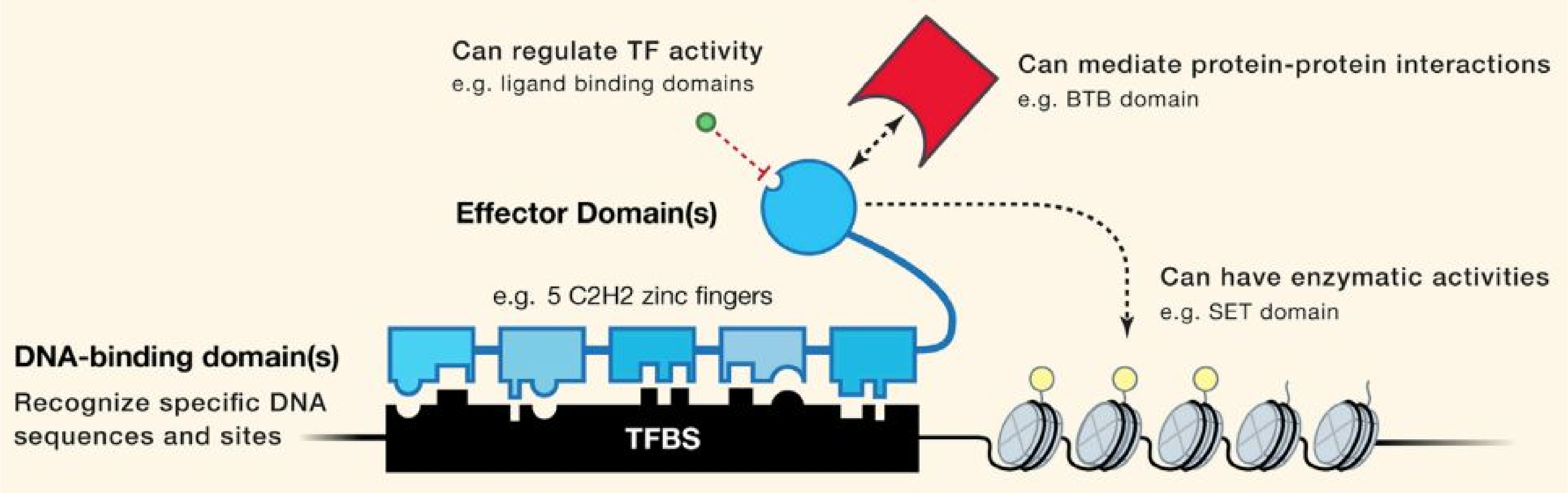
Lecture 7: Regulatory genomics

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

Transcriptional regulation by TFs


Can regulate TF activity
e.g. ligand binding domains
Can mediate protein-protein interactions
e.g. BTB domain
Effector Domain(s)
e.g. 5 C2H2 zinc fingers
Can have enzymatic activities
e.g. SET domain
DNA-binding domain(s)
Recognize specific DNA sequences and sites
TFBS

Lambert (2018) Cell

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.

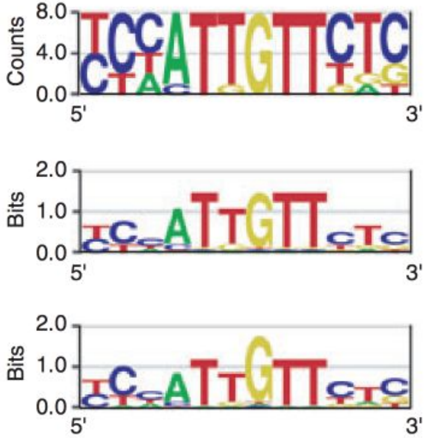
HindII bind to GTYRAC.

- 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

A	002700000010
C	464100000505
G	000001800112
T	422087088261

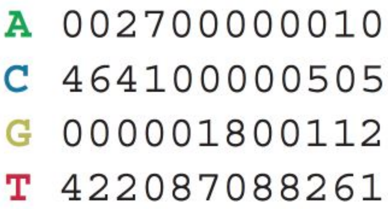
Position frequency matrix



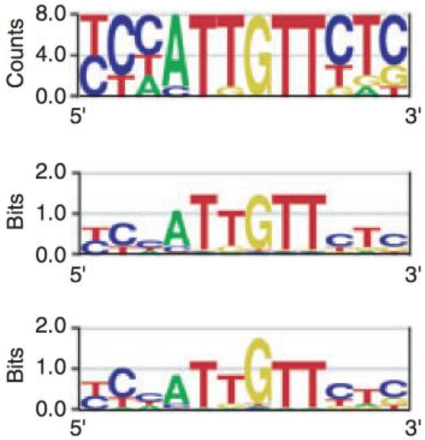
Sequence logo

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.

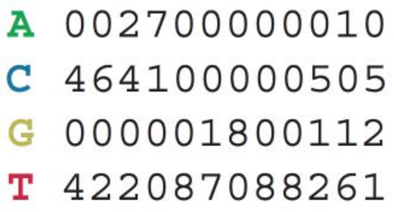
- $f_{b,i}$: frequency of base b at position i.
- Perfectly conserved: 2 bits of information.
- Two of the four bases occur 50% of the time each: 1 bit.
- All four bases occur equally often: no information.

HindII bind to GTYRAC.

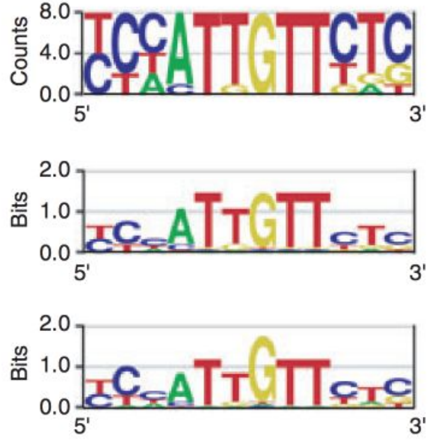
- What is its information content?

