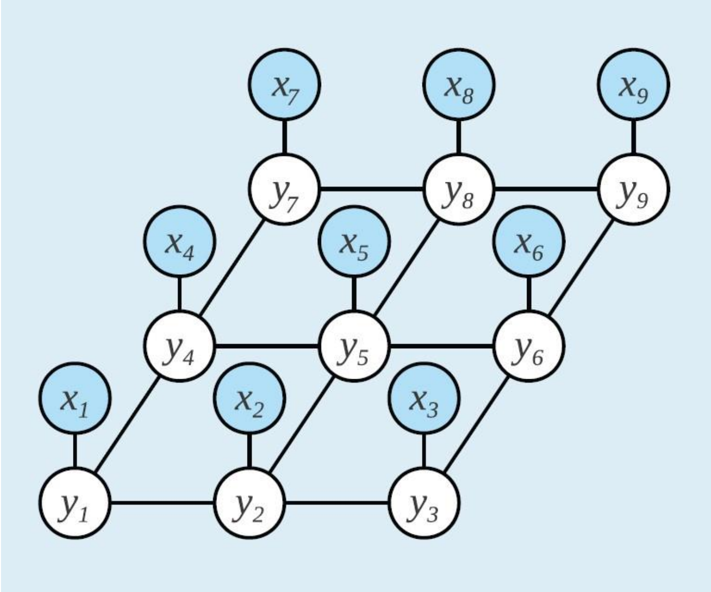
Probabilistic Graphical Models in Bioinformatics

Lecture 10: Expectation-maximization; regulatory motif finding

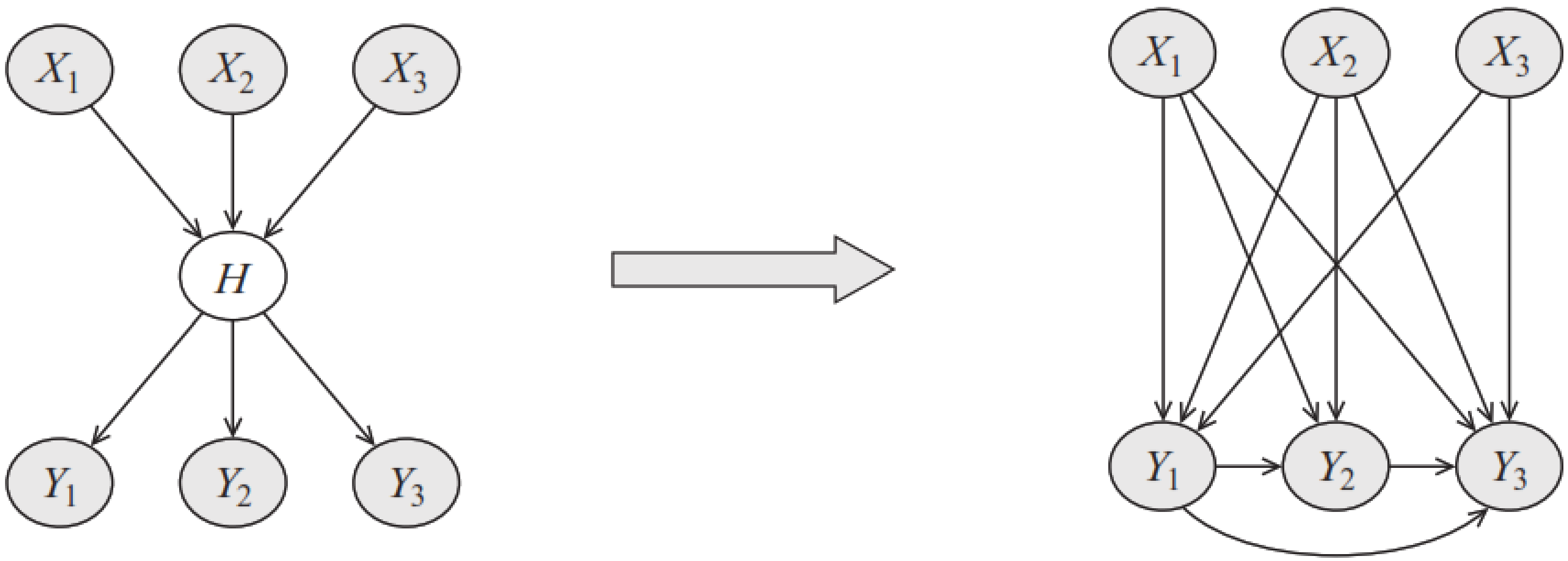


Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

Ordibehesht 18, 1398
(May 8, 2019)

