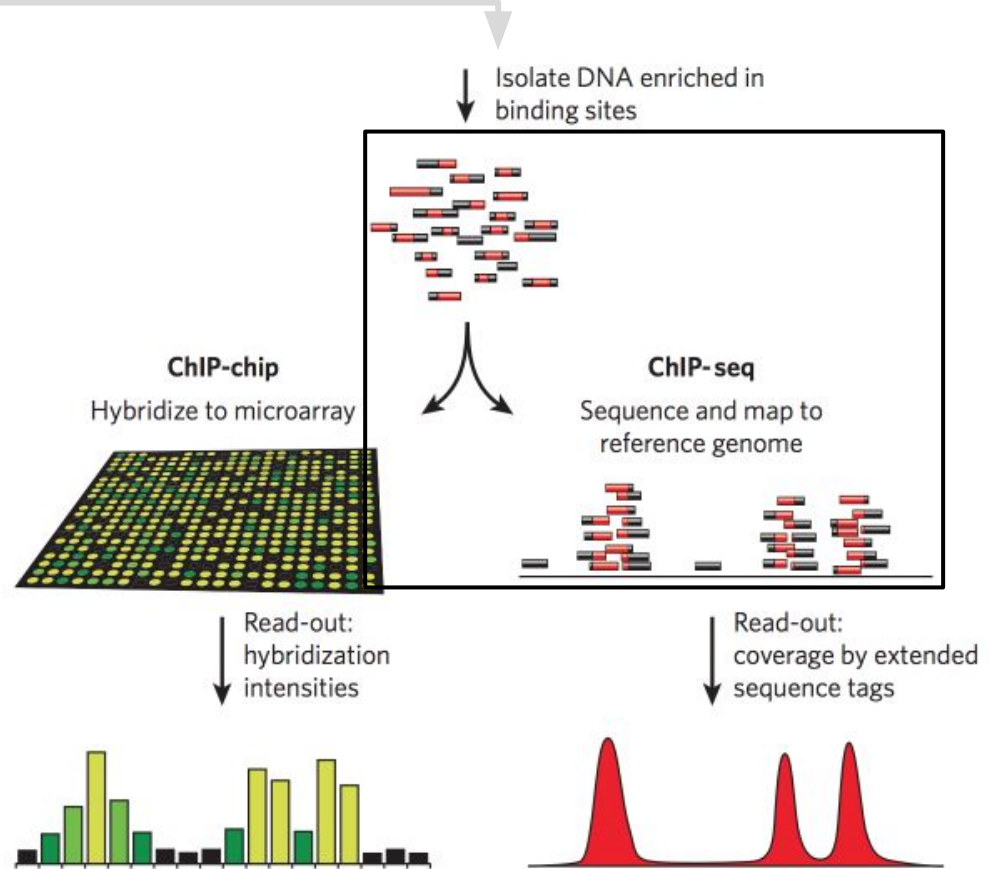
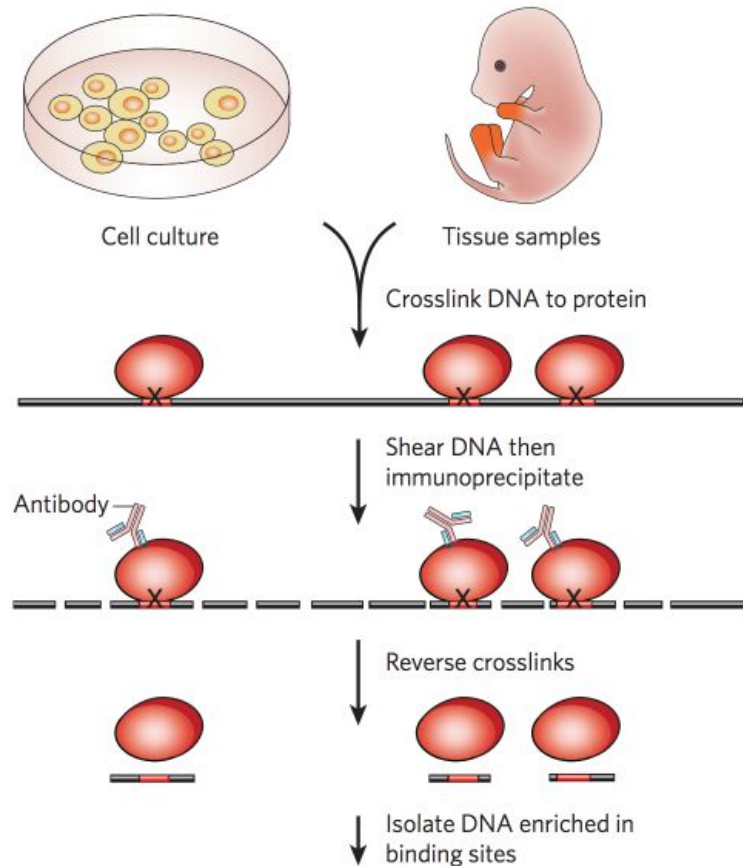


