ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

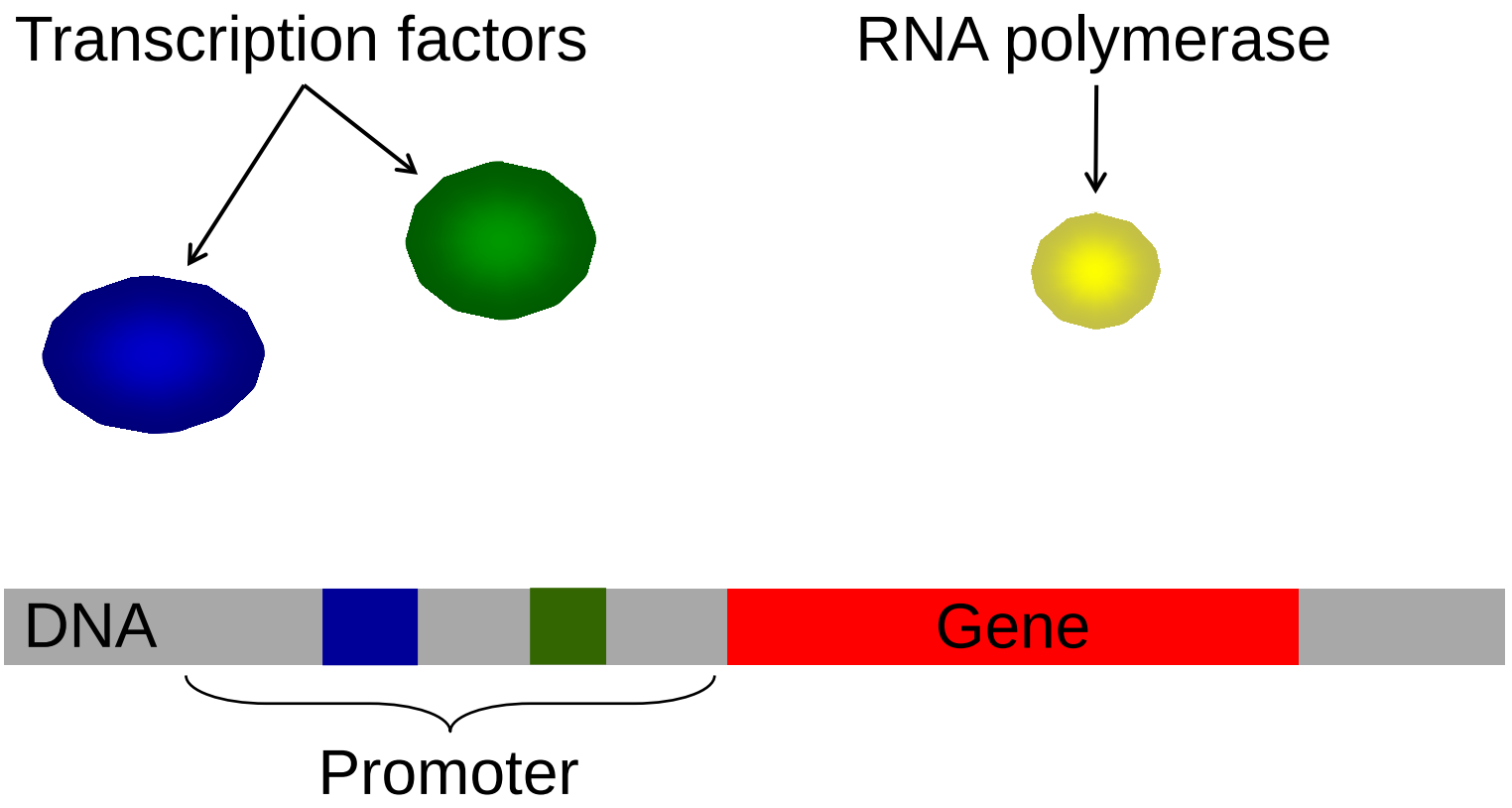
Expectation Maximization algorithm and motif finding

