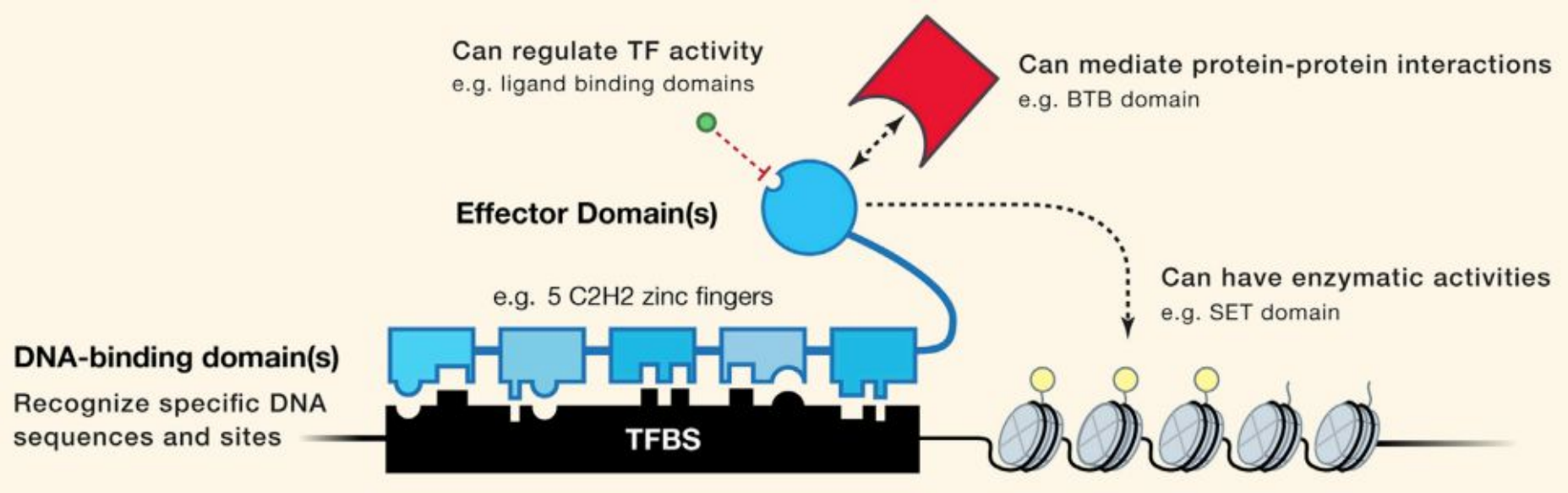
Lectures 12-13: Regulatory genomics

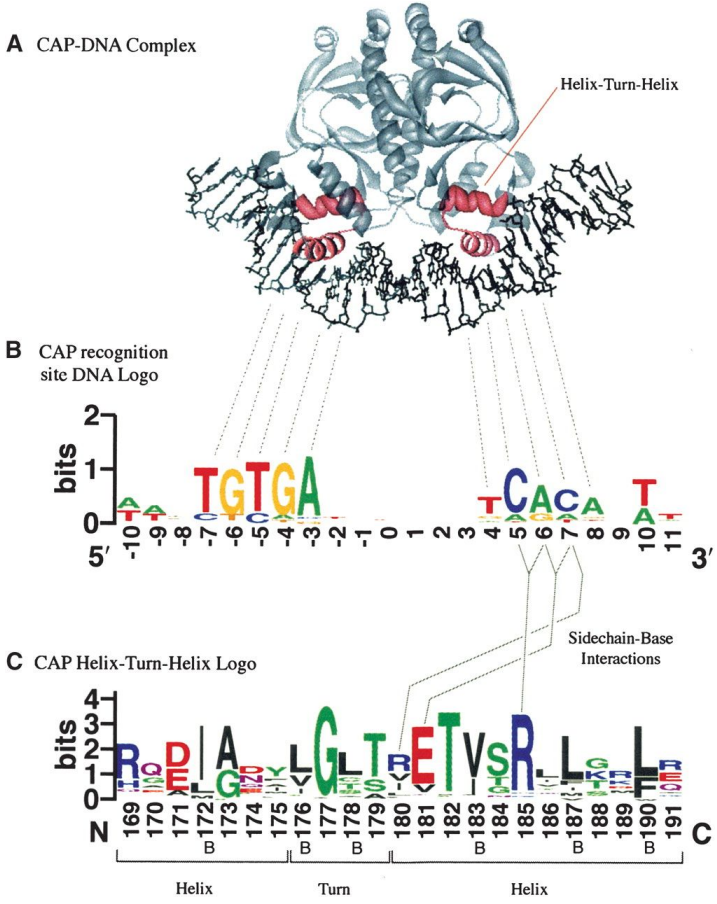
- DNA-binding sites/motifs
  - ChIP-seq
  - Position-weight matrices
  - Motif-finding
    - Expectation-Maximization
    - Gibbs Sampling

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.

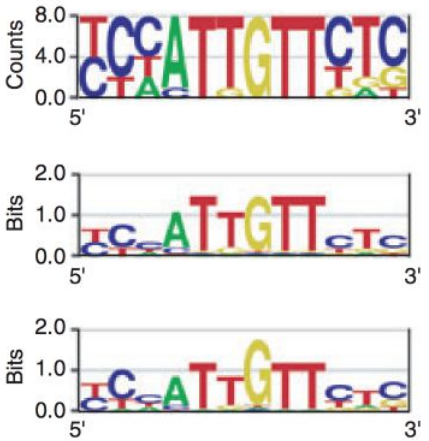
```
HEM13   CCCATTGTTCTC
HEM13   TTTCTGGTTCTC
HEM13   TCAATTGTTTAG
ANB1    CTCATTGTTGTC
ANB1    TCCATTGTTCTC
ANB1    CCTATTGTTCTC
ANB1    TCCATTGTTCGT
ROX1    CCAATTGTTTTG

        YCHATTGTTCTC
```

Motif instance ⟶ Motif

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

Position frequency matrix

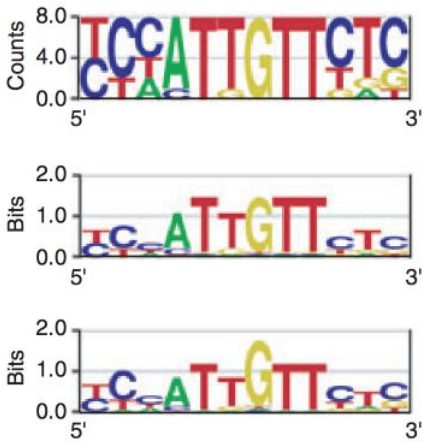


Sequence logo

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.