Week 06: Regulatory genomics

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

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



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 Maximum likelihood



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.

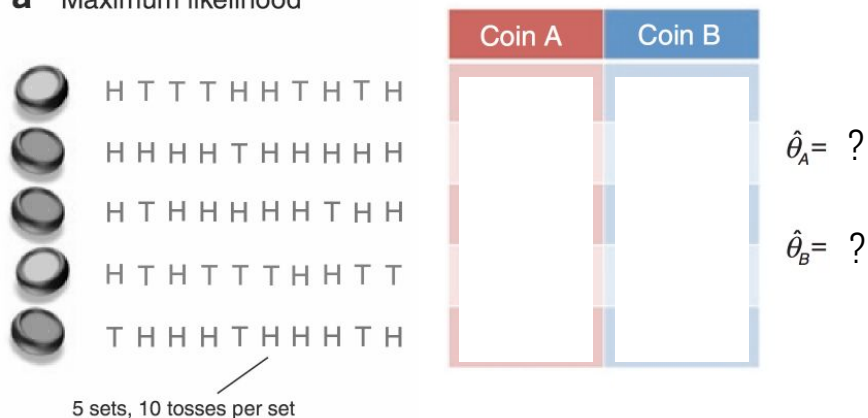
$x = (x_1, x_2, \dots, x_5) \mid x_i \in \{0, 1, \dots, 10\}$ is the no. of heads observed during the i th set of tosses.

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

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

Expectation-Maximization algorithm (EM)

a Maximum likelihood



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

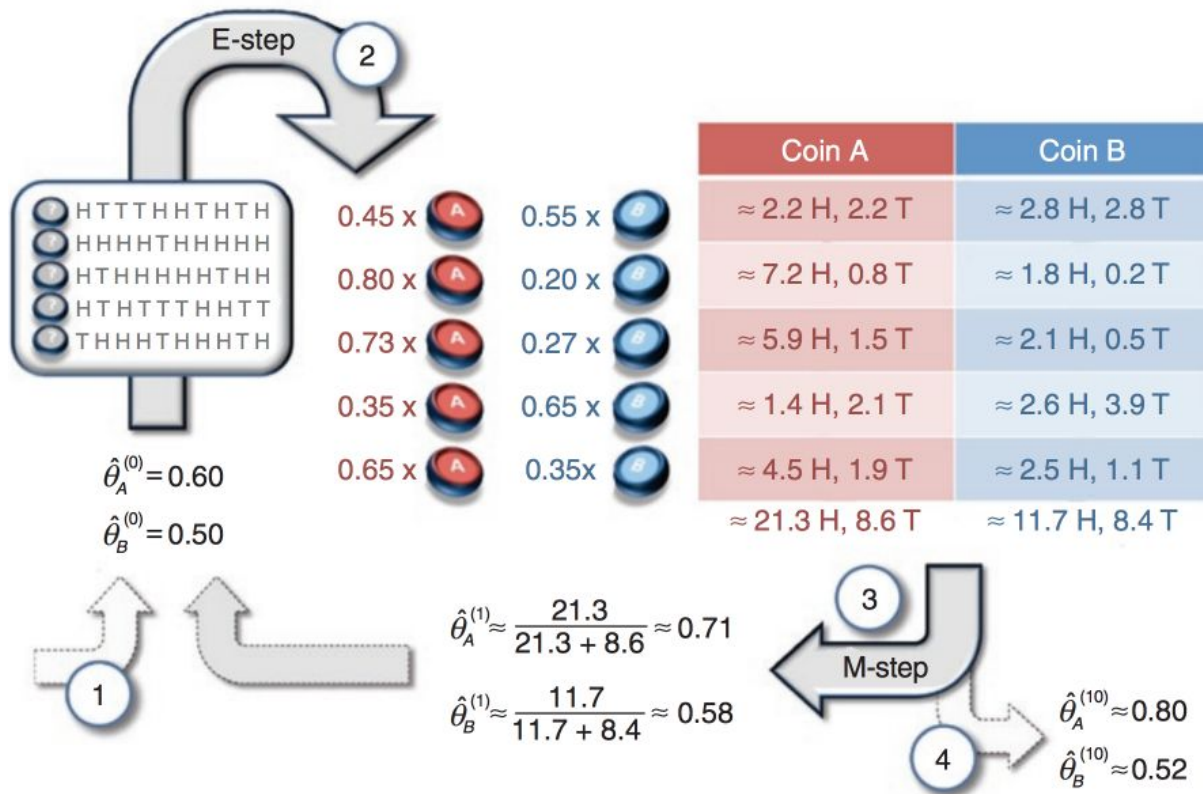
$z = (z_1, z_2, \dots, z_5) \mid z_i \in \{A, B\}$ is the identity of the coin used during the i th set of tosses. [Hidden variables / Latent factors]