	Known structure (Parameter estimation)	**Unknown structure (Structure learning)**
Fully observable	*MLE* *Bayesian methods*	*Constraint-based methods* *Score-based methods e.g. hill climbing*
Partially observable		

Hidden variables can greatly simplify the structure

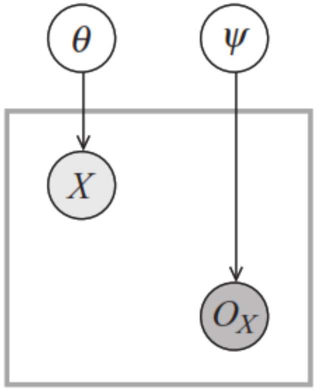


17 parameters

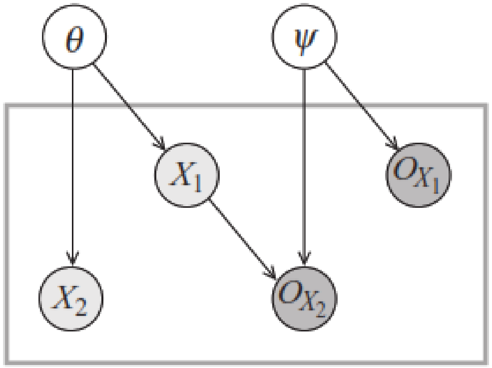
59 parameters

Observation mechanism

- **Missing completely at random (MCAR):** the observation mechanism is completely independent of observed and hidden variables.
  - The missing data are just a random subset of the data.



- **Missing at random (MAR):** the observation mechanism depends on some of the observed data but not directly on the hidden variables
  - Male participants of a study are more likely to tell their weight (weight is MAR).



- **Missing not at random (MNAR):** the observation mechanism depends on the hidden variables
  - People with low IQ values have missing observations for this variable (IQ is MNAR).

Expectation maximization (EM)

Expectation maximization (EM)

- EM is not a general-purpose algorithm for nonlinear function optimization.
- Tailored specifically to optimizing likelihood functions.

- Intuition
  - Parameter estimation from complete data is easy.
  - Guessing unobserved variables given parameters is possible (inference)
  - EM solves this “chicken and egg” problem in an iterative approach.

EM overview

- Pick an initial values for parameters

- Iterate
  - **E-step (expectation):** complete data using current parameters
  - **M-step (maximization):** estimate parameters from the current complete data

- Each iteration is guaranteed to improve the log-likelihood function

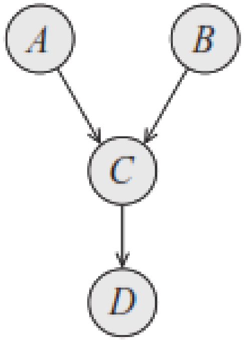
Expectation Maximization (EM)

University of TEHRAN

- In the fully observed case

m^{th} training example

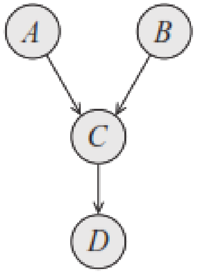
$$\hat{\theta}_{d^1|c^0} = \frac{M[d^1, c^0]}{M[c^0]} = \frac{\sum_{m=1}^{M} \mathbb{1}\{\xi[m]\langle D, C\rangle = \langle d^1, c^0\rangle\}}{\sum_{m=1}^{M} \mathbb{1}\{\xi[m]\langle C\rangle = c^0\}}$$



- Now consider $o = \langle a^1, ?\,, ?\,, d^0\rangle$ is given.
  - There are four possible realizations to B and C ($\langle b^1, c^1\rangle, \langle b^1, c^0\rangle, \langle b^0, c^1\rangle, \langle b^0, c^0\rangle$)
  - We do not know which of them is true. Question: how do we know which one is more likely?
    - We can compute distribution of hidden variables given observations (inference task), $Q(B, C) = P(B, C \mid o, \theta)$