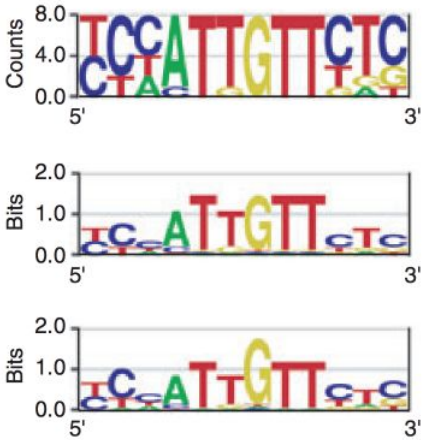
HindII bind to GTYRAC.

- What is its information content?

D'haeseleer (2016) Nat. Biotech.

Consensus sequence of DNA-binding sites

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

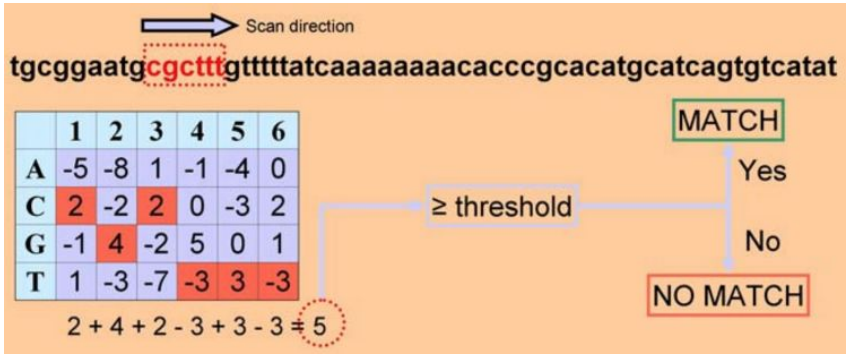


$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

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

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

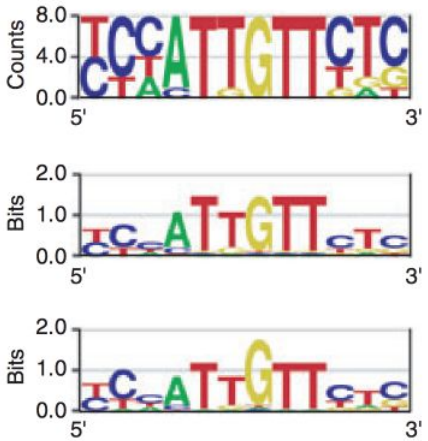
Position weight matrix (PWM).



D'haeseleer (2016) Nat. Biotech.

Consensus sequence of DNA-binding sites

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

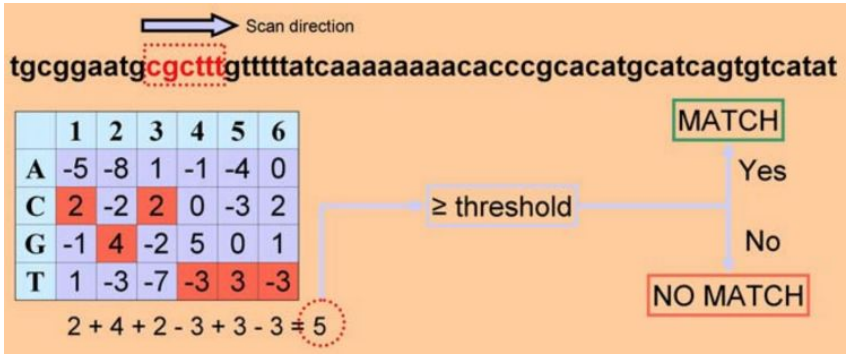


A generative model!

Assumptions:

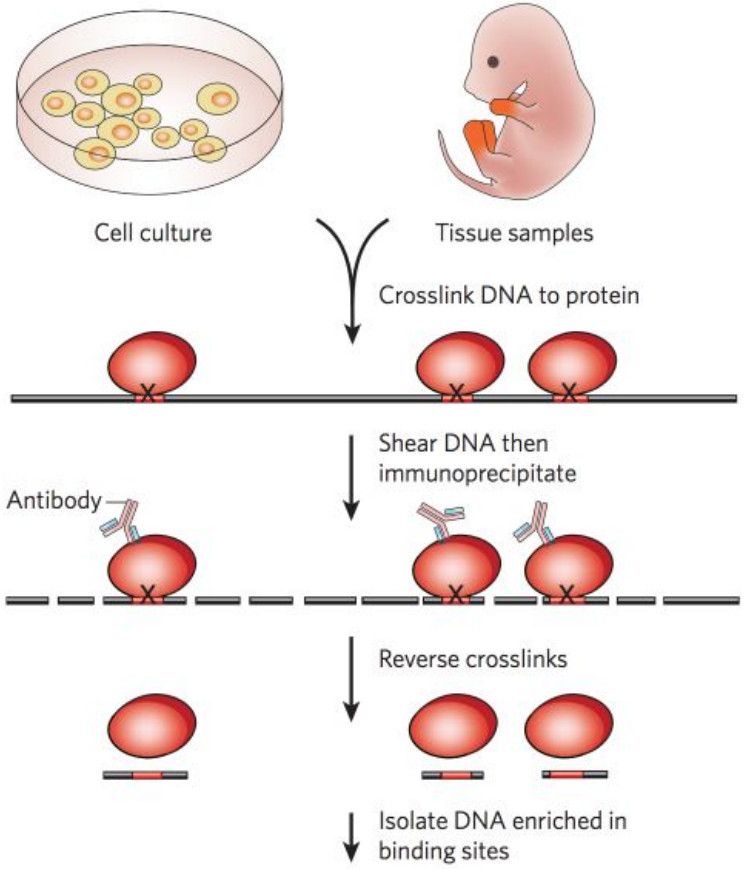
- Independence of positions
- Fixed spacing