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 know which coin was chosen.**)
 - Perform ten independent tosses with the selected coin.

Expectation-Maximization algorithm (EM)

b Expectation maximization



E-step:

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

M-step:

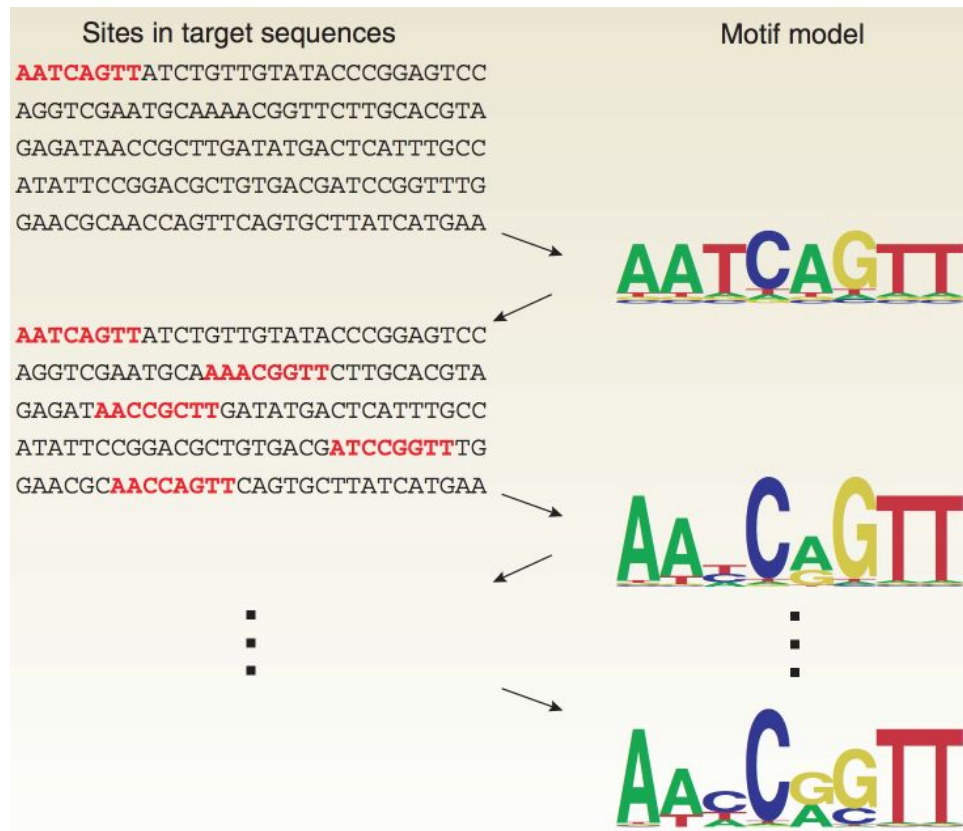
- Estimate new parameters $\theta^{(t+1)}$ given current estimates of hidden variables & parameters.

Repeat until convergence.

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

Expectation-Maximization algorithm (EM)

1. Define the probabilistic model and the likelihood function $P(X | \theta)$.
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.