First observation: $o = \langle a^1, ?, ?, d^0 \rangle$

$P(a^1, d^0 \mid \theta)$



$$
\begin{aligned}
\boldsymbol{\theta}_{a^1} &= 0.3 & \boldsymbol{\theta}_{b^1} &= 0.9 \\
\boldsymbol{\theta}_{d^1|c^0} &= 0.1 & \boldsymbol{\theta}_{d^1|c^1} &= 0.8 \\
\boldsymbol{\theta}_{c^1|a^0,b^0} &= 0.83 & \boldsymbol{\theta}_{c^1|a^1,b^0} &= 0.6 \\
\boldsymbol{\theta}_{c^1|a^0,b^1} &= 0.09 & \boldsymbol{\theta}_{c^1|a^1,b^1} &= 0.2,
\end{aligned}
$$

$$
\begin{aligned}
Q(\langle b^1, c^1 \rangle) &= 0.3 \cdot 0.9 \cdot 0.2 \cdot 0.2/0.2196 = 0.0492 \\
Q(\langle b^1, c^0 \rangle) &= 0.3 \cdot 0.9 \cdot 0.8 \cdot 0.9/0.2196 = 0.8852 \\
Q(\langle b^0, c^1 \rangle) &= 0.3 \cdot 0.1 \cdot 0.6 \cdot 0.2/0.2196 = 0.0164 \\
Q(\langle b^0, c^0 \rangle) &= 0.3 \cdot 0.1 \cdot 0.4 \cdot 0.9/0.2196 = 0.0492,
\end{aligned}
$$

Second observation: $o = \langle ?,\ b^1, ?, d^1 \rangle$

$$Q'(A, C) = P(A, C \mid b^1, d^1, \boldsymbol{\theta})$$

$$
\begin{aligned}
Q'(\langle a^1, c^1 \rangle) &= 0.3 \cdot 0.9 \cdot 0.2 \cdot 0.8/0.1675 = 0.2579 \\
Q'(\langle a^1, c^0 \rangle) &= 0.3 \cdot 0.9 \cdot 0.8 \cdot 0.1/0.1675 = 0.1290 \\
Q'(\langle a^0, c^1 \rangle) &= 0.7 \cdot 0.9 \cdot 0.09 \cdot 0.8/0.1675 = 0.2708 \\
Q'(\langle a^0, c^0 \rangle) &= 0.7 \cdot 0.9 \cdot 0.91 \cdot 0.1/0.1675 = 0.3423.
\end{aligned}
$$

$$\tilde{\boldsymbol{\theta}}_{d^1|c^0} = \frac{\bar{M}_{\boldsymbol{\theta}}[d^1, c^0]}{\bar{M}_{\boldsymbol{\theta}}[c^0]}$$

$$
\begin{aligned}
\bar{M}_{\boldsymbol{\theta}}[d^1, c^0] &= Q'(\langle a^1, c^0 \rangle) + Q'(\langle a^0, c^0 \rangle) \\
&= 0.1290 + 0.3423 = 0.4713
\end{aligned}
$$

$$\tilde{\boldsymbol{\theta}}_{d^1|c^0} = \frac{0.4713}{1.4057} = 0.3353.$$

$$
\begin{aligned}
\bar{M}_{\boldsymbol{\theta}}[c^0] &= Q(\langle b^1, c^0 \rangle) + Q(\langle b^0, c^0 \rangle) + Q'(\langle a^1, c^0 \rangle) + Q'(\langle a^0, c^0 \rangle) \\
&= 0.8852 + 0.0492 + 0.1290 + 0.3423 = 1.4057.
\end{aligned}
$$

Sufficient statistics

- A function $s(\xi)$ from instances to a vector in $\mathcal{R}^k$ if for any two datasets D and D' and any $\theta \in \Theta$ we have that

$$\sum_{x[i]\in D} s(x[i]) = \sum_{x[i]\in D'} s(x[i]) \quad \Rightarrow \quad L(\theta : D) = L(\theta : D')$$

- $\sum_{x[i]\in D} s(x[i])$ is called the sufficient statistics of data set D.

- Multinomial example:
  - $\langle M[1], \ldots, M[K]\rangle$ is the sufficient statistics from the data
  - It is computed by instance level statistics

k^{th} position

$$s(x^k) = (0, \ldots, 0, 1, 0, \ldots, 0)$$

$$L(\theta : D) = \prod_k \theta_k^{M[k]}$$

Fully *vs* partially observed case

- For each data case $\langle o[m], h[m] \rangle$
  - we define weight as $Q(h[m]) = P(h[m] \mid o[m], \theta)$

- *Expected sufficient statistics*

$$\bar{M}_{\boldsymbol{\theta}}[\boldsymbol{y}] = \sum_{m=1}^{M} \sum_{\boldsymbol{h}[m] \in Val(\boldsymbol{H}[m])} Q(\boldsymbol{h}[m]) \boldsymbol{I}\{\xi[m]\langle \boldsymbol{Y} \rangle = \boldsymbol{y}\}.$$

$$M[c^1] = \sum_{m=1}^{M} \boldsymbol{I}\{\xi[m]\langle C \rangle = c^1\}.$$

Fully observed case

$$\bar{M}_{\boldsymbol{\theta}}[c^1] = \sum_{m=1}^{M} P(c^1 \mid \boldsymbol{o}[m], \boldsymbol{\theta}).$$