Niko Beerenwinkel

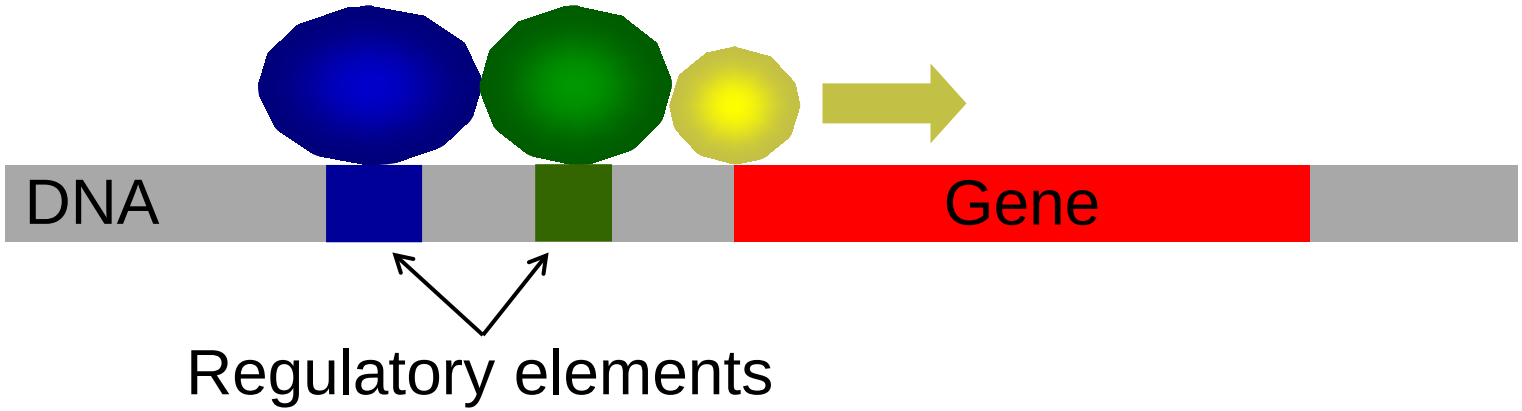
Outline

- Gene regulation
- Motif finding
- Finite mixture model
- EM algorithm
- EM algorithm in general