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: $p_{c,k}$ (probability of character c in column k).
- The "background" (i.e. sequence outside the motif) is given by $p_{c,0}$ (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) Z_{ij} .

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 $m = L - W + 1$

$Z =$

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

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

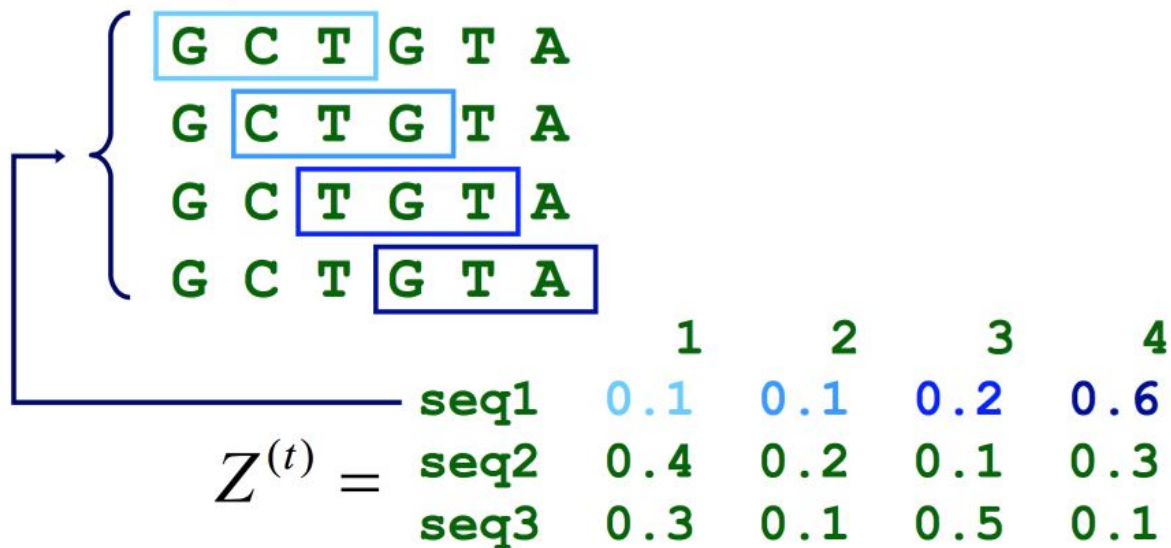
Motif-finding using MEME

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: width 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  $p^{(t)} < \epsilon$ 
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: $Z^{(t)} = E[Z \mid X, p^{(t-1)}]$
- For example:



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 $p^{(t)} < \epsilon$

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

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

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

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 $p^{(t)} < \epsilon$

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

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



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

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



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

Before motif

Motif

After motif

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

	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

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

$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} = 0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

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

- 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

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

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

		0	1	2	3
$p =$	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

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

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

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

...

...

...

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

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 $p^{(t)} < \epsilon$

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: $Z^{(t)} = E[Z \mid X, p^{(t-1)}]$

