variables / Latent factors**]

A coin-flipping experiment

- θ_A & θ_B are the biases of two coins A & B.
- **Goal**: Estimate $\theta = (\theta_A, \theta_B)$ by repeating the following procedure five times:
  - Randomly choose one of the two coins (with equal probability; **but you don't no which coin was chosen**.)
  - Perform ten independent tosses with the selected coin.

Do & Batzoglou (2008) Nat. Biotech.

Expectation-Maximization algorithm (EM)



E-step:

- Estimate $P(x_i, z_i \mid \theta^{(t)})$ and the expected values of the hidden variables.

M-step:

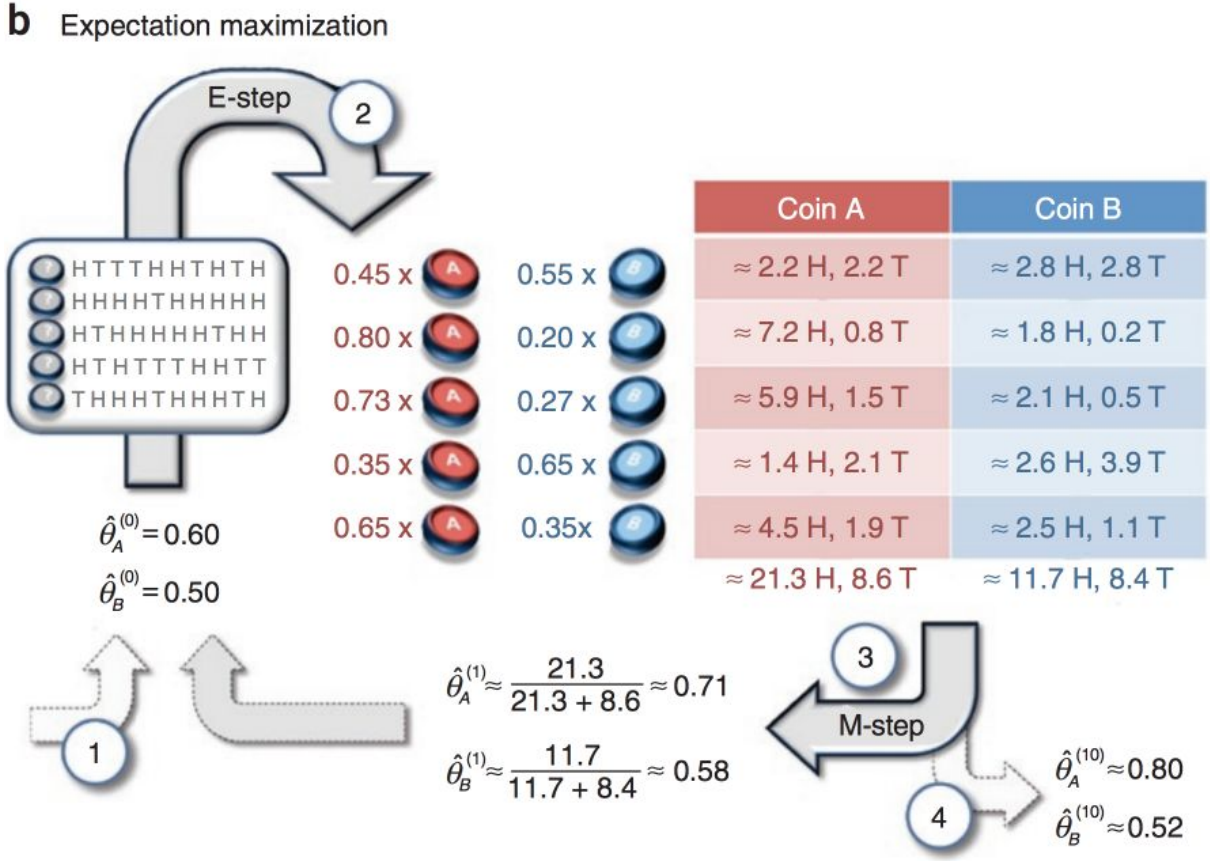
- Estimate new parameters $\theta^{(t+1)}$ given current estimates of hidden variables & parameters.

Repeat until convergence.

$P(x_i, z_i \mid \theta^{(t)})$: Likelihood function, from here on also going to be written as $P(X, Z \mid \theta)$.

Do & Batzoglou (2008) Nat. Biotech.

Expectation-Maximization algorithm (EM)



E-step:

- Estimate $P(x_i, z_i \mid \theta^{(t)})$ and the expected values of the hidden variables.

M-step:

- Estimate new parameters $\theta^{(t+1)}$ given current estimates of hidden variables & parameters.

Repeat until convergence.

$P(x_i, z_i \mid \theta^{(t)})$: Likelihood function, from here on also going to be written as $P(X, Z \mid \theta)$.

Do & Batzoglou (2008) Nat. Biotech.