Partially observed case

The EM algorithm for Bayesian networks

- Begin with some initial parameter assignment θ^0
- The algorithm then iterates for $t = 0, 1, \ldots$
  - Expectation (E-step): the algorithm uses the current parameters θ^t to compute the *expected sufficient statistics.*

Current set of parameters

$$\bar{M}_{\boldsymbol{\theta}^t}[x, \boldsymbol{u}] = \sum_m P(x, \boldsymbol{u} \mid \boldsymbol{o}[m], \boldsymbol{\theta}^t).$$

  - Maximization (M-step): treat the current *expected sufficient statistics* as observed, and perform *maximum likelihood estimation*

$$\boldsymbol{\theta}^{t+1}_{x|\boldsymbol{u}} = \frac{\bar{M}_{\boldsymbol{\theta}^t}[x, \boldsymbol{u}]}{\bar{M}_{\boldsymbol{\theta}^t}[\boldsymbol{u}]}$$

The EM algorithm for Bayesian networks

Procedure Expectation-Maximization (
$\mathcal{G}$, // Bayesian network structure over $X_1, \ldots, X_n$
$\boldsymbol{\theta}^0$, // Initial set of parameters for $\mathcal{G}$
$\mathcal{D}$ // Partially observed data set
)

for each $t = 0, 1 \ldots,$ until convergence
  // E-step
  $\{\bar{M}_t[x_i, \boldsymbol{u}_i]\} \leftarrow$ Compute-ESS($\mathcal{G}, \boldsymbol{\theta}^t, \mathcal{D}$)
  // M-step
  for each $i = 1, \ldots, n$
    for each $x_i, \boldsymbol{u}_i \in Val(X_i, \mathrm{Pa}^{\mathcal{G}}_{X_i})$
      $\theta^{t+1}_{x_i|\boldsymbol{u}_i} \leftarrow \frac{\bar{M}_t[x_i, \boldsymbol{u}_i]}{\bar{M}_t[\boldsymbol{u}_i]}$
return $\boldsymbol{\theta}^t$

Procedure Compute-ESS (
$\mathcal{G}$, // Bayesian network structure over $X_1, \ldots, X_n$
$\boldsymbol{\theta}$, // Set of parameters for $\mathcal{G}$
$\mathcal{D}$ // Partially observed data set
)

  // Initialize data structures
for each $i = 1, \ldots, n$
  for each $x_i, \boldsymbol{u}_i \in Val(X_i, \mathrm{Pa}^{\mathcal{G}}_{X_i})$
    $\bar{M}[x_i, \boldsymbol{u}_i] \leftarrow 0$
  // Collect probabilities from all instances
for each $m = 1 \ldots M$
  Run inference on $\langle \mathcal{G}, \boldsymbol{\theta} \rangle$ using evidence $\boldsymbol{o}[m]$
  for each $i = 1, \ldots, n$
    for each $x_i, \boldsymbol{u}_i \in Val(X_i, \mathrm{Pa}^{\mathcal{G}}_{X_i})$
      $\bar{M}[x_i, \boldsymbol{u}_i] \leftarrow \bar{M}[x_i, \boldsymbol{u}_i] + P(x_i, \boldsymbol{u}_i \mid \boldsymbol{o}[m])$
return $\{\bar{M}[x_i, \boldsymbol{u}_i] : \forall i = 1, \ldots, n, \forall x_i, \boldsymbol{u}_i \in Val(X_i, \mathrm{Pa}^{\mathcal{G}}_{X_i})\}$