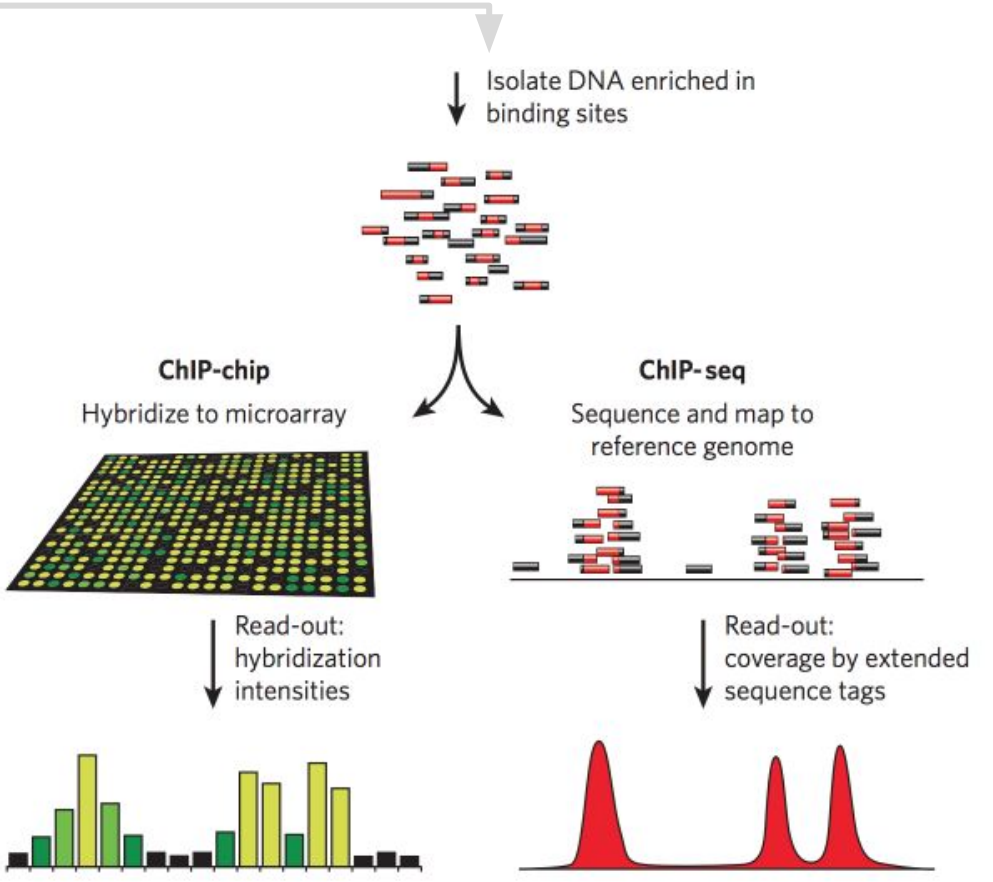
Position weight matrix (PWM).



D'haeseleer (2016) Nat. Biotech.

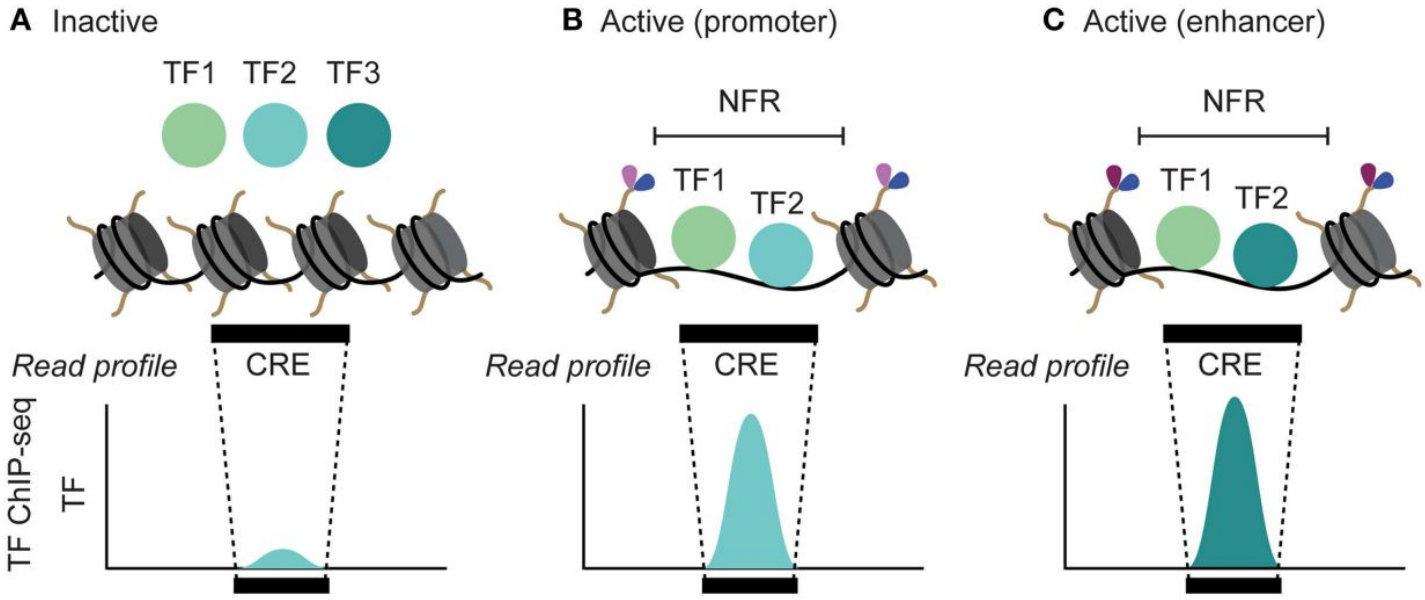
Mapping of regulatory elements using ChIP-chip and ChIP-seq