		0	1	2	3
$p =$	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

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 $p^{(t)} < \epsilon$

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

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

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

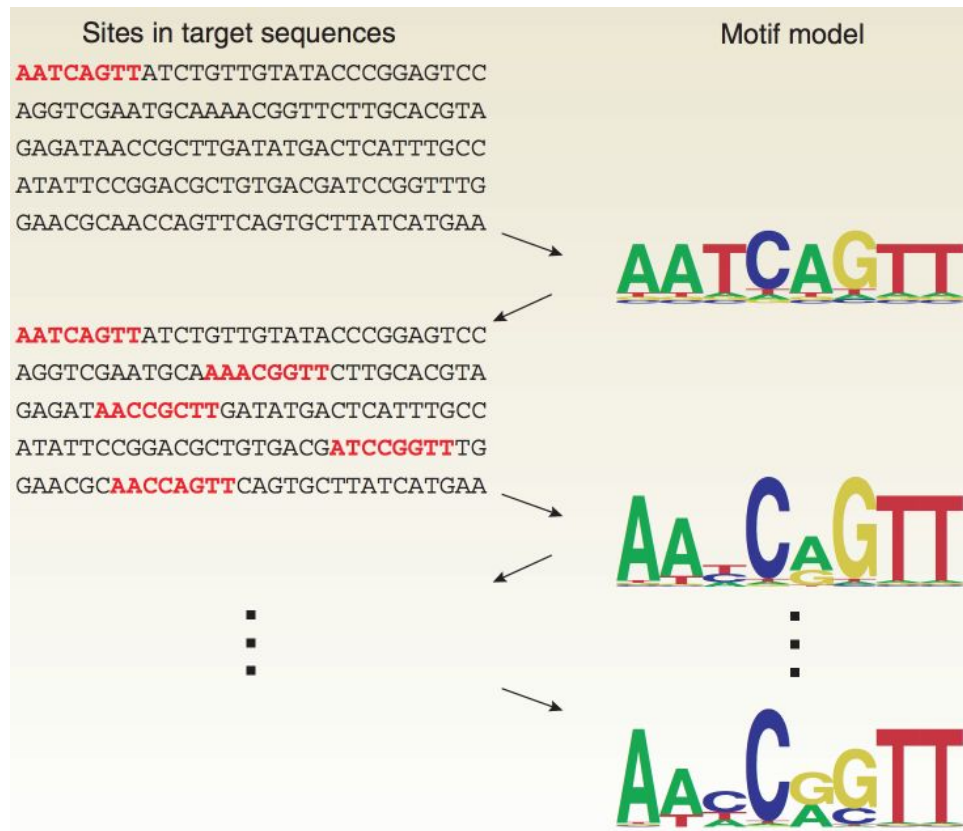
$$p^{(t)}_{A,1} = \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,3} + Z^{(t)}_{2,1} + Z^{(t)}_{3,3} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} \dots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 4}$$

$$p^{(t)}_{C,2} = \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,4} + Z^{(t)}_{2,3} + Z^{(t)}_{3,1} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} \dots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 4}$$

⋮

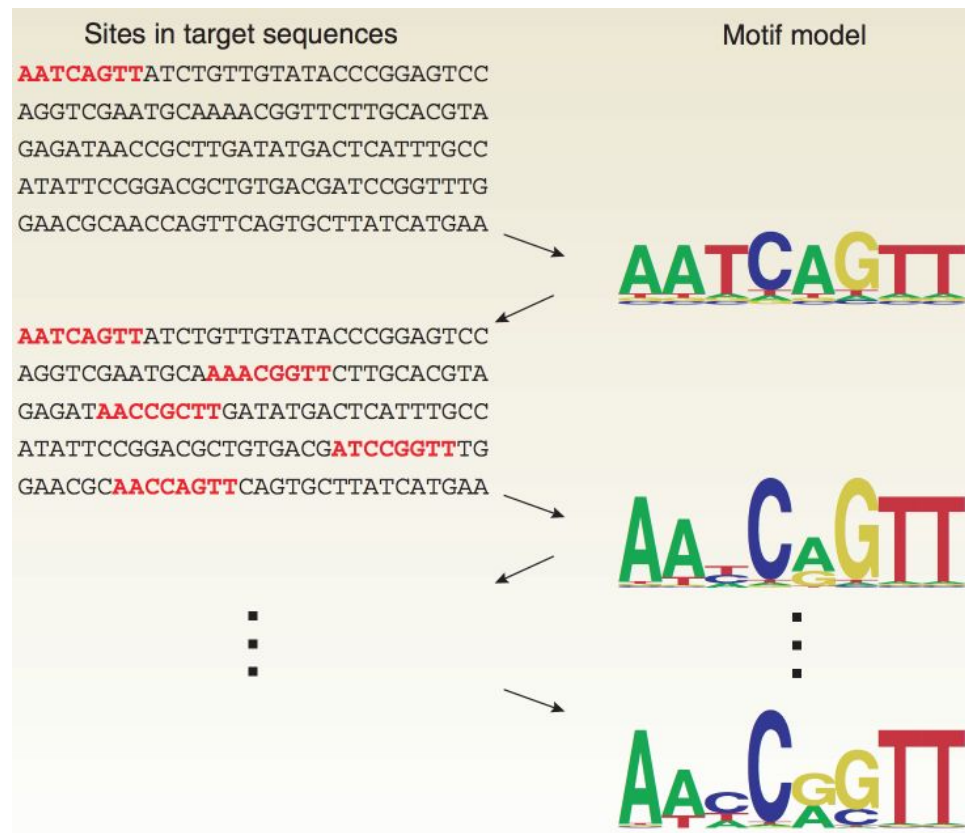
Expectation-Maximization algorithm (EM)

1. Define the probabilistic model and the likelihood function $P(X | \theta)$.
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.



