Expectation-Maximization & Baum-Welch

Slides: Roded Sharan, Jan 15; revised by Ron Shamir, Nov 15

CG

The goal

- Input: incomplete data originating from a probability distribution with some unknown parameters
- Want to find the parameter values that maximize the likelihood
- EM – approach that helps when maximum likelihood solution cannot be directly computed.
- Seeks a local maximum by iteratively solving two easier subproblems

Coin flipping: complete data

- Coins A, B with unknown heads probs. θ_A, θ_B
- Goal: estimate θ_A, θ_B
- Experiment: Repeat x5: choose A or B with prob. 1/2, flip 10 times, record results.

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

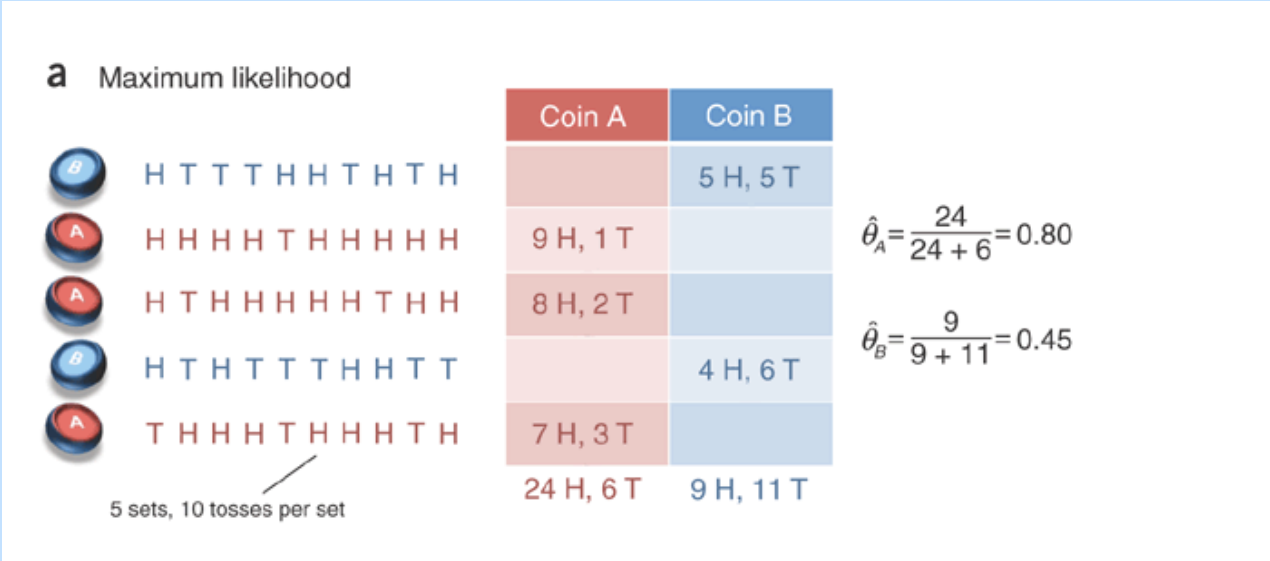
$x=(x_1,\ldots x_5)$: no of H in set 1,…5

$y=(y_1,\ldots y_5)$: coin used in set 1,…5

Do & Batzoglou, NBT 08

Coin flipping: complete data

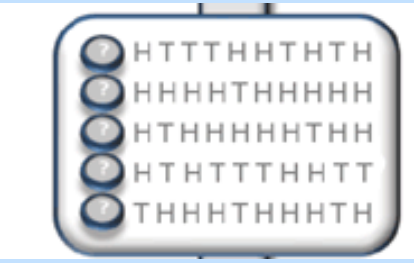
- Natural guess: θ_i = fraction of H in flips of coin i
- This is actually the ML solution: maximizes $P(x,y|\theta)$ (ex.)