D'haeseleer (2016) Nat. Biotech.

Consensus sequence of DNA-binding sites



A	002700000010
C	464100000505
G	000001800112
T	422087088261



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

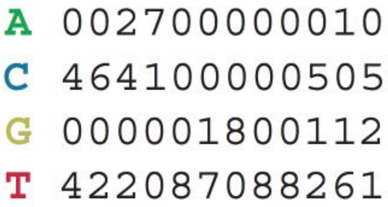
Relative entropy (a.k.a. Kullback-Leibler distance) to correct for background nucleotide frequencies.

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

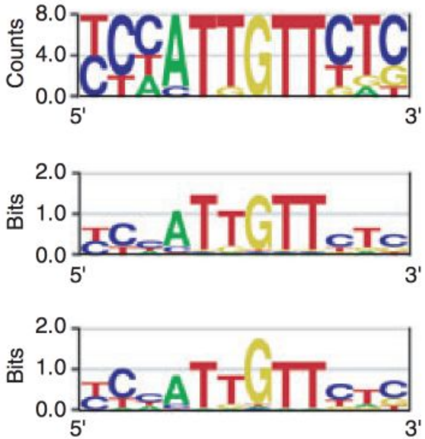
Position weight matrix (PWM).

D'haeseleer (2016) Nat. Biotech.

Consensus sequence of DNA-binding sites



A	0	0	2	7	0	0	0	0	0	0	1	0
C	4	6	4	1	0	0	0	0	0	5	0	5
G	0	0	0	0	0	1	8	0	0	1	1	2
T	4	2	2	0	8	7	0	8	8	2	6	1

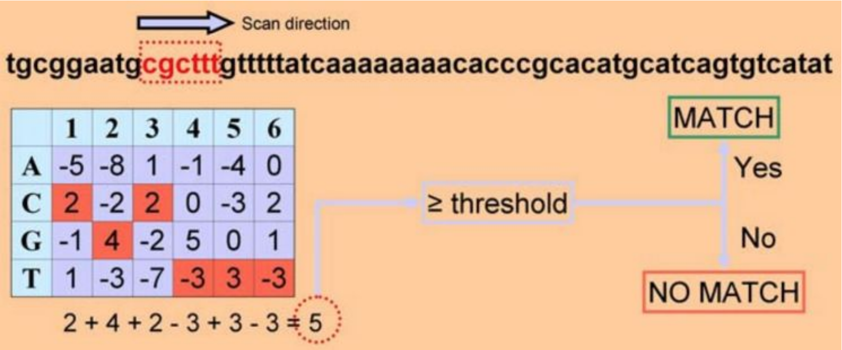


$$I_{seq}(i) = -\sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

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).