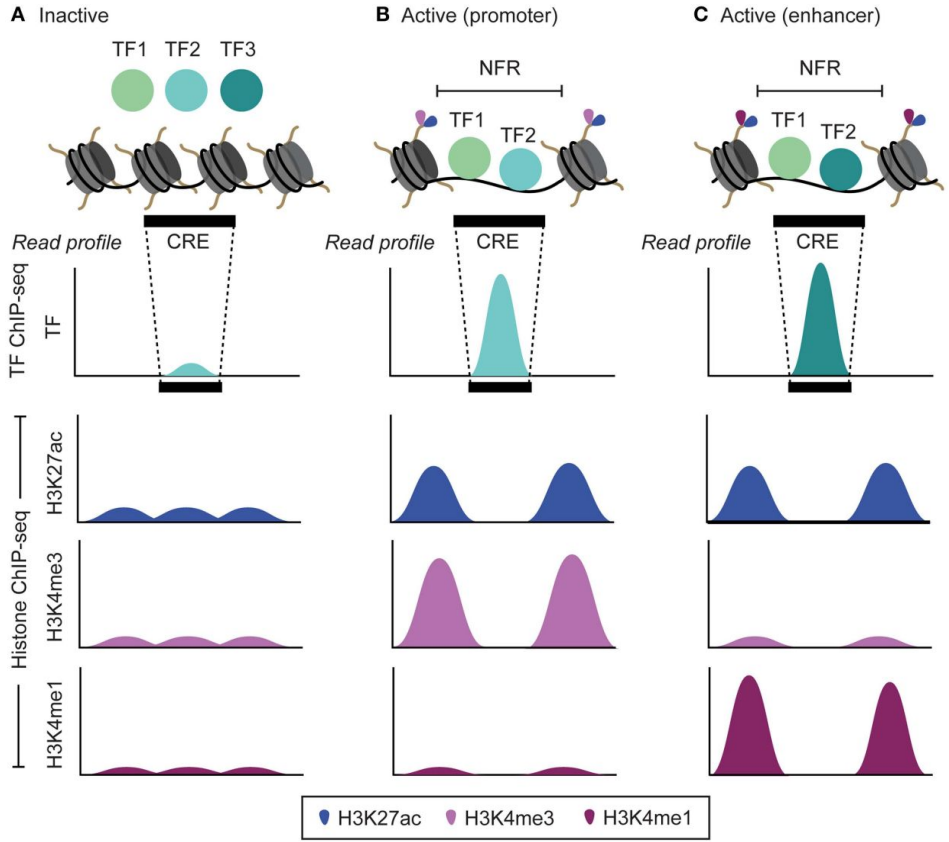
Visel (2009) Nature

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



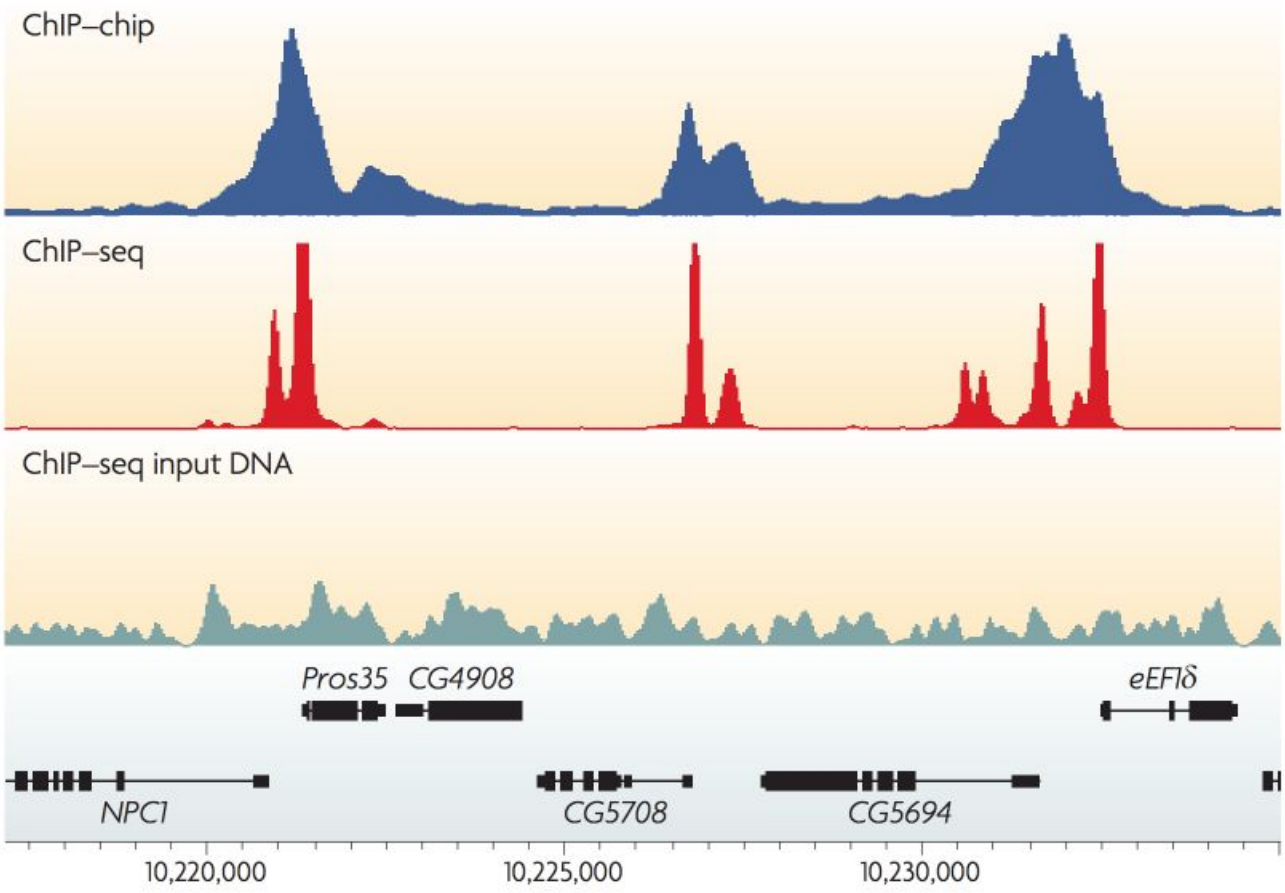
Pundhir (2015) Front. Genet.

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.

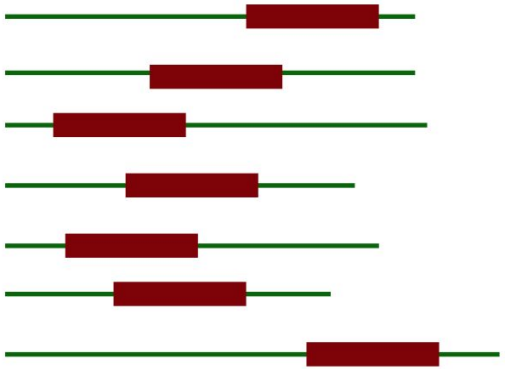
Mapping of regulatory elements using ChIP-chip and ChIP-seq

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:

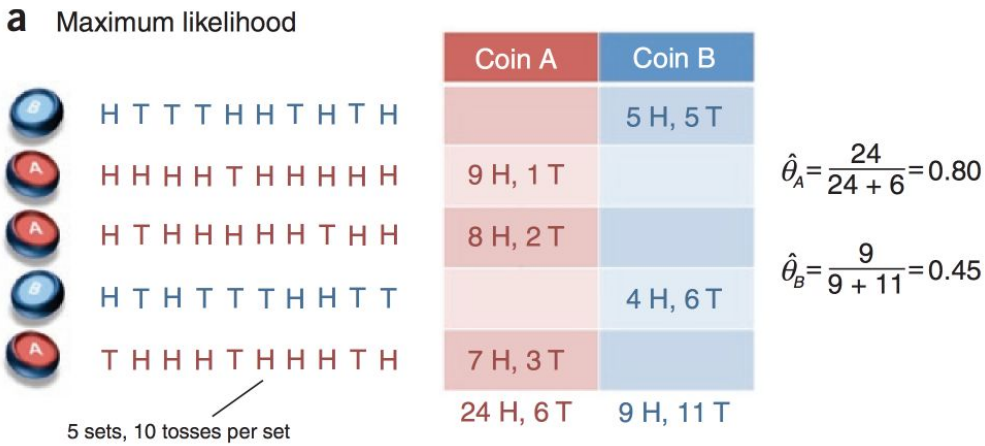
- Given: a set of sequences that are thought to contain occurrences of an unknown motif of interest
- Do:
  - infer a model (PWM) of the motif, and
  - predict the locations of the motif occurrences in the given sequences.



Expectation-Maximization: Iteratively refine positions / motif profile

Gibbs sampling: Iteratively sample positions / motif profile

Expectation-Maximization algorithm (EM)



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.