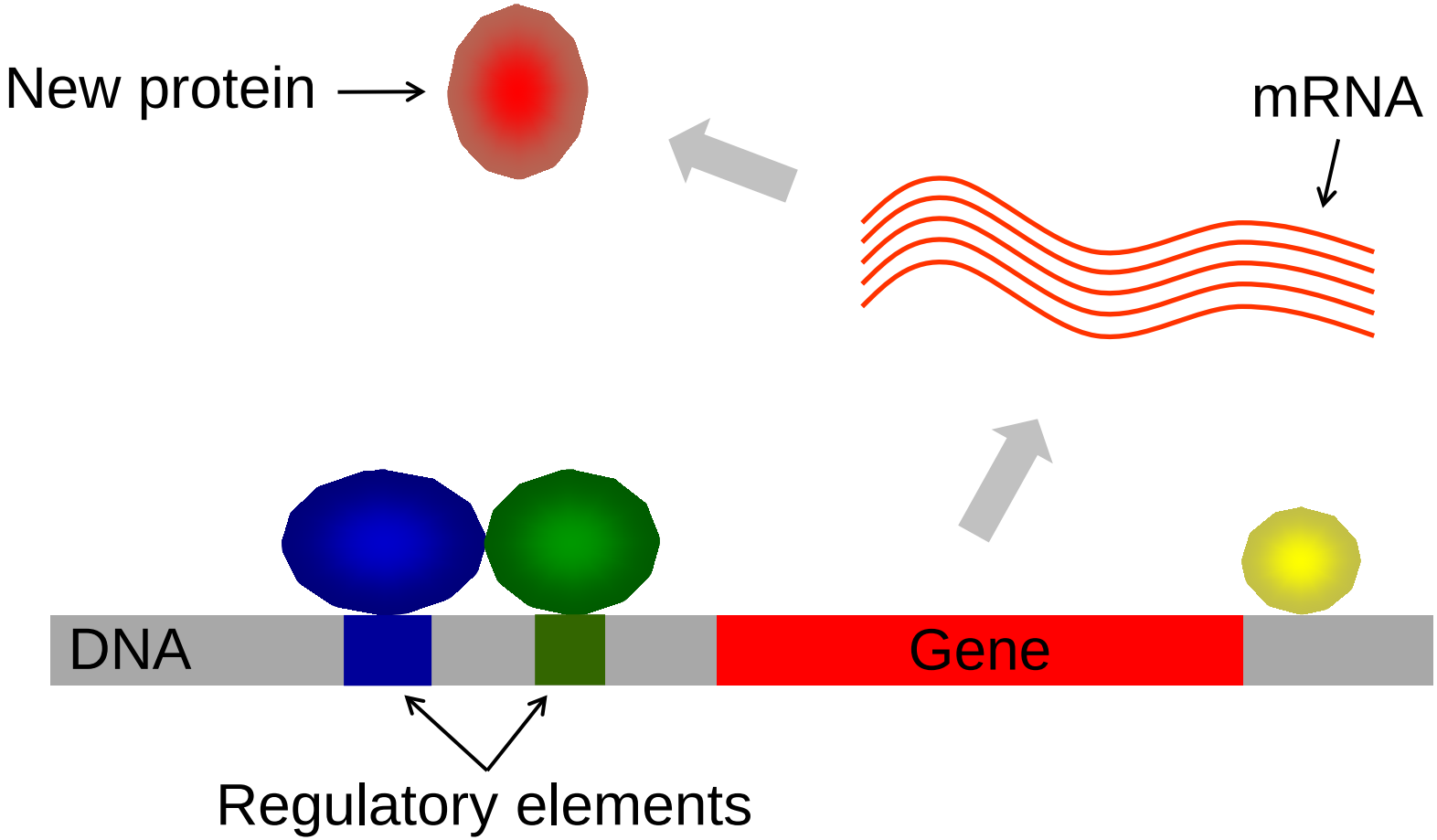
Gene regulation



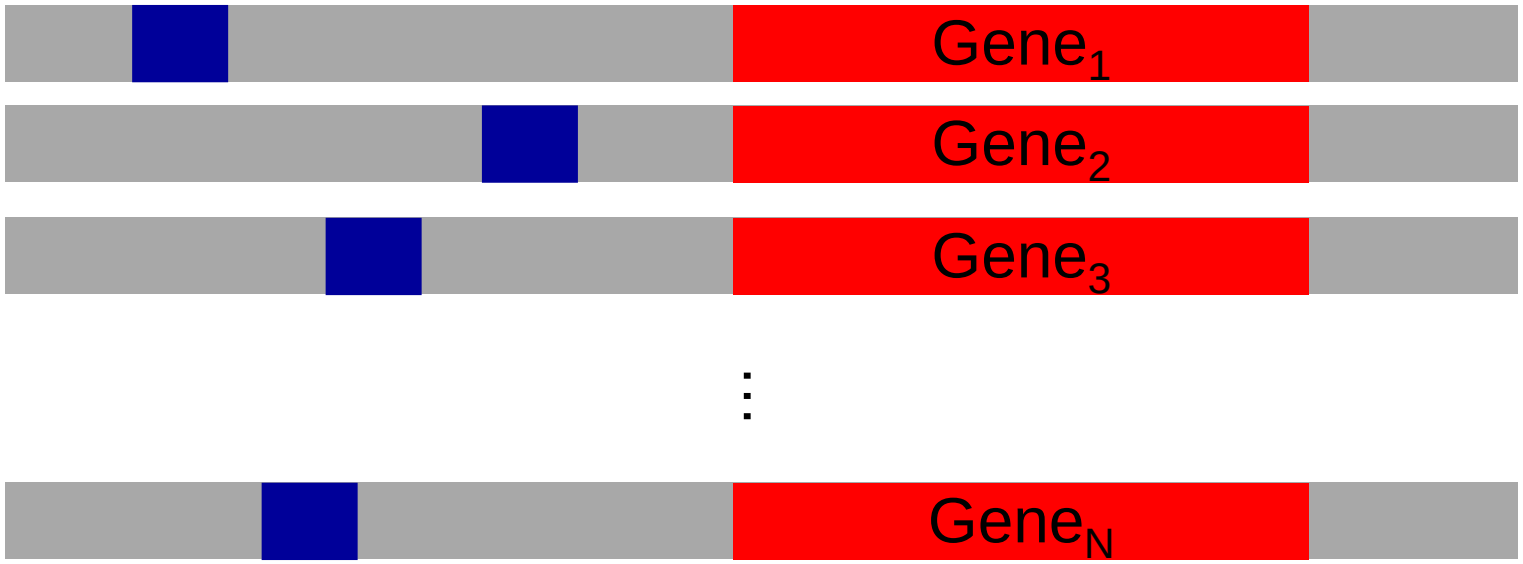
Transcription initiation



Translation



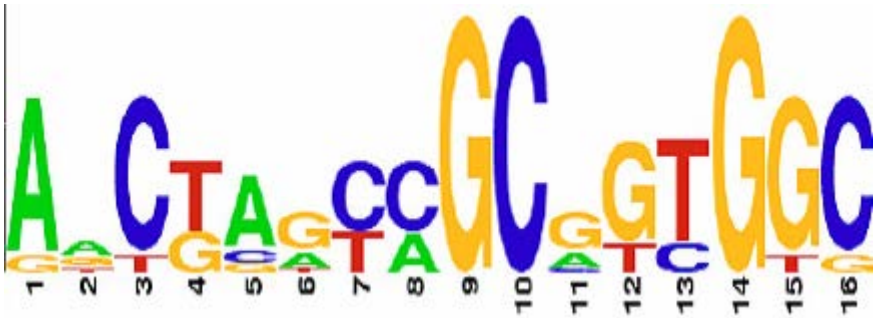
Finding regulatory DNA motifs

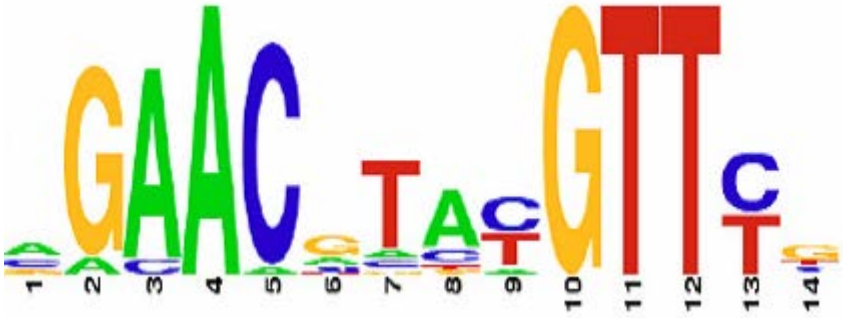


- Given a collection of genes with common expression, find a common transcription factor binding site motif

Regulatory motifs

- Short
- Constant size
- Highly variable
- Often repeated
- Gapless multiple alignment summarized in a sequence profile (logo)