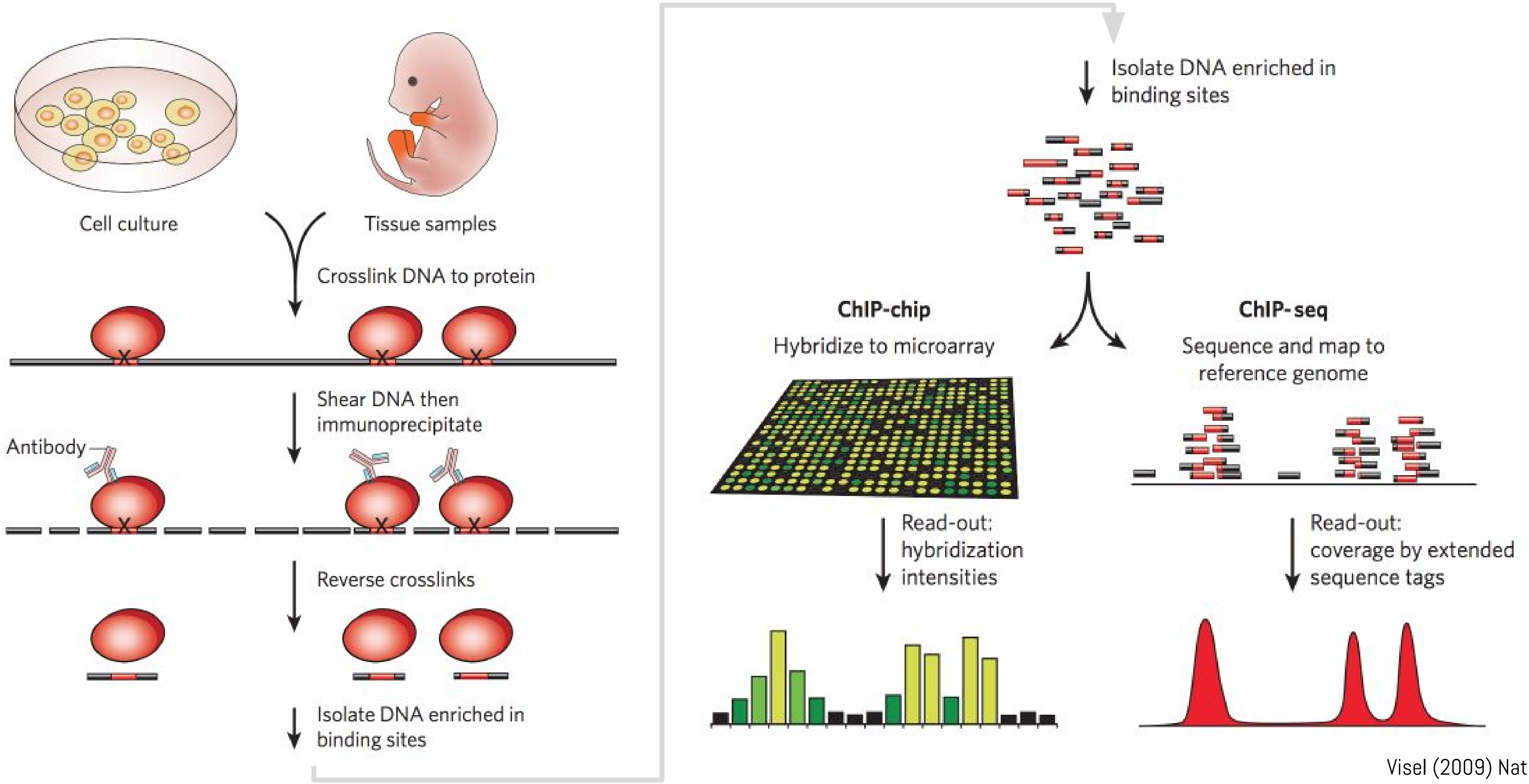


	1	2	3	4	5	6
A	-5	-8	1	-1	-4	0
C	2	-2	2	0	-3	2
G	-1	4	-2	5	0	1
T	1	-3	-7	-3	3	-3

$2 + 4 + 2 - 3 + 3 - 3 = 5$

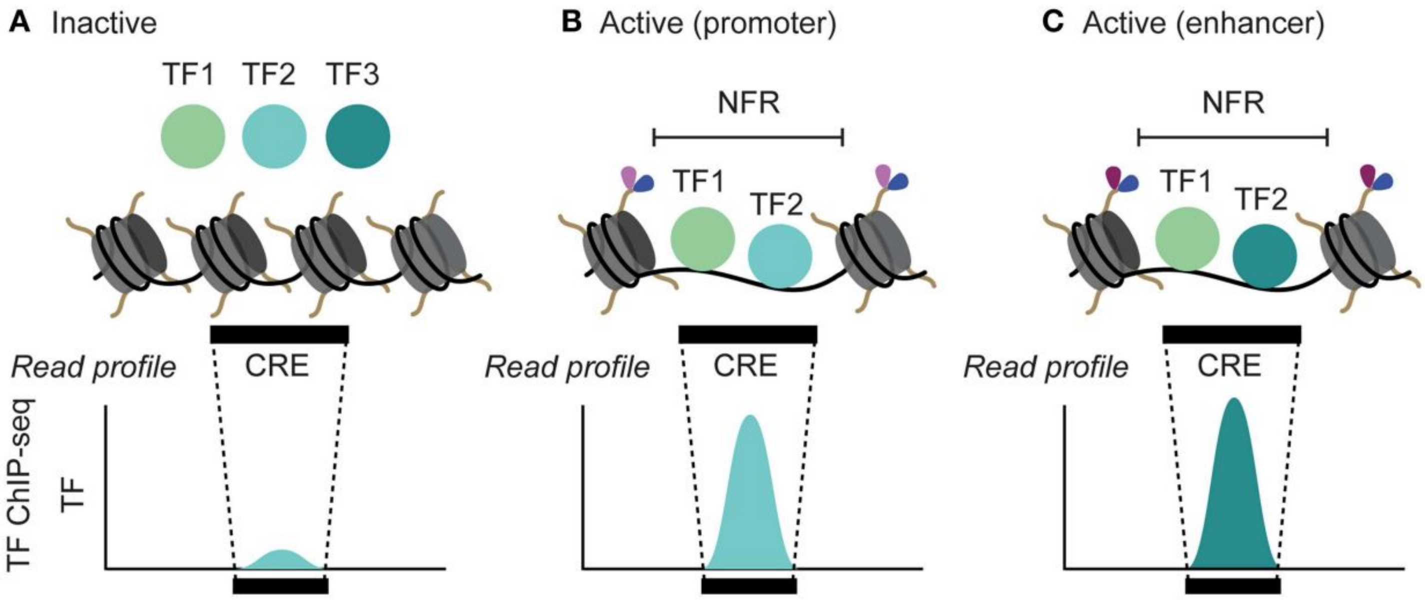
D'haeseleer (2016) Nat. Biotech.

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


Cell culture
Tissue samples
Crosslink DNA to protein
Shear DNA then immunoprecipitate
Antibody
Reverse crosslinks
Isolate DNA enriched in binding sites
Isolate DNA enriched in binding sites
ChIP-chip
Hybridize to microarray
ChIP-seq
Sequence and map to reference genome
Read-out: hybridization intensities
Read-out: coverage by extended sequence tags