Expectation-Maximization algorithm (EM)

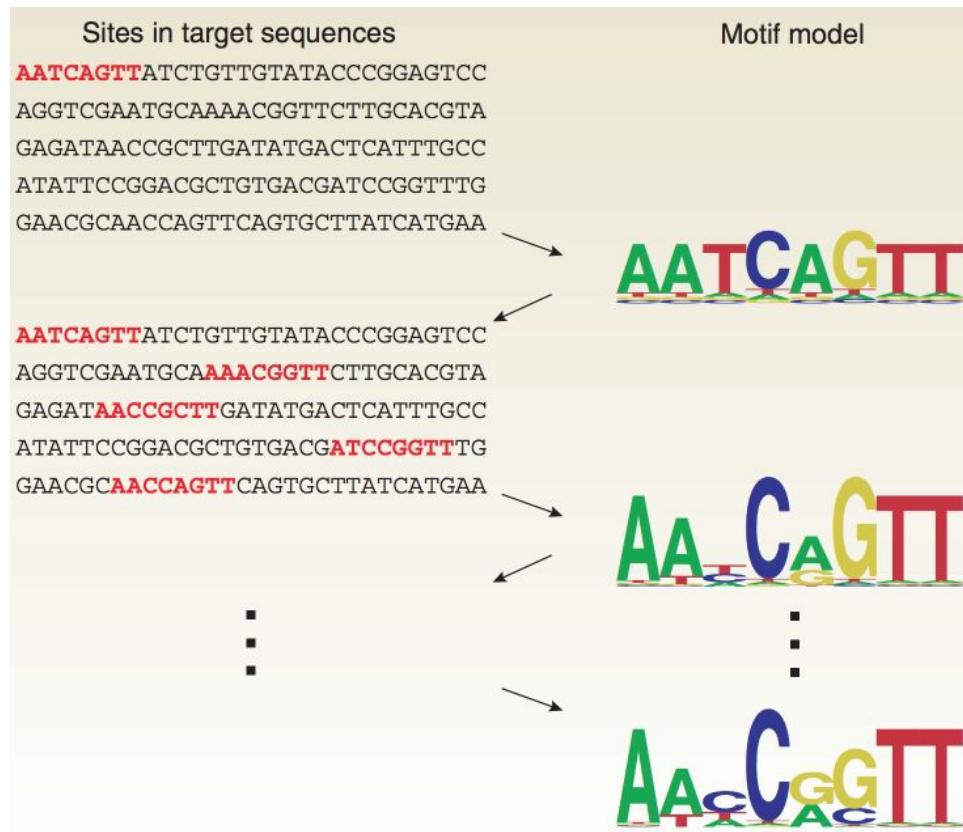
1. Assume zero or more motif occurrences per sequence.
2. Choosing the width of the motif.
3. Finding multiple motifs in a group of sequences.
4. Choosing good starting points for the parameters.
5. Using background knowledge to bias the parameters.



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.



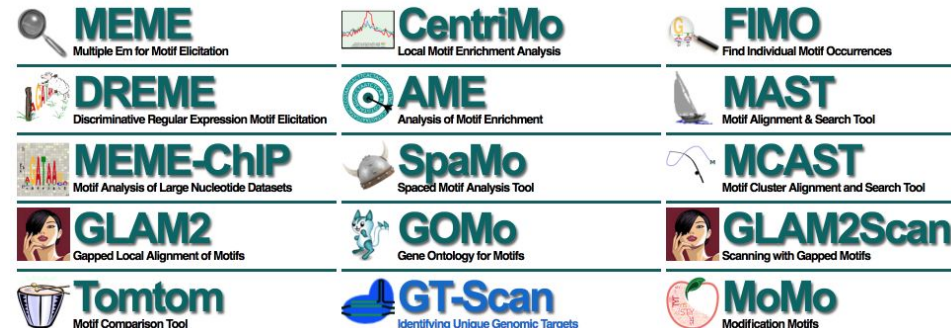
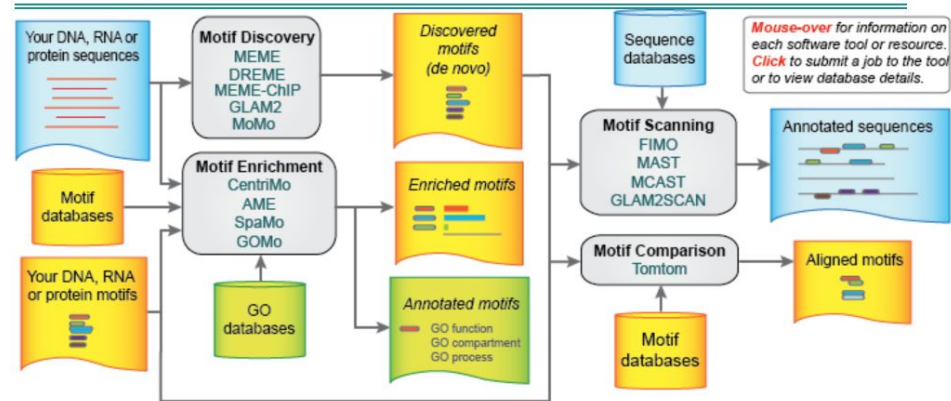
Motif-finding using MEME