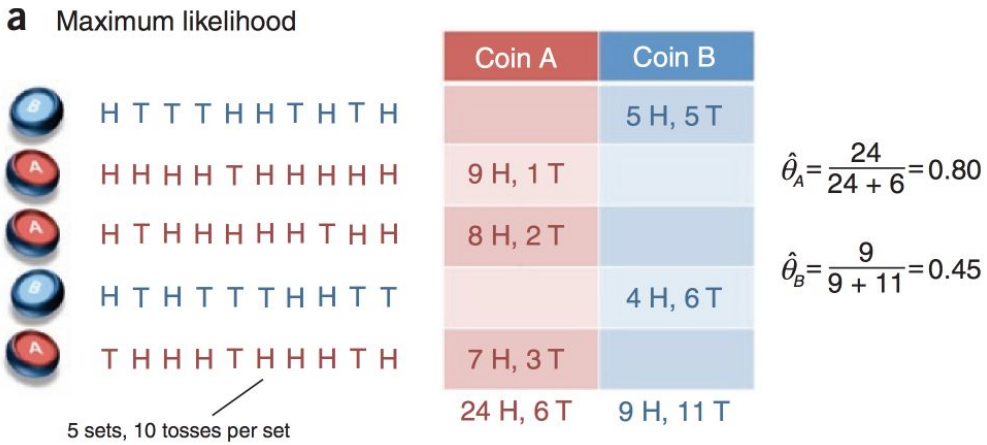
$x = (x_1, x_2, \ldots, x_5) \mid x_i \in \{0,1,\ldots,10\}$ is the no. of heads observed during the ith set of tosses.

$z = (z_1, z_2, \ldots, z_5) \mid z_i \in \{A,B\}$ 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.

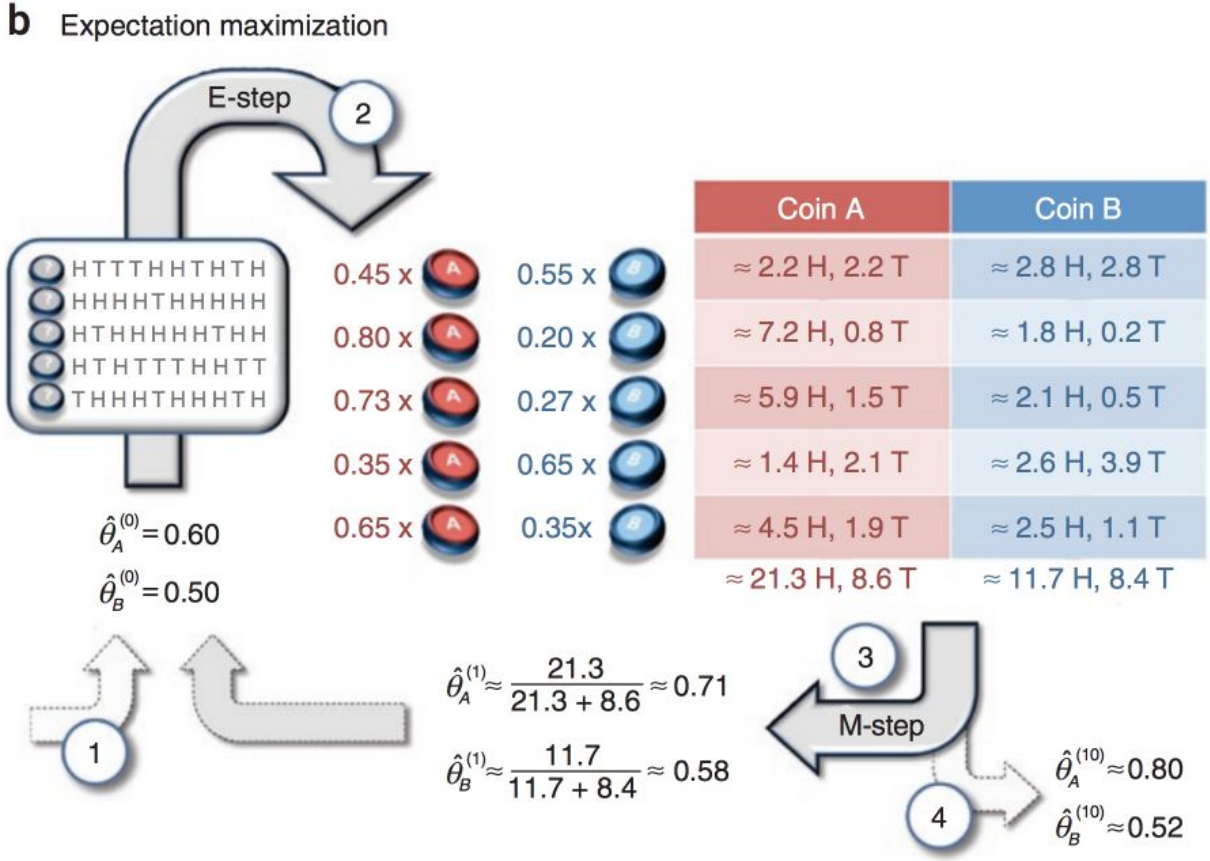
$z = (z_1, z_2, \ldots, z_5) \mid z_i \in \{A,B\}$ is the identity of the coin used during the ith set of tosses. [**Hidden variables / Latent factors**]

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), 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)



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.

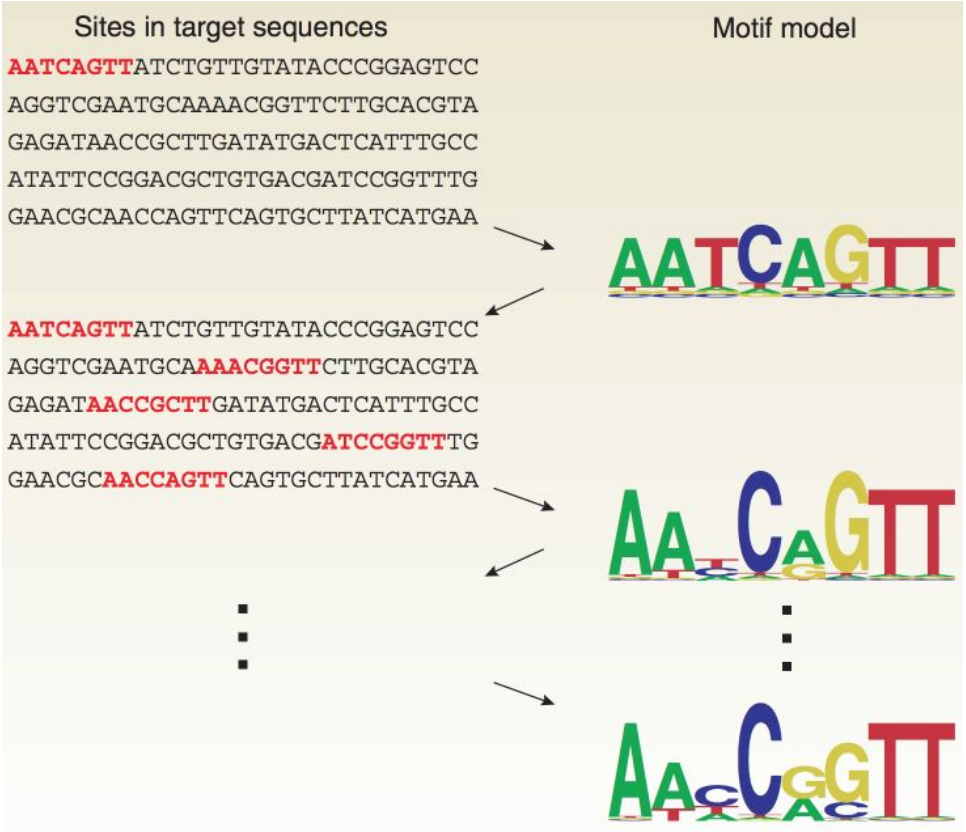
Repeat until convergence.