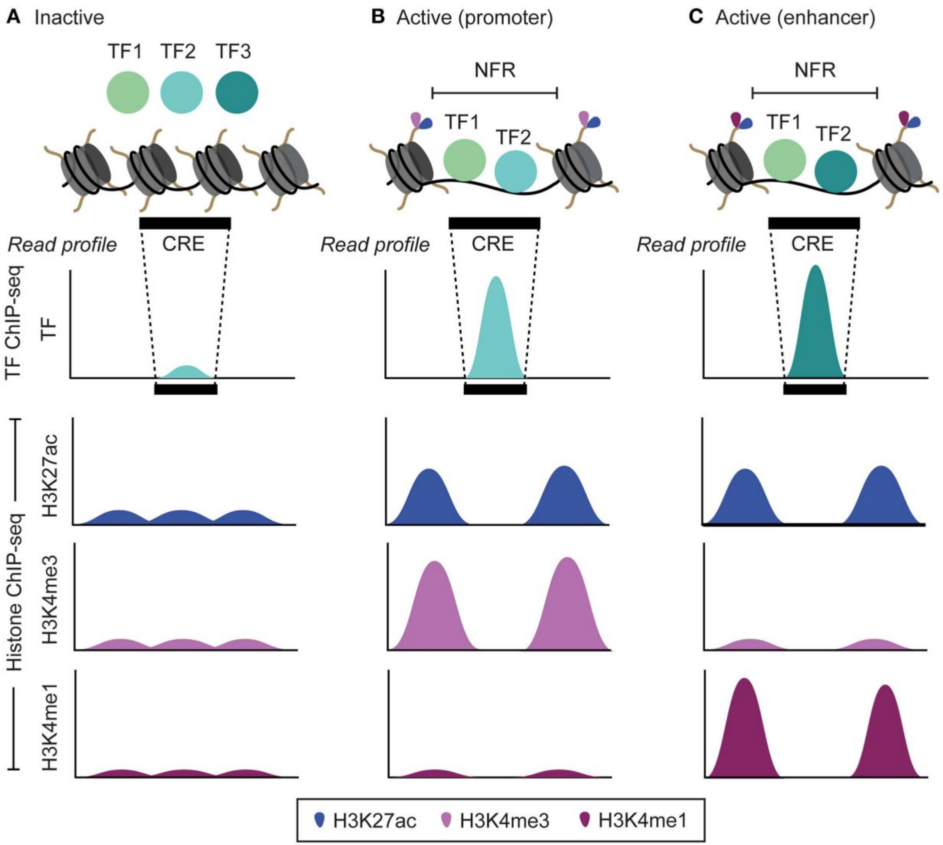
Visel (2009) Nature

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



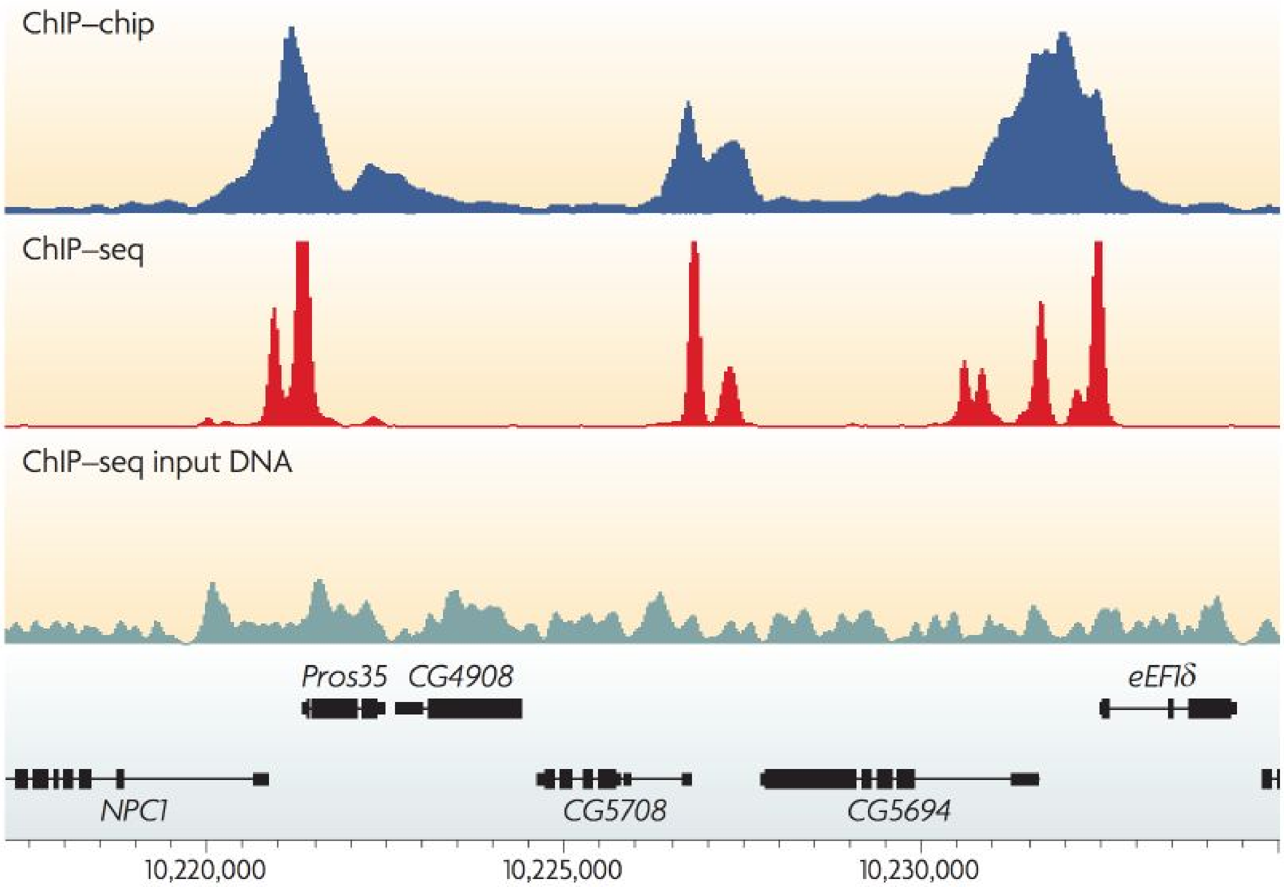
Pundhir (2015) Front. Genet.

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.

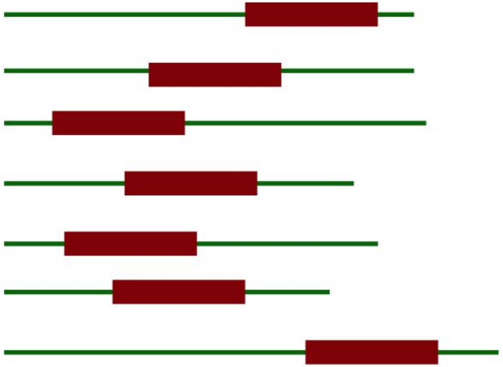
Mapping of regulatory elements using ChIP-chip and ChIP-seq

Sequences are not aligned.