EM – General case

- X: observed data, Z: hidden variables, θ: parameters
- Complete data: {X, Y}

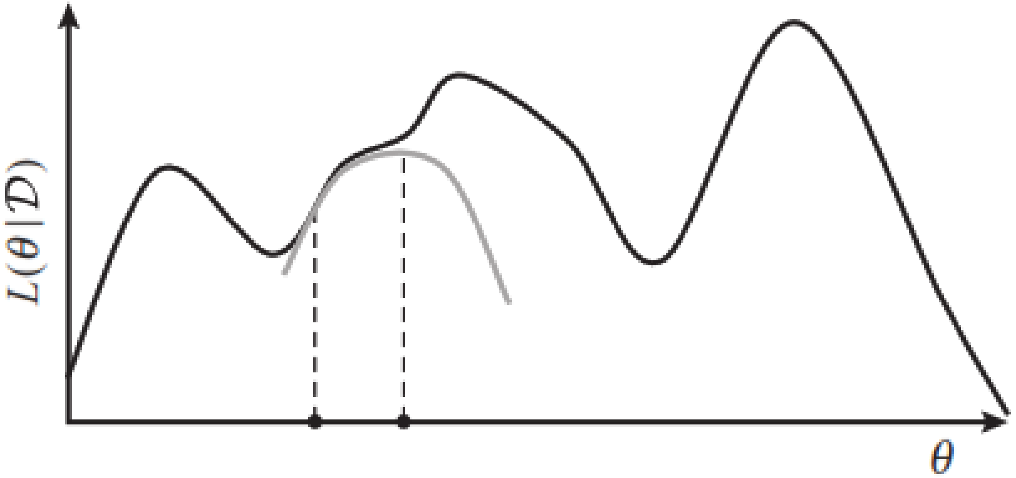
- **E-step:** we compute the expectation of the complete-data likelihood

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} \underbrace{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})}_{\text{Posterior of hidden variables given observations}} \ln \underbrace{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}_{\text{Complete-data likelihood}}.$$

- **M-step:** we maximize this function

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}).$$

Mathematical intuition



Theorem 19.5 *During iterations of the EM procedure of algorithm 19.2, we have that*

$$\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}) - \ell(\boldsymbol{\theta}^{t} : \mathcal{D}) \geq \boldsymbol{E}_{P(\mathcal{H}|\mathcal{D},\boldsymbol{\theta}^t)}\left[\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}, \mathcal{H})\right] - \boldsymbol{E}_{P(\mathcal{H}|\mathcal{D},\boldsymbol{\theta}^t)}\left[\ell(\boldsymbol{\theta}^{t} : \mathcal{D}, \mathcal{H})\right].$$

As a consequence, we obtain that:

$$\ell(\boldsymbol{\theta}^{t} : \mathcal{D}) \leq \ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}).$$

Each iteration improves the likelihood

Hard-assignment EM

- It iterates over two steps
  - **Completing the data:** for each data instance $o[m]$, select the single assignment $h[m]$ that maximizes $P(h \mid o[m], \theta^t)$.
  - Estimating the new parameters θ^{t+1} using the complete data
- The objective is to maximize the likelihood of the complete data

$$\max_{\boldsymbol{\theta}, \mathcal{H}} \ell(\boldsymbol{\theta} : \mathcal{H}, \mathcal{D}).$$

- **The "Soft-assignment" EM objective:** attempts to maximize $l(\theta : D)$, averaging over all possible completions of the data

Structure learning

- Both the scoring function and the search procedure is considerably more complicated in the case of incomplete data.
- Scoring structures
  - The likelihood score does not penalize more complex models
  - Recall Bayesian score

$$\text{score}_{\mathcal{B}}(\mathcal{G} : \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

where

$$P(\mathcal{D} \mid \mathcal{G}) = \int_{\Theta_{\mathcal{G}}} P(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G}) P(\boldsymbol{\theta}_{\mathcal{G}} \mid \mathcal{G}) d\boldsymbol{\theta}_{\mathcal{G}}.$$

Likelihood function does not decompose for incomplete data

  - Different approximations for dealing this issue such as Laplace approximation

$$\text{score}_{Laplace}(\mathcal{G} : \mathcal{D}) = \log P(\mathcal{G}) + \log P(\mathcal{D} \mid \tilde{\boldsymbol{\theta}}_{\mathcal{G}}, \mathcal{G}) + \frac{\dim(\boldsymbol{C})}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{C}|,$$

In practice expensive to compute

Structural EM

- E-step:
  - we use our current model to generate (perhaps implicitly) a completed data set
  - We may only need to compute expected sufficient statistics