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



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

Expectation-Maximization algorithm (EM)

The **likelihood** of a model is the probability that the observed data could have been generated by the model under consideration.

- Easier to optimize the logarithm of this probability (hence '**log likelihood**') with respect to the parameters of the model:

$$\log L(\text{model} \mid \text{data}) = \log \Pr(\text{data} \mid \text{model})$$
$$= \sum_i \log \Pr(\text{data}_i \mid \text{model})$$

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 base **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

$p =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

background | motif positions

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

G T C A G G
G A G A G T
A C G G A G
C C A G T C

$Z =$

	1	2	3	4
seq1	0	0	1	0
seq2	1	0	0	0
seq3	0	0	0	1
seq4	0	1	0	0

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)

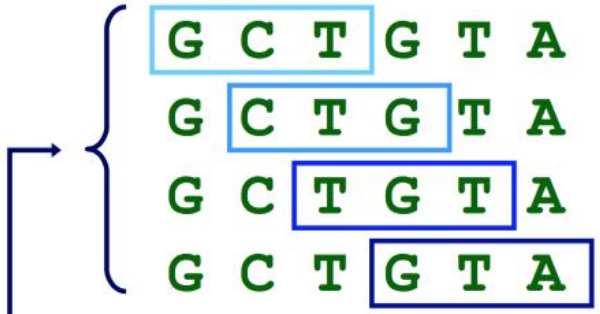
re-estimate $p^{(t)}$ from $Z^{(t)}$ (M-step)

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:



$Z^{(t)} =$

	1	2	3	4
seq1	0.1	0.1	0.2	0.6
seq2	0.4	0.2	0.1	0.3
seq3	0.3	0.1	0.5	0.1

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: $p^{(t)}$, $Z^{(t)}$

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

?

Motif-finding using MEME

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

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: $p^{(t)}$, $Z^{(t)}$

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

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

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

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

Motif-finding using MEME

Probability of a Sequence Given a Motif Starting Position

$$P(X_i \mid Z_{i,j} = 1, p) = \prod_{k=1}^{j-1} p_{c_k,0} \prod_{k=j}^{j+W-1} p_{c_k,k-j+1} \prod_{k=j+W}^{L} p_{c_k,0}$$

Before motif | Motif | 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

$$P(X_i \mid Z_{i,j} = 1, p) = \prod_{k=1}^{j-1} p_{c_k,0} \prod_{k=j}^{j+W-1} p_{c_k,k-j+1} \prod_{k=j+W}^{L} p_{c_k,0}$$

Before motif | Motif | After motif

$$X_i = \mathtt{G\ C\ T\ G\ T\ A\ G}$$

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

- 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

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

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

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

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

Motif-finding using MEME

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

$X_i =$ **G C T G T A G**

$p =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

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

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: $p^{(t)}$, $Z^{(t)}$

$$Z^{(t)}_{i,1} \propto P(X_i \mid Z_{i,1} = 1, p^{(t-1)}) = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z^{(t)}_{i,2} \propto P(X_i \mid Z_{i,2} = 1, p^{(t-1)}) = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

…

…

…

Normalize so that $\sum_{j=1}^{m} Z^{(t)}_{i,j} = 1$

Motif-finding using MEME

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

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

A C A G C A

$$Z^{(t)}_{1,1} = 0.1,\ Z^{(t)}_{1,2} = 0.7,\ Z^{(t)}_{1,3} = 0.1,\ Z^{(t)}_{1,4} = 0.1$$

A G G C A G

$$Z^{(t)}_{2,1} = 0.4,\ Z^{(t)}_{2,2} = 0.1,\ Z^{(t)}_{2,3} = 0.1,\ Z^{(t)}_{2,4} = 0.4$$

T C A G T C

$$Z^{(t)}_{3,1} = 0.2,\ Z^{(t)}_{3,2} = 0.6,\ Z^{(t)}_{3,3} = 0.1,\ Z^{(t)}_{3,4} = 0.1$$

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: $p^{(t)}$, $Z^{(t)}$

Motif-finding using MEME

M-step requires joint likelihood

$$\log P(X, Z \mid p) = \log \prod_i P(X_i, Z_i \mid p)$$

$$= \log \prod_i P(X_i \mid Z_i, p) P(Z_i \mid p)$$

$$= \log \prod_i \tfrac{1}{m} \prod_j P(X_i \mid Z_{i,j} = 1, p)^{Z_{i,j}}$$

$$= \sum_i \sum_j Z_{i,j} \log P(X_i \mid Z_{i,j} = 1, p) + n \log \tfrac{1}{m}$$

Motif-finding using MEME

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

$$p_{c,k}^{(t)} = \frac{n_{c,k} + d_{c,k}}{\sum_{b \in \{A,C,G,T\}} (n_{b,k} + d_{b,k})}$$

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

total # c's in the dataset

sum over positions where c appears

Motif-finding using MEME

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

A C A G C A

$$Z^{(t)}_{1,1} = 0.1,\ Z^{(t)}_{1,2} = 0.7,\ Z^{(t)}_{1,3} = 0.1,\ Z^{(t)}_{1,4} = 0.1$$

A G G C A G

$$Z^{(t)}_{2,1} = 0.4,\ Z^{(t)}_{2,2} = 0.1,\ Z^{(t)}_{2,3} = 0.1,\ Z^{(t)}_{2,4} = 0.4$$