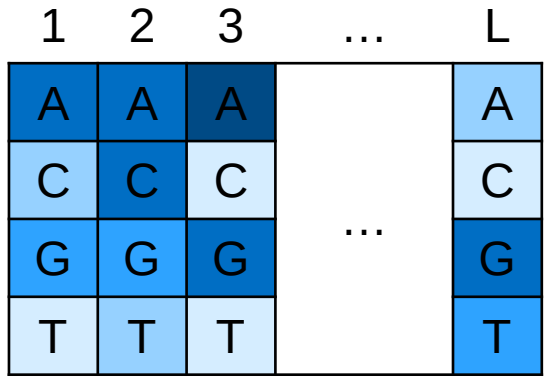


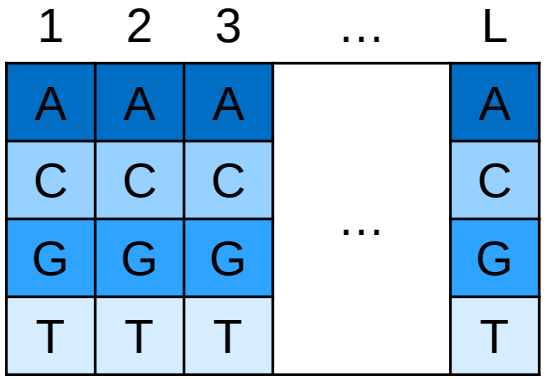
Motif-background mixing

motif

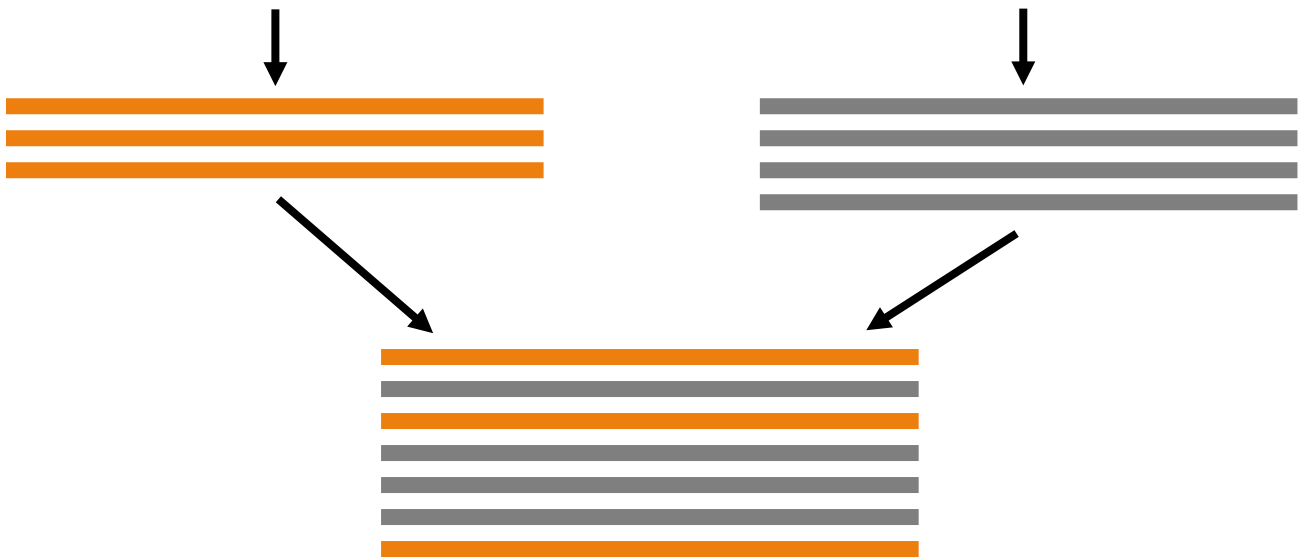
back-
ground

$\theta_1 =$





$= \theta_2$



Probabilistic motif finding using a finite mixture model

- Let $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$ be a set of sequences, each of length L, e.g., all subsequences of some promoter sequences.
- We regard $\mathcal{D}$ as a realization of $X = (X^{(1)}, \ldots, X^{(N)})$, where each $X^{(i)} = (X_1^{(i)}, \ldots, X_L^{(i)})$ and $X_n^{(i)} \in \mathcal{A} = \{A,C,G,T\}$.
- Each word $X^{(i)}$ is generated either
  - by the **motif model** with parameters $\theta_1 = (f_1, \ldots, f_L)$, where
    f_{nj} = Prob. of character $j \in \mathcal{A}$ at position n = 1, ..., L
    (positions are independent), or
  - by the **background model** with parameters $\theta_2 = f_0$, where
    f_{0j} = Prob. of character $j \in \mathcal{A}$ at any position,
    (positions are independent and identical)
- The motif model is used with probability λ_1, the background model with probability $\lambda_2 = 1 - \lambda_1$.

Hidden variables

- We cannot observe whether a word $X^{(i)}$ is generated by the motif (1) or by the background (2) model component.
- Define $Z = (Z^{(1)}, \ldots, Z^{(N)})$, $Z^{(i)} = (Z^{(i1)}, Z^{(i2)})$, where

$$Z^{(ik)} = \begin{cases} 1 & \text{if } X^{(i)} \text{ is from component } k \\ 0 & \text{otherwise} \end{cases}$$