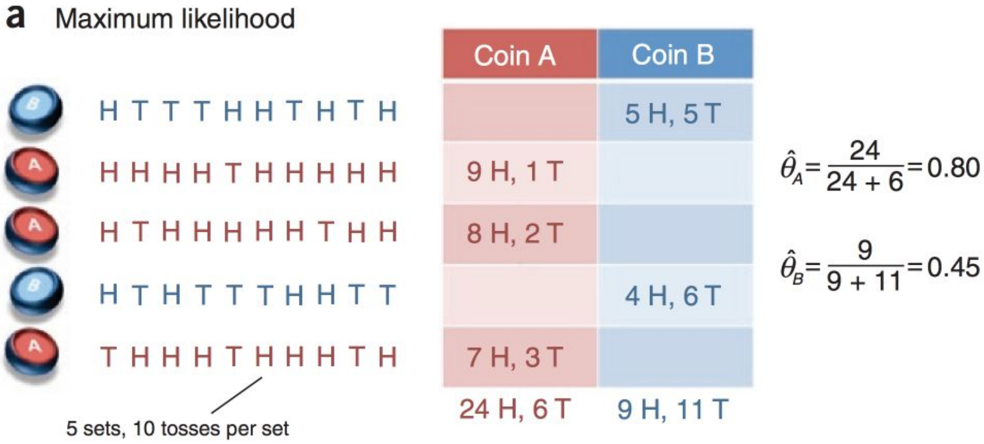
We don't know what the motif looks like.

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 algorithm (EM)



Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24 + 6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9 + 11} = 0.45$$

$x = (x_1, x_2, \ldots, x_5) \mid x_i \in \{0,1,\ldots,10\}$ 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.

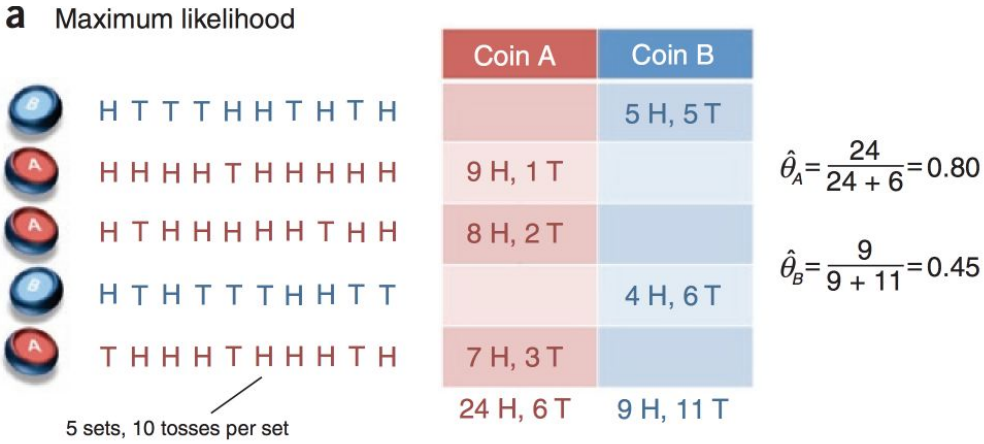
A coin-flipping experiment

- 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), and perform ten independent coin tosses with the selected coin.
    - Total of 50 coin 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)



		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

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

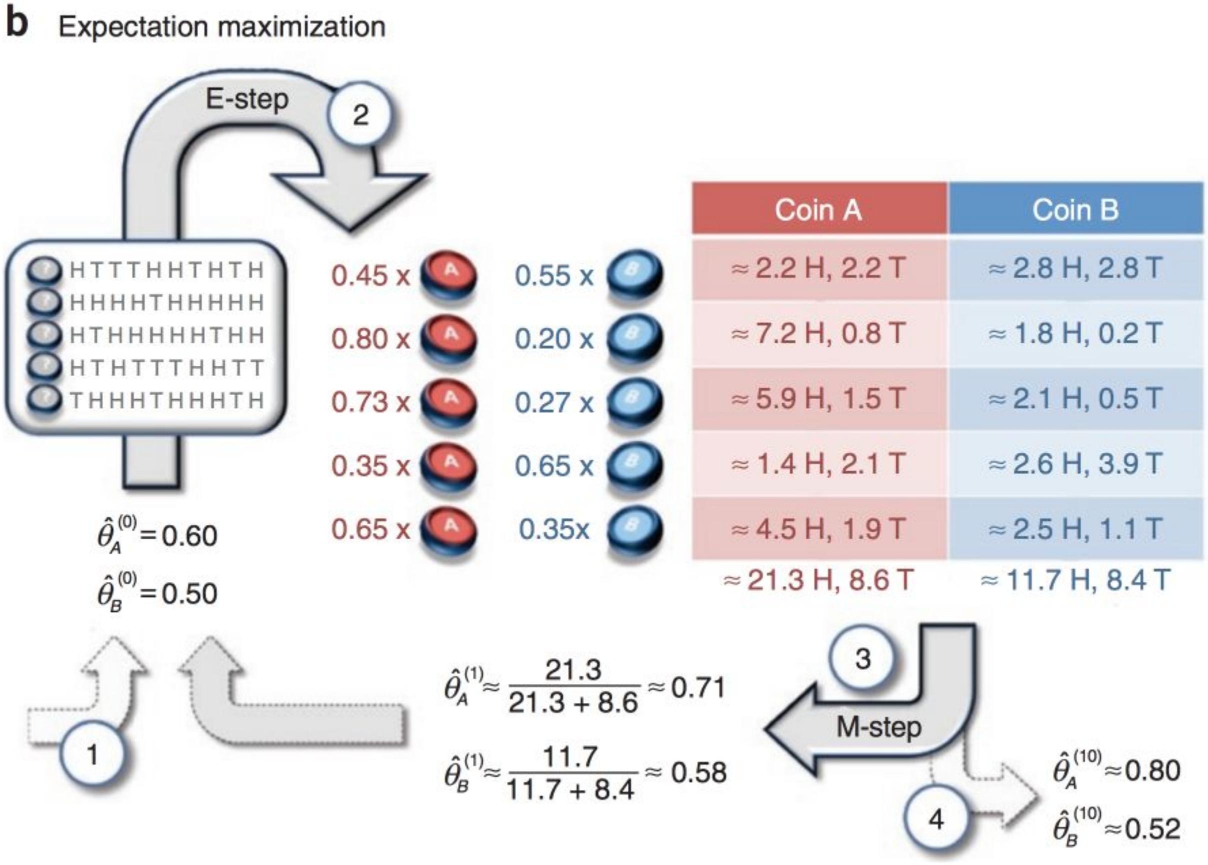
$z = (z_1, z_2, \ldots, z_5) \mid z_i \in \{A,B\}$ is the identity of the coin used during the ith set of tosses. [**Hidden variables / Latent factors**]