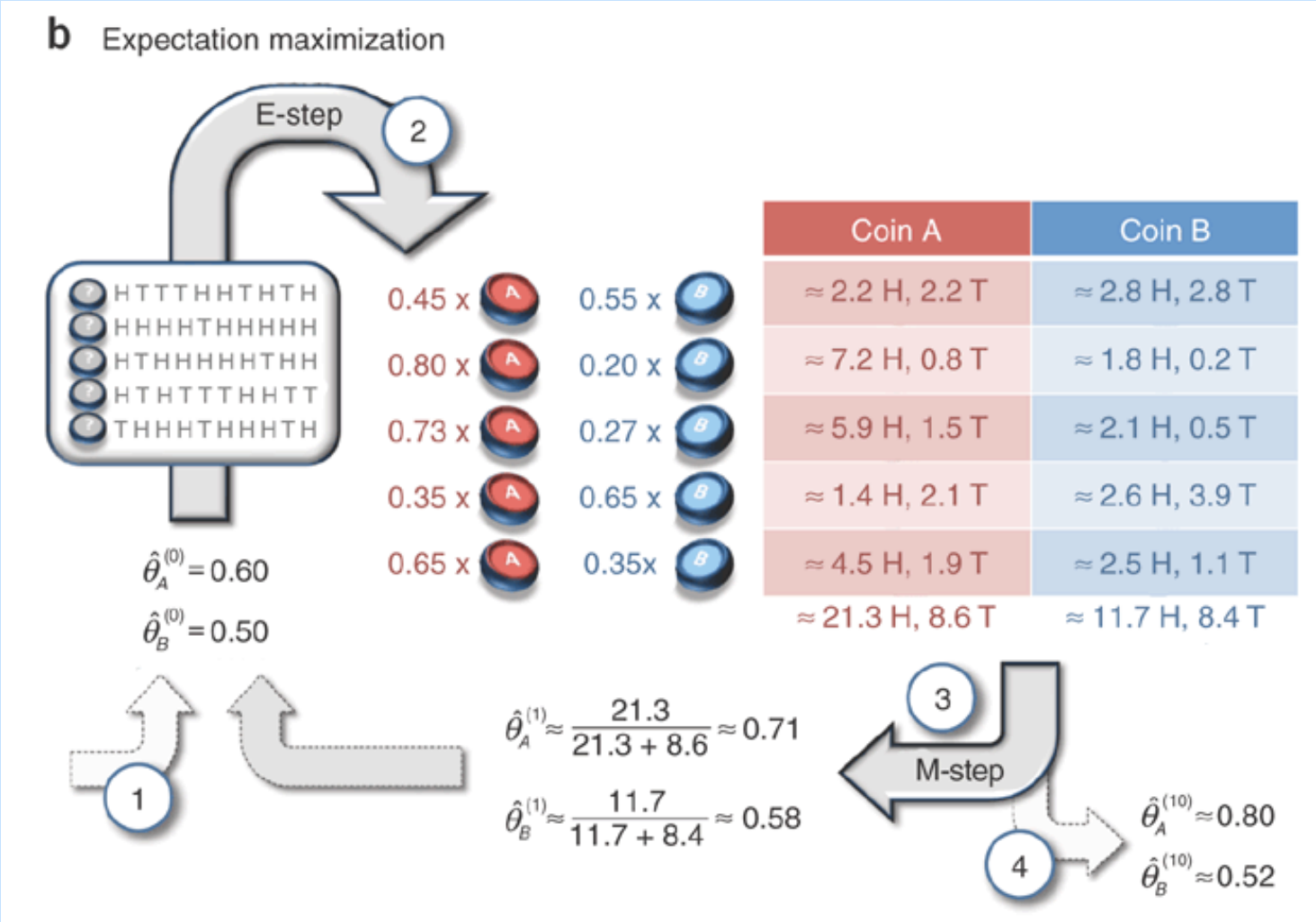
- What if we do not know which coin was used in each round?

Coin flipping: incomplete data

- Now $(y_1,\ldots y_5)$ are hidden / latent variables.
- Cannot compute H prob for each coin
- If we guessed Y correctly – we could.
- Idea: guess initial θ^0_A, θ^0_B
  - Use θ^t_A, θ^t_B to compute the most likely coin for each set, get new y
  - Use the resulting y to recompute θ_A, θ_B using ML, get θ^{t+1}_A, θ^{t+1}_B
  - Repeat till convergence
- EM: use probabilities rather than the single most likely completion y



Coin flipping: incomplete data

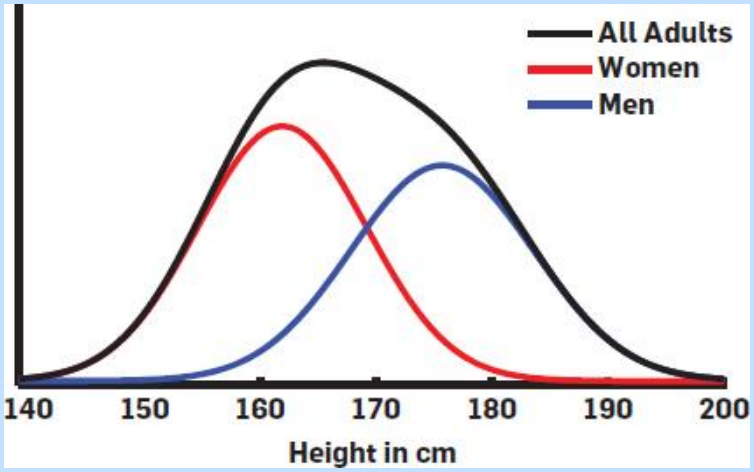


The probabilistic setting

<u>Input:</u> data X coming from a probabilistic model with hidden information y

<u>Goal:</u> Learn the model's parameters θ so that the likelihood is maximized.