T C A G T C

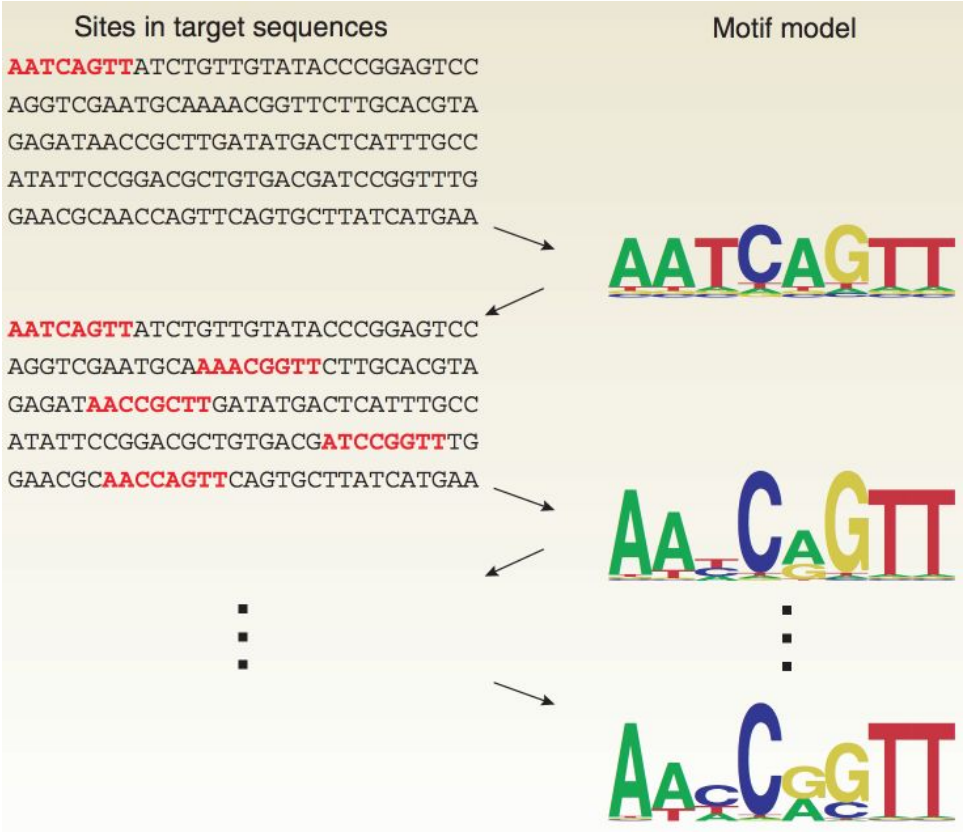
$$Z^{(t)}_{3,1} = 0.2,\ Z^{(t)}_{3,2} = 0.6,\ Z^{(t)}_{3,3} = 0.1,\ Z^{(t)}_{3,4} = 0.1$$

$$p^{(t)}_{\mathrm{A},1} = \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,3} + Z^{(t)}_{2,1} + Z^{(t)}_{3,3} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} \ldots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 4}$$

$$p^{(t)}_{\mathrm{C},2} =$$

Expectation-Maximization algorithm (EM)

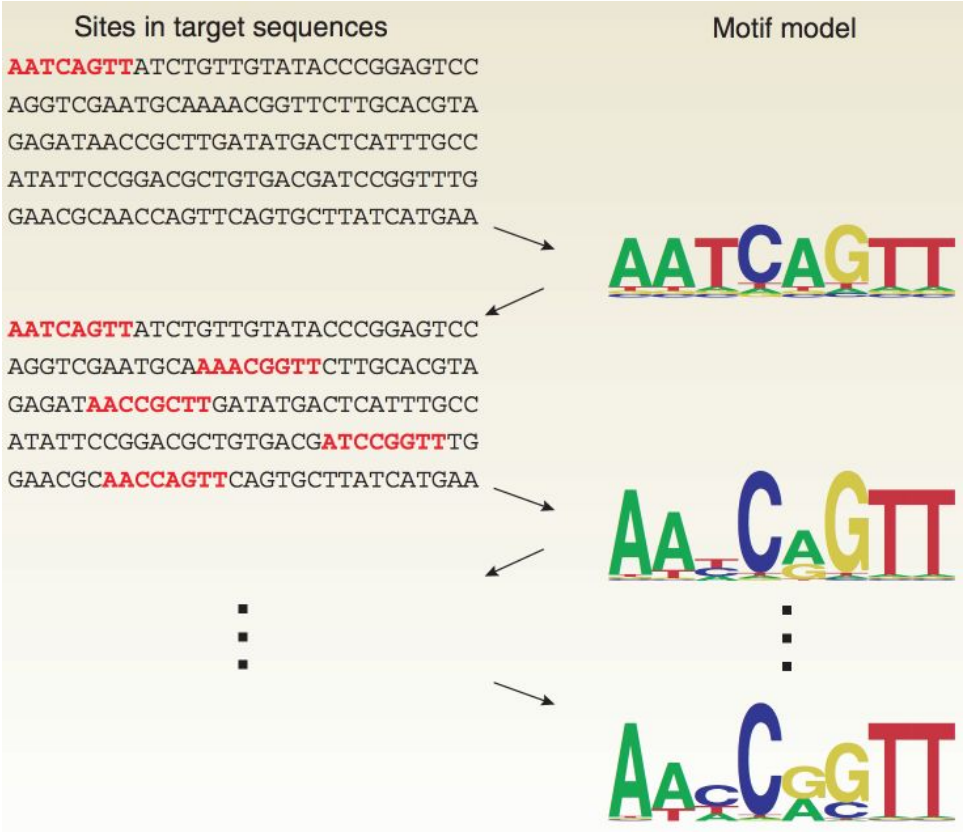
1. Define the probabilistic model and the likelihood function P(X | θ).
2. Identify the hidden variables (Z).
    a. Here, they are the locations of the motifs in each sequence.
3. Write the **E step**.
    a. Compute the expected values of the hidden variables given current parameter values.
4. Write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



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

Expectation-Maximization algorithm (EM)

1. Assume zero or more motif occurrences per sequence.
2. Choosing the width of the motif.
3. Finding multiple motifs in a group of sequences.
4. Choosing good starting points for the parameters.
5. Using background knowledge to bias the parameters.