- M-step:
  - Structure learning (e.g. hill climbing)
  - Parameter estimation

Algorithm 19.3 The structural EM algorithm for structure learning

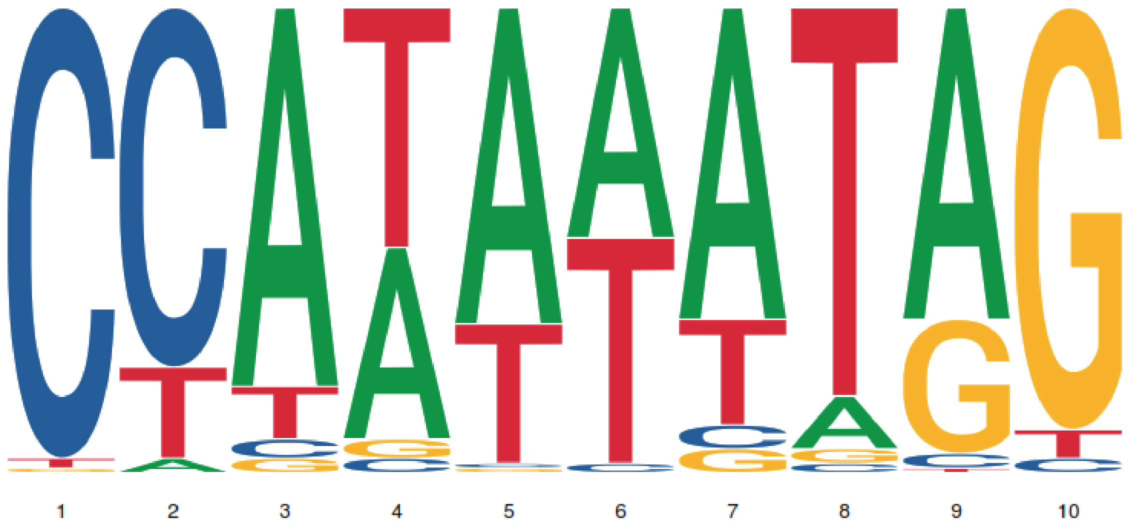
Procedure Structural-EM (
  $\mathcal{G}^0$, // Initial bayesian network structure over $X_1, \ldots, X_n$
  $\boldsymbol{\theta}^0$, // Initial set of parameters for $\mathcal{G}^0$
  $\mathcal{D}$ // Partially observed data set
)

for each $t = 0, 1 \ldots,$ until convergence
    // Optional parameter learning step
  $\boldsymbol{\theta}^{t'} \leftarrow$ Expectation-Maximization$(\mathcal{G}^t, \boldsymbol{\theta}^t, \mathcal{D})$
    // Run EM to generate expected sufficient statistics for $\mathcal{D}^*_{\mathcal{G}^t, \boldsymbol{\theta}^{t'}}$
  $\mathcal{G}^{t+1} \leftarrow$ Structure-Learn$(\mathcal{D}^*_{\mathcal{G}^t, \boldsymbol{\theta}^{t'}})$
  $\boldsymbol{\theta}^{t+1} \leftarrow$ Estimate-Parameters$(\mathcal{D}^*_{\mathcal{G}^t, \boldsymbol{\theta}^{t'}}, \mathcal{G}^{t+1})$
return $\mathcal{G}^t, \boldsymbol{\theta}^t$

- If we use the BIC score, the structural EM will improve the BIC score at each iteration.
- Unlike the case EM, no proof the structure it finds is a local maximum.

	Known structure (Parameter estimation)	**Unknown structure (Structure learning)**
Fully observable	*MLE* *Bayesian methods*	*Constraint-based methods* *Score-based methods e.g. hill climbing*
Partially observable	Gradient ascent Expectation maximization	Structural EM