Mixture of two Gaussians



Kalai et al. Disentangling Gaussians CACM 2012

Our input generates the black distribution. We want to color each sample red/blue and find the parameters of the two distributions to maximize the data probability. (assume σ known)

$$P(y_i = 1) = p_1 \; ; P(y_i = 2) = p_2 = 1 - p_1$$

$$P(x_i \mid y_i = j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu_j)^2}{2\sigma^2} \right)$$

$$\theta = (p_1, \mu_1, \mu_2)$$

CG

The likelihood function

$$P(y_i = 1) = p_1 \; ; P(y_i = 2) = p_2 = 1 - p_1$$

$$P(x_i \mid y_i = j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu_j)^2}{2\sigma^2} \right)$$

$$L(\theta) = \prod_i P(x_i \mid \theta) = \prod_i \sum_j P(x_i, y_i = j \mid \theta)$$

$$\log L(\theta) = \sum_i \log\left(\sum_j \frac{p_j}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu_j)^2}{2\sigma^2} \right) \right)$$

To be continued…

CG

KL divergence

Def: The Kullback-Liebler divergence (aka relative entropy) of discrete probability distributions P and Q:

$$D_{KL}(P \| Q) = \sum_{i \in +} P(x_i) \cdot \log \frac{P(x_i)}{Q(x_i)}$$

$i \in +$: sum over x s.t P(x)>0

Q(x)=0 → P(x)=0, 0log0=0

Lemma: KL divergence is nonnegative

log(x)≤x-1 for all x>0 with equality iff x=1

$$-D_{KL}(P \| Q) = \sum_{i \in +} P(x_i) \cdot \log \frac{Q(x_i)}{P(x_i)} \leq \sum_{i \in +} P(x_i) \cdot \left(\frac{Q(x_i)}{P(x_i)} - 1 \right)$$

$$= \sum_{i \in +} Q(x_i) - \sum_{i \in +} P(x_i) = \sum_{i \in +} Q(x_i) - 1 \leq 0$$

CG

With equality iff P≡Q

The EM algorithm (i)

Goal: max log P(x|θ)=log (Σ P(x,y|θ))

Strategy: guess an initial θ and iteratively adjust it, making sure that the likelihood always improves.
Assume we have a model θ^t that we wish to improve to a new value.

Bayes rule: P(x|θ) = P(x,y|θ) / P(y|x,θ)

Take log and multiply both sides by $P(y \mid x, \theta^t)$

$$P(y \mid x, \theta^t) \cdot \log P(x \mid \theta) = P(y \mid x, \theta^t) \cdot \log P(x, y \mid \theta) - P(y \mid x, \theta^t) \cdot \log P(y \mid x, \theta)$$

$$\sum_y P(y \mid x, \theta^t) \cdot \log P(x \mid \theta) = \sum_y P(y \mid x, \theta^t) \cdot \log P(x, y \mid \theta) - \sum_y P(y \mid x, \theta^t) \cdot \log P(y \mid x, \theta)$$

$$\log P(x \mid \theta) = \sum_y P(y \mid x, \theta^t) \cdot \log P(x, y \mid \theta) - \sum_y P(y \mid x, \theta^t) \cdot \log P(y \mid x, \theta)$$

The EM algorithm (ii)

$$\log P(x\,|\,\theta) = \sum_y P(y\,|\,x,\theta^t)\cdot\log P(x,y\,|\,\theta) - \sum_y P(y\,|\,x,\theta^t)\cdot\log P(y\,|\,x,\theta)$$

$$\log P(x\,|\,\theta^t) = \sum_y P(y\,|\,x,\theta^t)\cdot\log P(x,y\,|\,\theta^t) - \sum_y P(y\,|\,x,\theta^t)\cdot\log P(y\,|\,x,\theta^t)$$