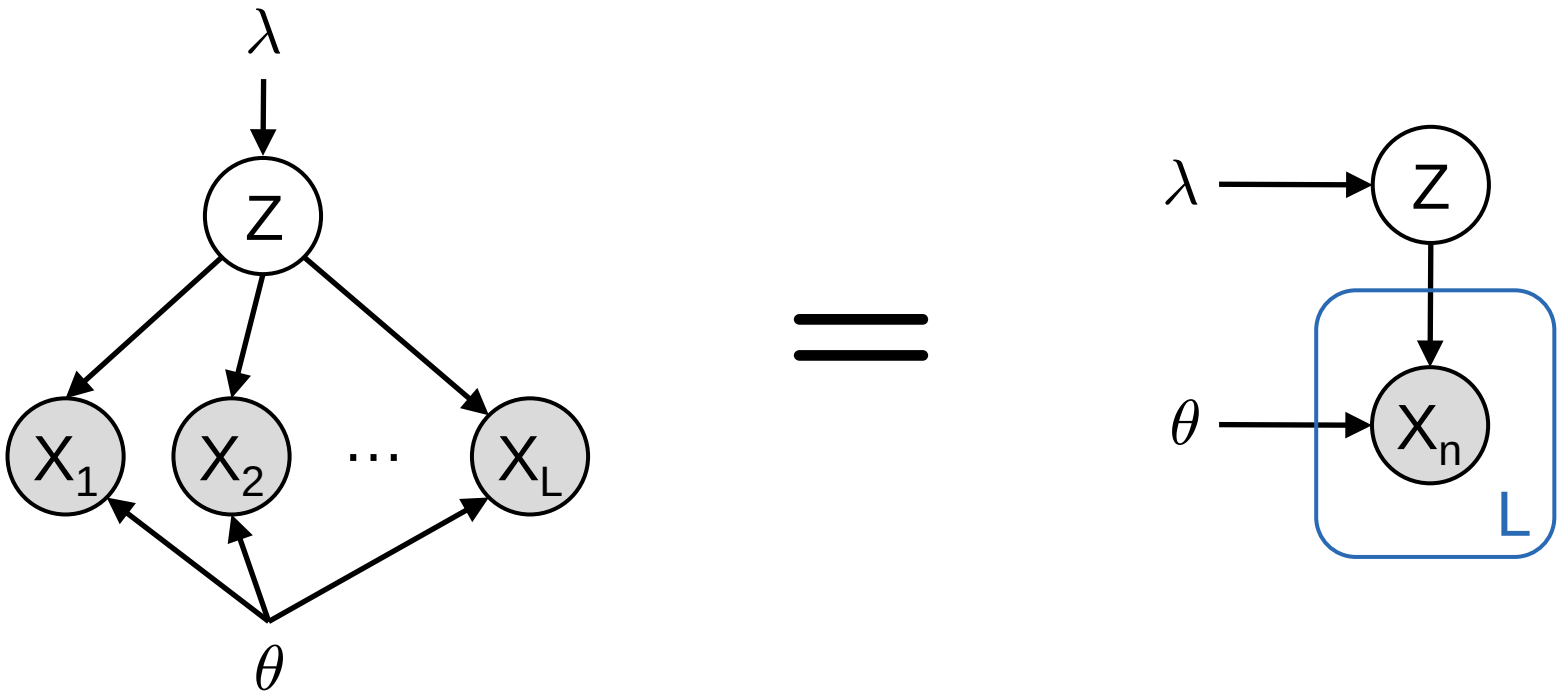
and $P(Z^{(ik)} = 1) = \lambda_k$. Note that $Z^{(i1)} + Z^{(i2)} = 1$. Then

$$P(X^{(i)}, Z^{(ik)} = 1 \mid \theta, \lambda) = \lambda_k P(X^{(i)} \mid \theta_k)$$

$$= \begin{cases} \lambda_1 f_{1X_1^{(i)}} \cdots f_{LX_L^{(i)}} & \text{if } k = 1 \\ \lambda_2 f_{0X_1^{(i)}} \cdots f_{0X_L^{(i)}} & \text{if } k = 2 \end{cases}$$

Graphical model representation

- $X = (X_1, \ldots, X_L)$ is an observed random variable.
- Z is a hidden (or latent) random variable representing missing data.



Joint probability

- The joint probability of a single word and its membership is

$$P(X^{(i)}, Z^{(i)} \mid \theta, \lambda) = P(Z^{(i)} \mid \lambda) P(X^{(i)} \mid Z^{(i)}, \theta)$$

$$= \prod_{k=1,2} \left[\lambda_k P(X^{(i)} \mid \theta_k) \right]^{Z^{(ik)}}$$

- The joint probability of the observed joint r.v. X and the hidden joint r.v. Z is

$$P(X, Z \mid \theta, \lambda) = \prod_{i=1}^{N} P(X^{(i)}, Z^{(i)} \mid \theta, \lambda)$$

$$= \prod_{i=1}^{N} \prod_{k=1,2} \left[\lambda_k P(X^{(i)} \mid \theta_k) \right]^{Z^{(ik)}}$$

Complete-data log-likelihood

- The log-likelihood of the hidden data is

$$\ell_{\text{hid}}(\theta, \lambda) := \log P(X, Z \mid \theta, \lambda)$$

$$= \sum_{i=1}^{N} \log P(X^{(i)}, Z^{(i)} \mid \theta, \lambda)$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{2} Z^{(ik)} \log \left[\lambda_k P(X^{(i)} \mid \theta_k) \right]$$

Observed likelihood

- With $\mathcal{Z}$ = {(0,1), (1,0)}, the likelihood of the observed data is

$$L_{\text{obs}}(\theta, \lambda) := P(X \mid \theta, \lambda) = \sum_{Z} P(X, Z \mid \theta, \lambda)$$

$$= \sum_{Z^{(1)} \in \mathcal{Z}} \cdots \sum_{Z^{(N)} \in \mathcal{Z}} \prod_{i} P(X^{(i)}, Z^{(i)} \mid \theta, \lambda)$$

- Usually, the observed log-likelihood

$$\ell_{\text{obs}}(\theta, \lambda) = \log L_{\text{obs}}(\theta, \lambda)$$

is much harder to maximize than the hidden log-likelihood.