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), and perform ten independent coin tosses with the selected coin.
  - **Not told which coin was chosen.**

Do & Batzoglou (2008) Nat. Biotech.

Expectation-Maximization algorithm (EM)



Coin A	Coin B
≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3+8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7+8.4} \approx 0.58$$

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.

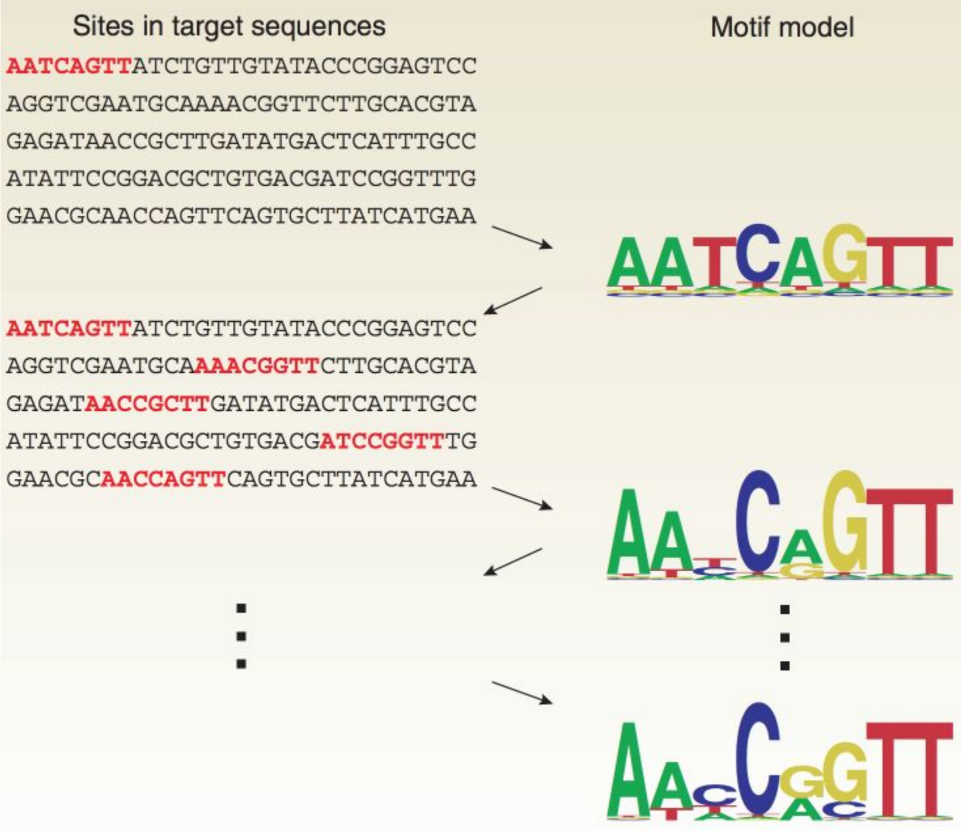
Repeat until convergence.

$P(x_i, z_i \mid \theta^{(t)})$: 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)

1. Define the probabilistic model and the likelihood function $P(X \mid \theta)$.
2. Identify the hidden variables (Z).
    a. Here, they are the locations of the motifs in each sequence.
3. Write the E step.
    a. Compute the expected values of the hidden variables given current parameter values.
4. Write the M step.
    a. Determine new parameters given the expected values of the hidden variables.
5. Repeat until convergence.



Sites in target sequences

Motif model

AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA

AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCA**AAACGGTT**CTTGCACGTA
GAGAT**AACCGCTT**GATATGACTCATTTGCC
ATATTCCGGACGCTGTGACG**ATCCGGTT**TG
GAACGC**AACCAGTT**CAGTGCTTATCATGAA

D'haeseleer (2016) Nat. Biotech.
Gitter @ U. Wisconsin

Motif-finding using MEME

- MEME: Multiple EM for Motif Elicitation
- A motif is:
    - assumed to have a fixed width, **W**
    - represented by a matrix of probabilities: $\mathbf{p_{c,k}}$ (probability of character **c** in column **k**).
- The "background" (i.e. sequence outside the motif) is given by $\mathbf{p_{c,0}}$ (probability of character **c** in the background).
- Data is a collection of sequences, denoted **X**.
- Motif starting positions are represented by a matrix indicator variables (0/1) $\mathbf{Z_{i,j}}$.