Want $\quad P(x\,|\,\theta) > P(x\,|\,\theta^t)$

Define $\quad \Delta = \log P(x|\theta) - \log P(x|\theta^t)$

Define $\quad Q(\theta\,|\,\theta^t) = \sum_y P(y\,|\,x,\theta^t)\cdot\log P(x,y\,|\,\theta)$

$$\Delta = Q(\theta\,|\,\theta^t) - Q(\theta^t\,|\,\theta^t) + \sum_y P(y\,|\,x,\theta^t)\cdot\underbrace{\log\frac{P(y\,|\,x,\theta^t)}{P(y\,|\,x,\theta)}}_{\text{KL Divergence } \geq 0}$$

$$\Rightarrow\ \Delta \geq\ Q(\theta\,|\,\theta^t) - Q(\theta^t\,|\,\theta^t)$$

CG

The EM algorithm (iii)

Main component:

$$Q(\theta \mid \theta^t) = \sum_y P(y \mid x, \theta^t) \cdot \log P(x, y \mid \theta)$$

log P(x,y|θ) is called the complete log likelihood function

➔ Q is the expectation of the complete log likelihood over the distribution of y given the current parameters $\theta^\dagger$

<u>The algorithm:</u>

repeat

- E-step: Calculate the Q function
- M-step: Maximize $Q(\theta|\theta^\dagger)$ with respect to θ
- Stopping criterion: improvement in log likelihood ≤ ε

Note: local optimum guaranteed to be reached, not global.
Starting point matters! Try many..

CG

Back to the Gaussian mixture model

$$Q(\theta \mid \theta^t) = \sum_y P(y \mid x, \theta^t) \cdot \log P(x, y \mid \theta)$$

$$P(x, y \mid \theta) = \prod_i P(x_i, y_i \mid \theta) = \prod_i \prod_j P(x_i, y_i = j \mid \theta)^{y_{ij}}$$

$$y_{ij} = \begin{cases} 1 & y_i = j \\ 0 & y_i \neq j \end{cases}$$

$$\log P(x, y \mid \theta) = \sum_i \sum_j y_{ij} \log P(x_i, y_i = j \mid \theta)$$

$$Q(\theta \mid \theta^t) = \sum_y P(y \mid x, \theta^t) \sum_i \sum_j y_{ij} \log P(x_i, y_i = j \mid \theta) =$$

$$\sum_i \sum_j \sum_y P(y \mid x, \theta^t) y_{ij} \log P(x_i, y_i = j \mid \theta)$$

CG

Application (cont.)

$$Q(\theta \mid \theta^t) = \sum_i \sum_j P(y_{ij} = 1 \mid x_i, \theta^t) \log P(x_i, y_i = j \mid \theta)$$

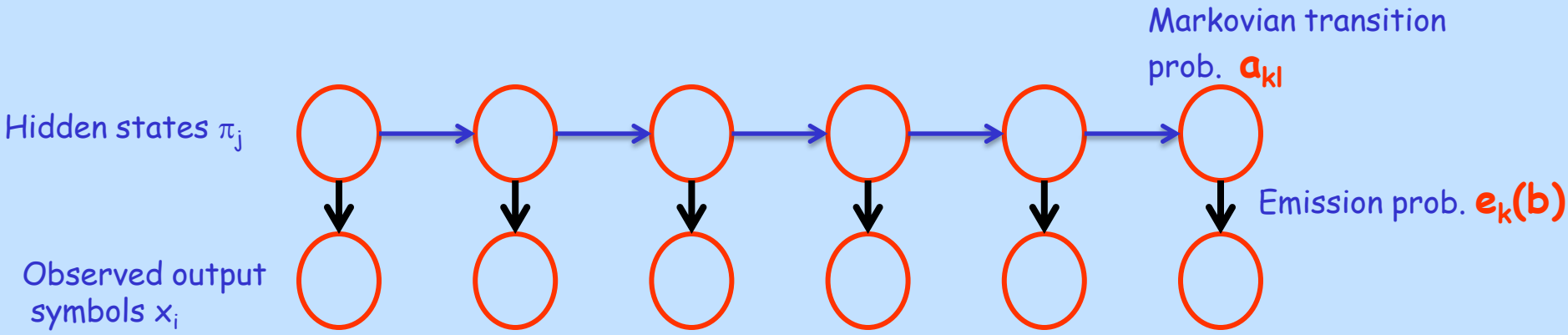
$$w_{ij}^t := P(y_{ij} = 1 \mid x_i, \theta^t) = \frac{P(x_i, y_i = j \mid \theta^t)}{\sum_j P(x_i, y_i = j \mid \theta^t)}$$

$$Q(\theta \mid \theta^t) = \sum_i \sum_j w_{ij}^t \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma + \log p_j - \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

Now write the derivatives and equate to zero to get the optimal parameters $\theta^{t+1} = (\mu_1^{t+1}, \mu_2^{t+1}, p_1^{t+1})$