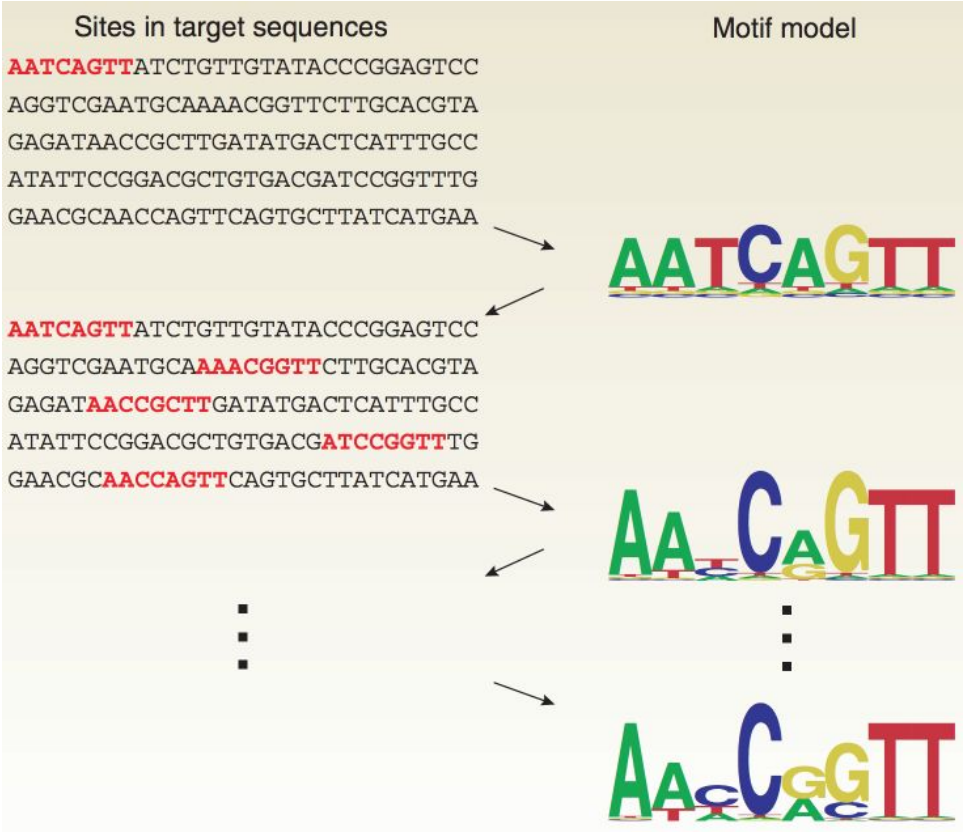


D'haeseleer (2016) Nat. Biotech.

Motif-finding using MEME

MEME:

- EM is susceptible to local maxima; so, try multiple starting points.
- Motif must be similar to some subsequence in data set
- For every distinct subsequence of length W in the training set
  - derive an initial p matrix from this subsequence
  - run EM for 1 iteration
- Choose motif model (i.e. p matrix) with highest likelihood.
- Run EM to convergence.

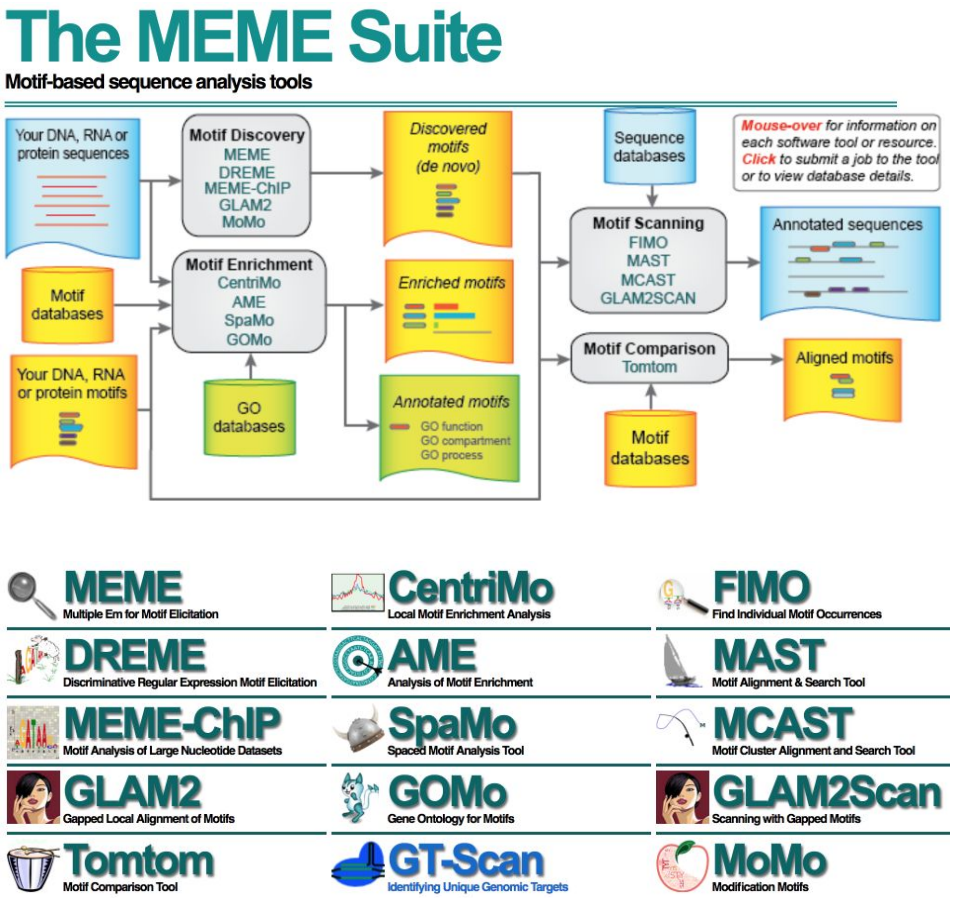


D'haeseleer (2016) Nat. Biotech.
Gitter @ U. Wisconsin

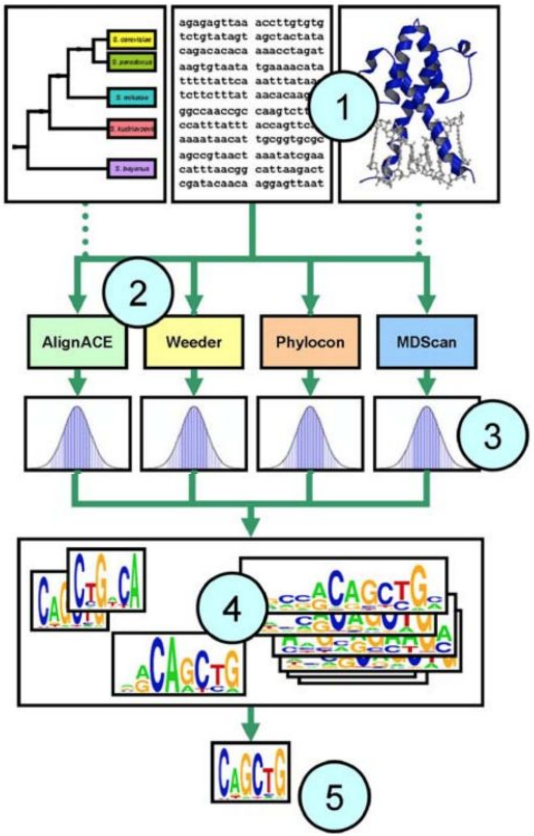
Motif-finding using MEME

MEME:

- Lawrence & Reilly (1990) “An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences", *Proteins*.
- Bailey & Elkan (1994) "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*.
- http://meme-suite.org/



Practical strategies for finding motifs



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.