Expected complete-data log-likelihood

- For any distribution q(Z) of the hidden data Z,

$$\begin{aligned}
\ell_{\text{obs}}(\theta, \lambda) &= \log \sum_Z P(X, Z \mid \theta, \lambda) \\
&= \log \sum_Z q(Z) \frac{P(X, Z \mid \theta, \lambda)}{q(Z)} \\
&= \log \mathrm{E}_q \left[P(X, Z \mid \theta, \lambda) / q(Z) \right] \\
&\geq \mathrm{E}_q \left[\log \left\{ P(X, Z \mid \theta, \lambda) / q(Z) \right\} \right] \\
&= \mathrm{E}_q \left[\ell_{\text{hid}}(\theta, \lambda) \right] - \mathrm{E}_q \left[\log q(Z) \right]
\end{aligned}$$

Jensen's inequality →

Expectation Maximization (EM) algoritm

- Basic idea: We iteratively maximize the lower bound

$$\ell_{\text{obs}}(\theta, \lambda) \geq \mathrm{E}_q \left[\ell_{\text{hid}}(\theta, \lambda)\right] - \mathrm{E}_q \left[\log q(Z)\right]$$

- **E step**: maximize w.r.t. q, i.e., set q = P(Z | X, θ, λ)

$$Z^{\text{new}} = \mathrm{E}\left[Z \mid X, \theta^{\text{old}}, \lambda^{\text{old}}\right]$$

- **M step**: maximize w.r.t. (θ, λ)

$$(\theta^{\text{new}}, \lambda^{\text{new}}) = \underset{\theta,\lambda}{\operatorname{argmax}}\, \mathrm{E}_{Z^{\text{old}}|X} \left[\ell_{\text{hid}}(\theta, \lambda)\right]$$

Motif mixture model: E step

- The expected value of the missing data is

$$\gamma_{ik} = \mathrm{E}\left[Z^{(ik)} \mid X, \theta, \lambda\right]$$

$$= \frac{\lambda_k P(X^{(i)} \mid \theta_k)}{\lambda_1 P(X^{(i)} \mid \theta_1) + \lambda_2 P(X^{(i)} \mid \theta_2)},$$

the *responsibility* of component k for observation i.

Motif mixture model: expected hidden log-likelihood

$$\begin{aligned}
\mathsf{E}\left[\ell_{\mathsf{hid}}(\theta,\lambda)\right] &= \mathsf{E}\left\{\sum_i \sum_k Z^{(ik)} \log\left[\lambda_k P(X^{(i)} \mid \theta_k)\right]\right\} \\
&= \sum_{i,k} \mathsf{E}\left[Z^{(ik)} \mid X^{(i)}, \theta, \lambda\right] \log\left[\lambda_k P(X^{(i)} \mid \theta_k)\right] \\
&= \sum_{i,k} \gamma_{ik} \log\left[\lambda_k P(X^{(i)} \mid \theta_k)\right] \\
&= \sum_{i,k} \gamma_{ik} \log \lambda_k + \sum_{i,k} \gamma_{ik} \log P(X^{(i)} \mid \theta_k)
\end{aligned}$$

Motif mixture model: M step, λ

- Maximization w.r.t. λ (left sum) yields

$$\hat{\lambda}_k = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ik}$$

Motif mixture model: M step, θ

- For maximizing w.r.t. θ (right sum), let
  - c_{nj} be the expected count of letter j in motif position n,

$$c_{nj} = \sum_{i=1}^{N} \gamma_{i1} \mathbb{I}\{X_n^{(i)} = j\}$$

  - and c_{0j} the expected count of letter j in any background position,

$$c_{0j} = \sum_{i=1}^{N} \sum_{n=1}^{L} \gamma_{i2} \mathbb{I}\{X_n^{(i)} = j\}$$

- Then the argument $\theta = (\theta_1, \theta_2) = (f_1, \ldots, f_L, f_0)$ maximizing the expected hidden log-likelihood is given by

$$\hat{f}_{nj} = \frac{c_{nj}}{\sum_{j'} c_{nj'}}$$

Summary: EM algorithm

1. Initialize parameters (θ, λ)
2. Repeat
   - **E step**: Compute expectation of missing data
     $Z^{\text{new}} = \text{E}[Z \mid X, \theta^{\text{old}}, \lambda^{\text{old}}]$
   - **M step**: Maximize expected hidden log-likelihood
     $(\theta^{\text{new}}, \lambda^{\text{new}}) = \text{argmax}_{(\theta,\lambda)} \text{E}_{Z^{\text{old}}|X} [\ell_{\text{hid}}(\theta, \lambda)]$
3. Until change in parameters or in likelihood is small

- *O*(NL) per iteration
- The EM algorithm is only guaranteed to find an (approximate) *local* maximum of the likelihood function.
- Generally, different starting solutions give different results.

The EM algorithm in general

- Consider any probabilistic (graphical) model with observed data X, hidden data Z, and parameters θ.
- We want to maximize the likelihood

$$L_{\text{obs}}(\theta) = P(X \mid \theta) = \sum_{Z} P(X, Z \mid \theta)$$

- We will decompose the log-likelihood

$$\ell_{\text{obs}}(\theta) = \log P(X \mid \theta)$$

into two terms, one of which is the lower bound on the log-likelihood derived before.