A motif model of length 3

$$p = \begin{array}{c|c|ccc} & 0 & 1 & 2 & 3 \\ \mathtt{A} & 0.25 & 0.1 & 0.5 & 0.2 \\ \mathtt{C} & 0.25 & 0.4 & 0.2 & 0.1 \\ \mathtt{G} & 0.25 & 0.3 & 0.1 & 0.6 \\ \mathtt{T} & 0.25 & 0.2 & 0.2 & 0.1 \end{array}$$

background

motif positions

Given sequences L = 6.
Possible starting positions m = L – W + 1

$$\begin{array}{l} \mathtt{G\ T\ \boxed{C\ A\ G}\ G} \\ \mathtt{\boxed{G\ A\ G}\ A\ G\ T} \\ \mathtt{A\ C\ G\ \boxed{G\ A\ G}} \\ \mathtt{C\ \boxed{C\ A\ G}\ T\ C} \end{array}$$

$$Z = \begin{array}{lcccc} & 1 & 2 & 3 & 4 \\ \text{seq1} & 0 & 0 & 1 & 0 \\ \text{seq2} & 1 & 0 & 0 & 0 \\ \text{seq3} & 0 & 0 & 0 & 1 \\ \text{seq4} & 0 & 1 & 0 & 0 \end{array}$$

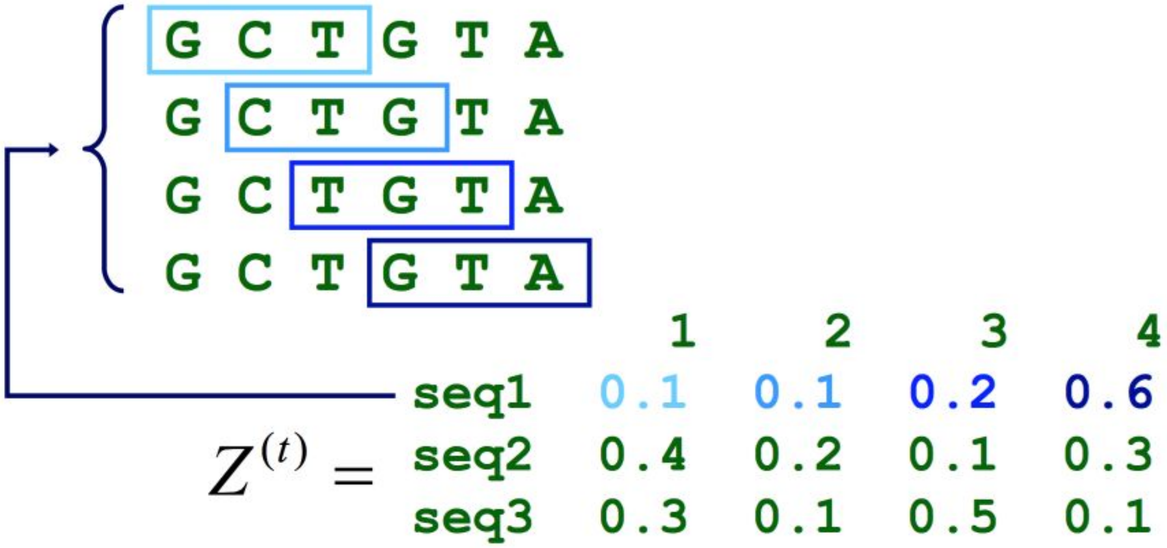
Motif-finding using MEME

1. Define the probabilistic model and the likelihood function $P(X \mid \theta)$.
2. Identify the hidden variables (Z).
   a. Here, they are the locations of the motifs in each sequence.
3. Write the E step.
   a. Compute the expected values of the hidden variables given current parameter values.
4. Write the M step.
   a. Determine new parameters given the expected values of the hidden variables.
5. Repeat until convergence.

```
given: length parameter W, set of sequences
  t=0
  set initial values for $p^{(0)}$
  do
    ++t
    re-estimate $Z^{(t)}$ from $p^{(t-1)}$ (E-step)
    re-estimate $p^{(t)}$ from $Z^{(t)}$ (M-step)
  until change in $p^{(t)} < \varepsilon$
return: $p^{(t)}$, $Z^{(t)}$
```

Motif-finding using MEME

- **E-step**: compute the expected values of Z given X and $p^{(t-1)}$
- Expected values: $Z^{(t)}$ □= E[Z | X, $p^{(t□-1)}$]
- For example:



$$Z^{(t)} = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \text{seq1} & 0.1 & 0.1 & 0.2 & 0.6 \\ \text{seq2} & 0.4 & 0.2 & 0.1 & 0.3 \\ \text{seq3} & 0.3 & 0.1 & 0.5 & 0.1 \end{array}$$

```
given: length parameter W, set of sequences
  t=0
  set initial values for $p^{(0)}$
  do
    ++t
    re-estimate $Z^{(t)}$ from $p^{(t-1)}$ (E-step)
    re-estimate $p^{(t)}$ from $Z^{(t)}$ (M-step)
  until change in $p^{(t)} < \varepsilon$
return: $p^{(t)}$, $Z^{(t)}$
```

Gitter @ U. Wisconsin

Motif-finding using MEME

- **<u>E-step</u>**: compute the expected values of Z given X and $p^{(t-1)}$
- Expected values: $Z^{(t)}$ □= E[Z | X, $p^{(t□-1)}$]
- Applying Bayes rule to: $P(Z_{i,j} = 1 \mid X_i, p^{(t-1)})$

$$Z_{i,j}^{(t)} = \frac{P(X_i \mid Z_{i,j} = 1, p^{(t-1)}) P(Z_{i,j} = 1)}{\sum_{k=1}^{m} P(X_i \mid Z_{i,k} = 1, p^{(t-1)}) P(Z_{i,k} = 1)}$$

```
given: length parameter W, set of sequences
  t=0
  set initial values for $p^{(0)}$
  do
    ++t
    re-estimate $Z^{(t)}$ from $p^{(t-1)}$ (E-step)
    re-estimate $p^{(t)}$ from $Z^{(t)}$ (M-step)
  until change in $p^{(t)} < \varepsilon$
return: $p^{(t)}$, $Z^{(t)}$
```

$$Z_{i,j}^{(t)} = \frac{P(X_i \mid Z_{i,j} = 1, p^{(t-1)})}{\sum_{k=1}^{m} P(X_i \mid Z_{i,k} = 1, p^{(t-1)})}$$