Motif finding

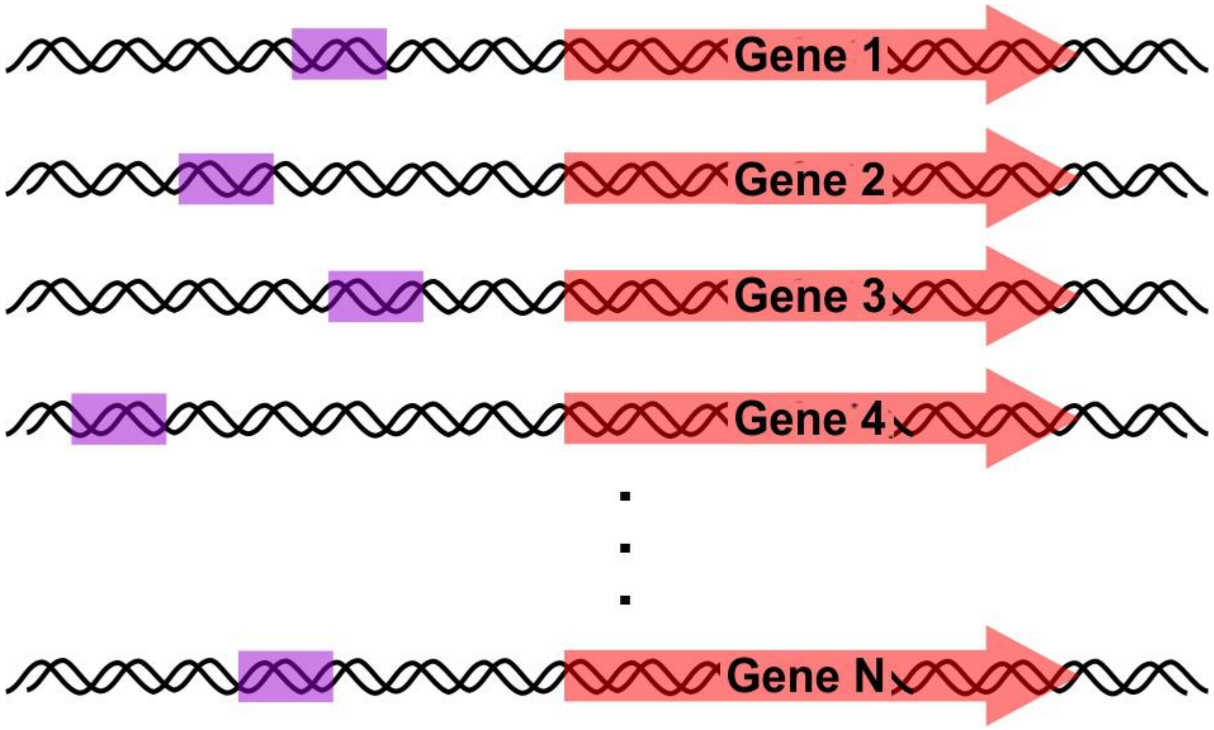


Regulatory motif

- A transcription factor regulates a gene by binding to a specific short DNA interval called a regulatory motif, or transcription factor binding site, in the gene's upstream region.

- For example, the transcription factor CCA1 binds to AAAATATCT in upstream of many genes

Finding regulatory motifs



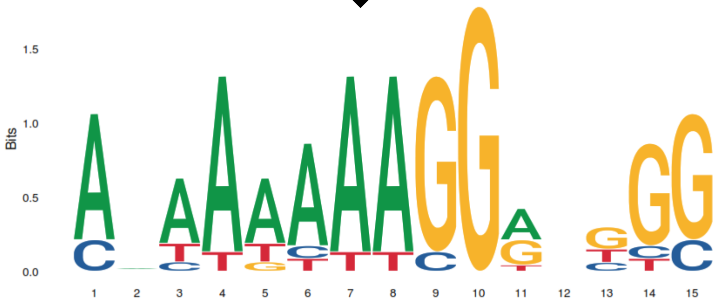
Given a collection of genes with common expression, find a common transcription factor binding motif!

University of TEHRAN

1 atgaccgggatactgataaaaaaaagggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
2 acccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataaaaaaaaaggggggga
3 tgagtatccctgggatgacttaaaaaaaagggggggtgctctcccgattttgaatatgtaggatcattcgccagggtccga
4 gctgagaattggatgaaaaaaaagggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
5 tcccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaagggggggcttatag
6 gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
7 cggttttggcccttgttagaggcccccgtaaaaaaaagggggggcaattatgagagagctaatctatcgcgtgcgtgttcat
8 aacttgagttaaaaaaaagggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
9 ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaagggggggaccgaaaggaag
10 ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga

Motif finding

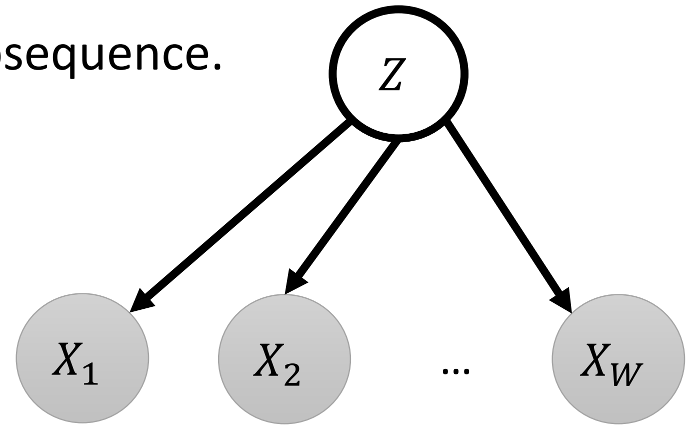
1 atgaccgggatactgat**AgAAgAAAGGttGGG**ggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
2 acccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaata**cAAtAAAAcGGcGGG**a
3 tgagtatccctgggatgactt**AAAAtAAtGGaGtGG**tgctctcccgattttgaatatgtaggatcattcgccagggtccga
4 gctgagaattggatg**cAAAAAAAGGGattG**tccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
5 tcccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaat**AtAAtAAAGGaaGGG**cttatag
6 gtcaatcatgttcttgtgaatggattt**AAcAAtAAGGGctGG**gaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
7 cggttttggcccttgttagaggcccccgt**AtAAAcAAGGaGGGc**caattatgagagagctaatctatcgcgtgcgtgttcat
8 aacttgagtt**AAAAAAtAGGGaGcc**ctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
9 ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcat**ActAAAAAGGaGcGG**accgaaaggaag
10 ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagctt**ActAAAAAGGaGcGG**a