Kullback-Leibler divergence

- The KL divergence between two probability distributions P(X) and Q(X) is defined as

$$\mathrm{D_{KL}}(P \parallel Q) = -\sum_X P(X) \log \frac{Q(X)}{P(X)}$$

- Properties:

$$\mathrm{D_{KL}}(P \parallel Q) \geq 0$$

$$\mathrm{D_{KL}}(P \parallel Q) = 0 \iff P = Q$$

- KL divergence measures the dissimilarity between P and Q.

Lower bound on the log-likelihood revisited

- For any distribution q, let us define

$$F(q,\theta) := \sum_Z q(Z) \log \frac{P(X,Z \mid \theta)}{q(Z)}$$

- F is the lower bound on the log-likelihood derived earlier:

$$\begin{aligned} F(q,\theta) &= \sum_Z q(Z) \left[\log P(X,Z \mid \theta) - \log q(Z)\right] \\ &= \mathrm{E}_q[\ell_{\mathrm{hid}}(\theta)] - \mathrm{E}_q[\log q(Z)] \end{aligned}$$

Decomposing the log-likelihood

$$
\begin{aligned}
F(q,\theta) &= \sum_Z q(Z) \log \frac{P(X,Z \mid \theta)}{q(Z)} = \\
&= \sum_Z q(Z) \log \left[P(X \mid \theta) \frac{P(Z \mid X,\theta)}{q(Z)} \right] \\
&= \sum_Z q(Z) \left[\log P(X \mid \theta) + \log \frac{P(Z \mid X,\theta)}{q(Z)} \right] \\
&= \log P(X \mid \theta) - \mathrm{D}_{\mathrm{KL}} \left[q(Z) \parallel P(Z \mid X,\theta) \right] \\
&= \ell_{\mathrm{obs}}(\theta) - \mathrm{D}_{\mathrm{KL}}(q \parallel P)
\end{aligned}
$$

Log-likelihood decomposition

- For any q, we have

$$\ell_{\text{obs}}(\theta) = F(q, \theta) + \mathrm{D}_{\mathrm{KL}}(q \parallel P)$$

- $\ell_{\text{obs}}(\theta) \geq F(\mathrm{q}, \theta)$, because $\mathrm{D}_{\mathrm{KL}} \geq 0$.
- F is called the evidence lower bound (ELBO).

E step

- In the E step, F is maximized w.r.t. q:

$$\max_q F(q, \theta^{\text{old}}) = \max_q \left[\ell_{\text{obs}}(\theta^{\text{old}}) - \mathrm{D}_{\mathrm{KL}}(q \parallel P) \right]$$

- Because ℓ_{obs} is independent of q, this optimization problem is solved by q(Z) = P(Z | X, θ^{old}).

- After the E step,

$$F(q, \theta^{\text{old}}) = \ell_{\text{obs}}(\theta^{\text{old}})$$

M step

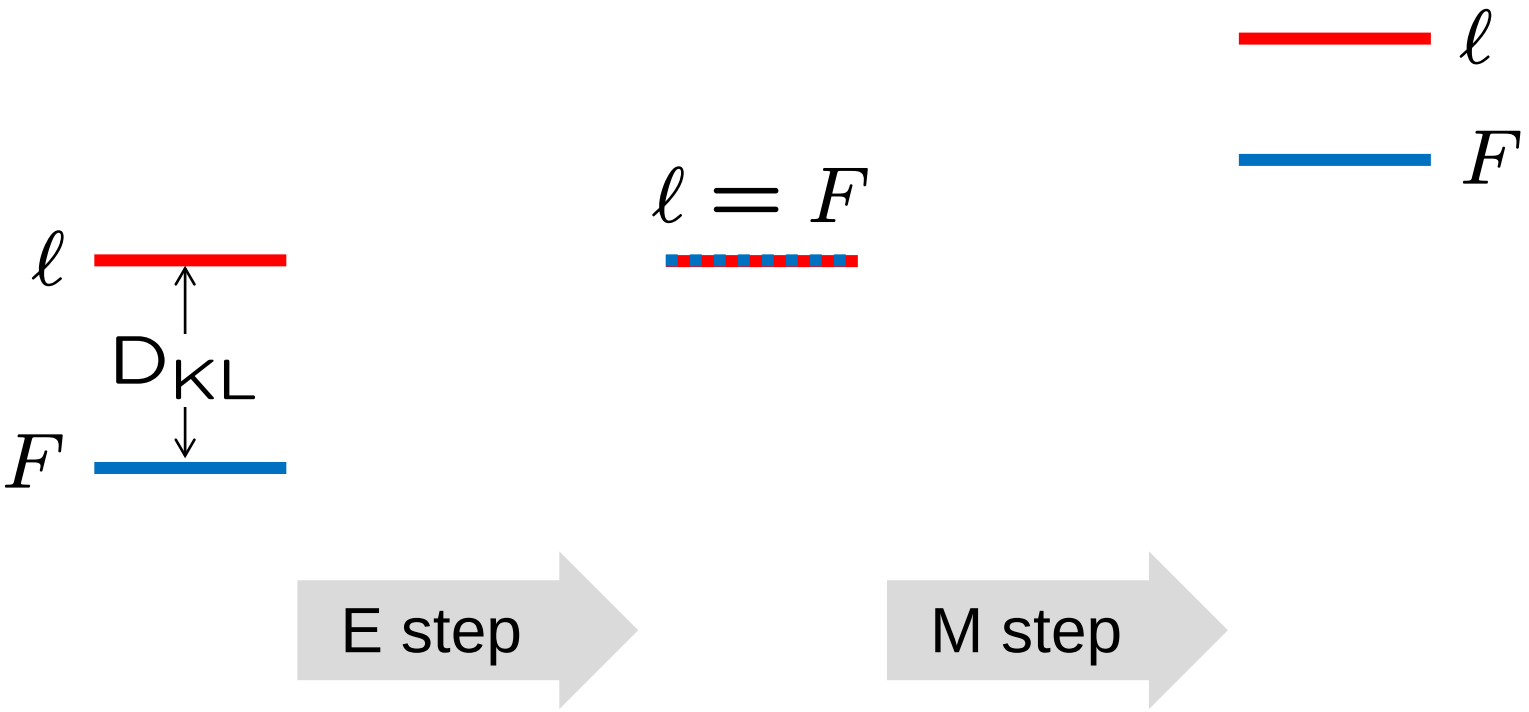
- In the M step, $q^{\text{old}}(Z) = P(Z \mid X, \theta^{\text{old}})$ is fixed, and F is maximized w.r.t. θ:

$$\max_{\theta} F(q^{\text{old}}, \theta) = \max_{\theta} \left\{ \mathrm{E}_{q^{\text{old}}}[\ell_{\text{hid}}(\theta)] - \underbrace{\mathrm{E}_{q^{\text{old}}}[\log q^{\text{old}}(Z)]}_{\text{entropy}} \right\}$$

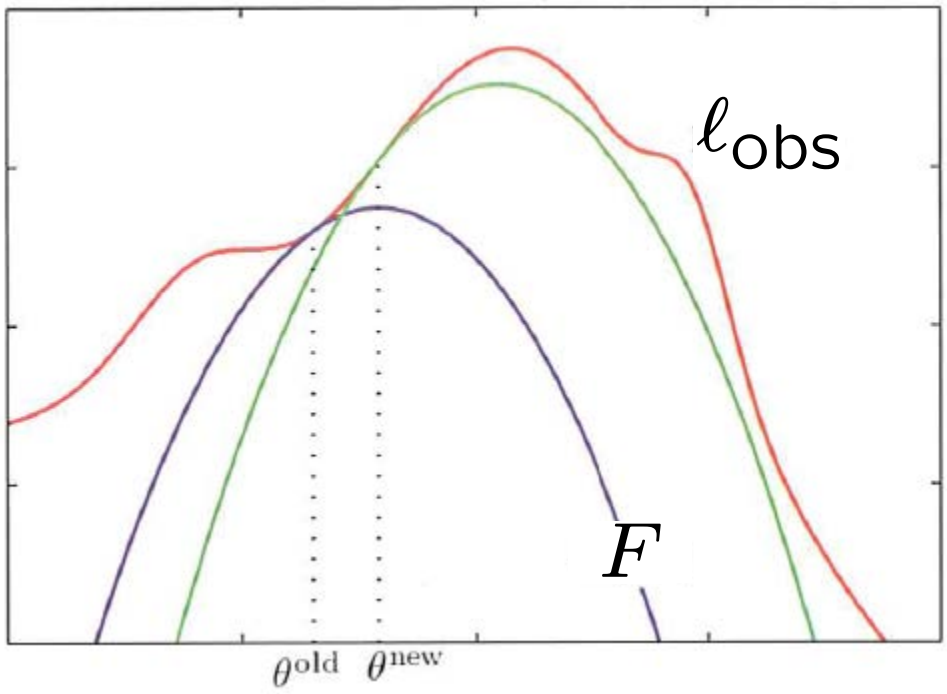
- The entropy term on the right does not depend on θ.
- Maximization will increase F and hence ℓ_{obs}.
- But, in general, the increase in ℓ_{obs} will be greater, because

$$\mathrm{D}_{\mathrm{KL}}(P(Z \mid X, \theta^{\text{old}}) \parallel P(Z \mid X, \theta^{\text{new}}) > 0$$

Summary of the general EM algorithm



EM algorithm in parameter space



- F is tangential to ℓ.
- For a large class of models (mixtures of exponential family components), F is convex and much easier to optimize.

Summary

- Regulatory DNA sequences contain specific, short, conserved sequence segments called motifs.
- Motifs can be detected as specific model parameters of a mixture model generating motif and background words.
- Parameter estimation in the presence of unobserved (missing or hidden) data can be accomplished by the Expectation Maximization (EM) algorithm.
- The EM algorithm iteratively computes the expectation of the missing data (E step) and maximizes the expected hidden log-likelihood of the data (M step).

References

- Bishop CM. Pattern Recognition and Machine Learning. Section 9.4.
- Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. ISMB 1994, pp. 28-36. PDF
  - Software and web service: https://meme-suite.org/meme/
- Beerenwinkel N and Siebourg J. Statistics, probability, and computational science. In M. Anisimova, editor, *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*, chapter 3, pages 77–110. Springer, New York, 2012. Section 3.