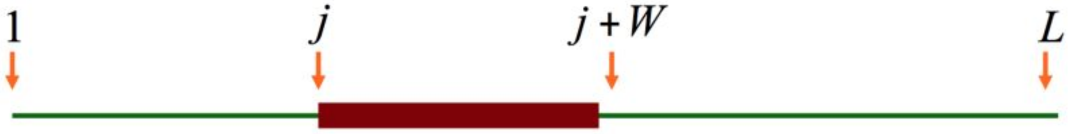
Assuming that it is equally likely that the motif will start in any position

$$P(Z_{i,j} = 1) = \frac{1}{m}$$

Gitter @ U. Wisconsin

Motif-finding using MEME

Probability of a Sequence Given a Motif Starting Position

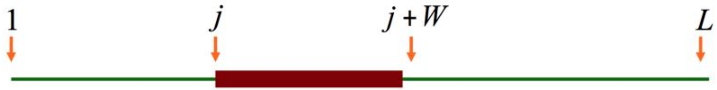


$$P(X_i \mid Z_{i,j}=1,p)=\underbrace{\prod_{k=1}^{j-1} p_{c_k,0}}_{\text{Before motif}}\ \underbrace{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}_{\text{Motif}}\ \underbrace{\prod_{k=j+W}^{L} p_{c_k,0}}_{\text{After motif}}$$

- **X_i** is the i th sequence
- **$Z_{i,j}$** is 1 if motif starts at position j in sequence i
- **c_k** is the base at position k in sequence i

Motif-finding using MEME

Probability of a Sequence Given a Motif Starting Position



$$P(X_i \mid Z_{i,j} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k,0}}_{\text{Before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}_{\text{Motif}} \underbrace{\prod_{k=j+W}^{L} p_{c_k,0}}_{\text{After motif}}$$

- X_i is the i th sequence
- $Z_{i,j}$ is 1 if motif starts at position j in sequence i
- c_k is the base at position k in sequence i

$$X_i = \text{G C } \boxed{\text{T G T}} \text{ A G}$$

$$p = \begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline \text{A} & 0.25 & 0.1 & 0.5 & 0.2 \\ \text{C} & 0.25 & 0.4 & 0.2 & 0.1 \\ \text{G} & 0.25 & 0.3 & 0.1 & 0.6 \\ \text{T} & 0.25 & 0.2 & 0.2 & 0.1 \end{array}$$

$$P(X_i \mid Z_{i,3} = 1, p) =$$

$$p_{\text{G},0} \times p_{\text{C},0} \times p_{\text{T},1} \times p_{\text{G},2} \times p_{\text{T},3} \times p_{\text{A},0} \times p_{\text{G},0} =$$

$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

$$P(X_i \mid Z_{i,1} = 1, p^{(t-1)}) \quad ?$$

Gitter @ U. Wisconsin

Motif-finding using MEME

- **M-step**: Estimate $p^{(t)}$ given X and $Z^{(t)}$.
- $p_{c,k}$ represents the probability of character c in position k.
- k=0 represents the background.

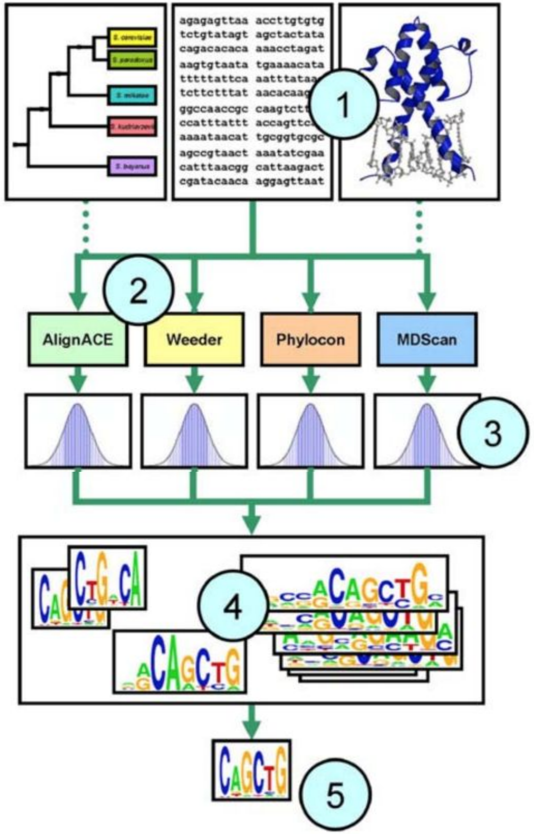
$$p^{(t)}_{c,k} = \frac{n_{c,k} + d_{c,k}}{\sum\limits_{b\in\{A,C,G,T\}} (n_{b,k} + d_{b,k})}$$

$$n_{c,k} = \begin{cases} \sum\limits_{i} \sum\limits_{\{j|X_{i,j+k-1}=c\}} Z^{(t)}_{i,j} & k > 0 \\ n_c - \sum\limits_{j=1}^{W} n_{c,j} & k = 0 \end{cases}$$

total # c's in the dataset

sum over positions where c appers

Practical strategies



1 **Assemble input data.** Results may be improved by restricting the input to high-confidence sequences. Some algorithms achieve improved performance by using phylogenetic conservation information from orthologous sequences or information about protein DNA-binding domains.

2 **Choose several motif discovery programs for the analysis.** For recommended programs see Figure 3.

3 **Test the statistical significance of the resulting motifs.** Use control calculations to estimate the empirical distribution of scores produced by each program on random data.

4 **Clustering and post-processing the motifs.** Motif discovery analyses often produce many similar motifs, which may be combined using clustering. Phylogenetic conservation information may be used to filter out statistically significant, but non-conserved motifs that are more likely to correspond to spurious sequence patterns.

5 **Interpretation of motifs.** Algorithms exist for linking motifs to transcription factors and for combining motif discovery with expression data.

MacIsaac & Fraenkel (2006) PLoS Comp. Biol.