Expectation Maximization for motif finding

- Given:
  - A collection of sequences denoted as over the alphabet $\mathcal{A} = \{A, C, G, T\}$.
  - *W:* length of motif

- idea:
  - breaks up the sequences into *N* (overlapping) subsequences of length *W*. The data is $D = \{x[1], \ \ldots, x[N]\}$ where $x[i] = (x_1[i], \ldots, x_W[i])$.
  - Consider a two-component mixture model that assumes each subsequence is either an instance of the motif or background.
  - The hidden variable *Z* indicates which component generated the subsequence.
  - Perform EM for parameter estimation of the mixture model.



TTACCTTAAC
GATGTCTGTC
ACGGCGTTAG
CCCTAACGAG

TTAC	GATG	ACGG	CCCT
TACC	ATGT	CGGC	CCTA
ACCT	TGTC	GGCG	CTAA
CCTT	GTCT	GCGT	TAAC
CTTA	TCTG	CGTT	AACG
TTAA	CTGT	GTTA	ACGA
TAAC	TGTC	TTAG	CGAG

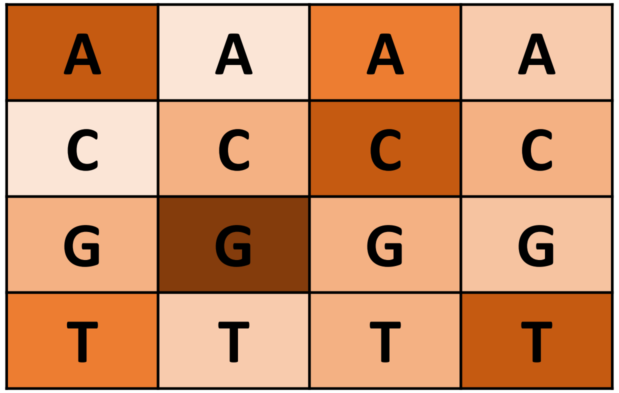
each subsequence is either an instance of the motif or background.

TTAC	GATG	ACGG	CCCT
TACC	ATGT	CGGC	CCTA
ACCT	TGTC	GGCG	CTAA
CCTT	GTCT	GCGT	TAAC
CTTA	TCTG	CGTT	AACG
TTAA	CTGT	GTTA	ACGA
TAAC	TGTC	TTAG	CGAG

$$\Pr(TTAC \mid motif) = f_{1T} f_{2T} f_{3A} f_{4C}$$

$$\Pr(TTAC \mid background) = f_{0T} f_{0T} f_{0A} f_{0C}$$

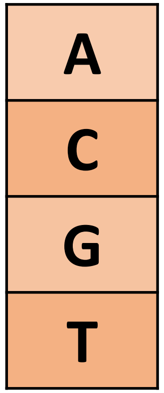
f_1 f_2 f_3 f_4



Motif model

$\theta_m = (f_1, \ldots, f_W)$

f_0



Background model

$\theta_{bg} = f_0$

Expectation Maximization for motif finding-2

- Conditional probability distributions:

$$P(\, z[i] = 1 \mid \lambda_m \,) = \lambda_m$$

$$P(\, x[i] \mid z[i] = 0 \,) = \prod_j f_{0, x_j[i]}$$

$$P(\, x[i] \mid z[i] = 1 \,) = \prod_j f_{1, x_j[i]}$$

- **Questions:**
    - Write the joint distribution of the following model.
    - Maximum likelihood estimation for λ_m, f_0, f_j.
    - Write down $P(\, z[i] = 1 \mid x[i] \,)$.

- You will examine the EM algorithm for this problem in detail in project 3.