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/>

The MEME Suite

Motif-based sequence analysis tools



Motif finding using Gibbs sampling instead of EM

EM can get trapped in local minima

- One approach to alleviate this limitation: try different (perhaps random) initial parameters.

Gibbs sampling exploits randomized search to a much greater degree:

- Can be viewed as a stochastic analog of EM for this task.
- In theory, Gibbs sampling is less susceptible to local minima than EM.

Motif finding using Gibbs sampling

- A motif is:
 - assumed to have a fixed width, W
 - represented by a matrix of probabilities: $p_{c,k}$ (probability of character c in column k).
- The "background" (i.e. sequence outside the motif) is given by $p_{c,0}$ (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) $Z_{i,j}$.

Motif
($W = 3$)

$p =$

	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

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 Gibbs sampling

1. Choose initial \mathbf{Z} containing the motif starting position in each sequence at random.
2. Loop through each sequence \mathbf{X}_i :
 - a. Update \mathbf{p} (position frequency matrix of background + motif) based on all sequences except \mathbf{X}_i .
 - b. Based on the *updated* \mathbf{p} , calculate the location of best match in sequence \mathbf{X}_i and update the corresponding row in \mathbf{Z} .
3. Repeat until convergence.

$Z =$						
			1	2	3	4
G	T	C	A	G	G	
G	A	G	A	G	T	seq1
A	C	G	G	A	G	seq2
C	C	A	G	T	C	seq3
						seq4
			0	0	1	0
			1	0	0	0
			0	0	0	1
			0	1	0	0

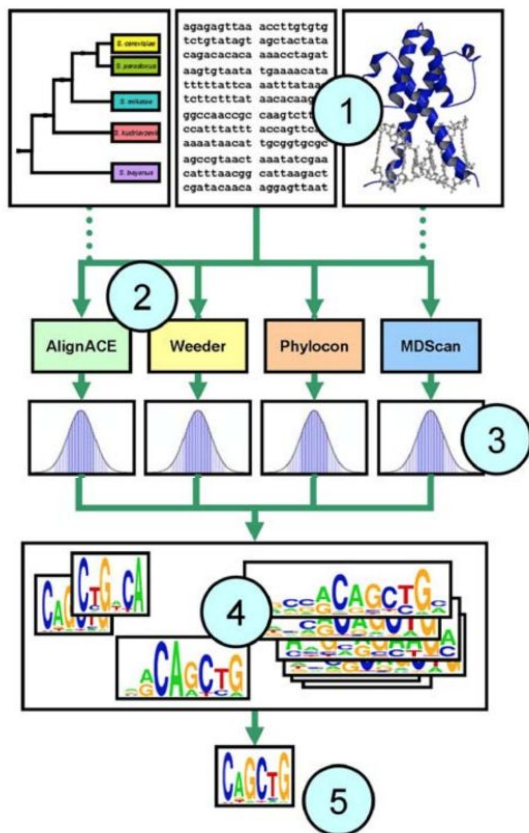
$$p_{c,k}^{(t)} = \frac{n_{c,k} + d_{c,k}}{\sum_{b \in \{A,C,G,T\}} (n_{b,k} + d_{b,k})}$$

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

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

$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} = 0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

Practical strategies for finding motifs



Assemble input data. Results may be improved by restricting the input to high-confidence sequences.

- 1 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.