EM for HMM: The Baum-Welch algorithm

Reminder: HMM



Model=(Σ, Q, Θ)

path $\Pi=\pi_1,\ldots,\pi_M$

Given sequence $X = (x_1,\ldots,x_M)$:

- $a_{kl} = P(\pi_i=l \mid \pi_{i-1}=k)$,
- $e_k(b) = P(x_i=b \mid \pi_i=k)$

$$P(X,\Pi) = a_{\pi_0,\pi_1} \cdot \Pi_{i=1\ldots L} e_{\pi_i}(x_i) \cdot a_{\pi_i,\pi_{i+1}}$$

Goal: Finding path Π^* maximizing $P(X,\Pi)$

Max likelihood in HMM

- $y=\pi$, $\theta = (a_{kl}, e_k(b))$

→ the log likelihood is

$$\log P(x \mid \theta) = \log \sum_{\pi} P(x, \pi \mid \theta)$$

And the Q function is:

$$Q(\theta \mid \theta^t) = \sum_{\pi} P(\pi \mid x, \theta^t) \cdot \log P(x, \pi \mid \theta)$$

Computing Q

$$P(x,\pi \mid \theta) = \prod_{k=1}^{M} \prod_{b} [e_k(b)]^{E_k(b,\pi)} \cdot \prod_{k=1}^{M} \prod_{l=1}^{M} a_{kl}^{A_{kl}(\pi)}$$

Emission probability, state k character b

Transition probability, state k to state l

Number of times we saw b from k in path π

Number of transitions from k to l in path π

CG

Computing Q (ii)

$$Q(\theta \mid \theta^t) = \sum_{\pi} P(\pi \mid x, \theta^t) \cdot \left[\sum_{k=1}^{M} \sum_{b} E_k(b, \pi) \cdot \log(e_k(b)) + \sum_{k=1}^{M} \sum_{l=1}^{M} A_{kl}(\pi) \cdot \log a_{kl} \right] =$$

$$= \sum_{k=1}^{M} \sum_{b} \sum_{\pi} P(\pi \mid x, \theta^t) \cdot E_k(b, \pi) \cdot \log(e_k(b)) + \sum_{k=1}^{M} \sum_{l=1}^{M} \sum_{\pi} P(\pi \mid x, \theta^t) \cdot A_{kl}(\pi) \cdot \log a_{kl}$$

$$\sum_{\pi} P(\pi \mid x, \theta^t) \cdot E_k(b, \pi) = E_k(b)$$

probability value expectation

$$\sum_{\pi} P(\pi \mid x, \theta^t) \cdot A_{kl}(\pi) = A_{kl}$$

probability value expectation

CG

Computing Q (iii)

- So we want to find a set of parameters θ^{t+1} that maximizes:

$$\sum_{k=1}^{M}\sum_{b} E_k(b)\cdot\log(e_k(b)) + \sum_{k=1}^{M}\sum_{l=1}^{M} A_{kl}\cdot\log a_{kl}$$

$f_k(i) = P(x_0,\ldots,x_i,\ \pi_i=k)$

$b_k(i) = P(x_{i+1},\ldots x_L \mid \pi_i=k)$

- $E_k(b)$, A_{kl} can be computed using forward/backward:

$P(\pi_i=k,\ \pi_{i+1}=l \mid x,\ \Theta^t) = [1/P(x)]\cdot f_k(i)\cdot a_{kl}\cdot e_l(x_{i+1})\cdot b_l(i+1)$

$A_{kl} = [1/P(x)]\cdot \sum_i f_k(i)\ \cdot\ a_{kl}\ \cdot\ e_l(x_{i+1})\ \cdot\ b_l(i+1)$

similarly, $E_k(b) = [1/P(x)]\ \cdot\ \sum_{\{i|x_i=b\}} f_k(i)\ \cdot\ b_k(i)$

- For maximization, select:

$$a_{ij} = \frac{A_{ij}}{\sum_k A_{ik}} \quad , \quad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

CG

Baum-Welch: EM for HMM

Maximize: $$\sum_{k=1}^{M}\sum_{b} E_k(b)\cdot\log(e_k(b)) + \sum_{k=1}^{M}\sum_{l=1}^{M} A_{kl}\cdot\log a_{kl}$$

Multiply and divide by same factor

$$a_{ij} = \frac{A_{ij}}{\sum_k A_{ik}} \quad (\text{denote as } a_{ij}^{\text{chosen}}), \quad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Difference between chosen set and some other:

$$\sum_{k=1}^{M}\sum_{l=1}^{M} A_{kl}\cdot\log\left(\frac{a_{kl}^{chosen}}{a_{kl}^{other}}\right) = \sum_{k=1}^{M}\sum_{k'} A_{kk'}\sum_{l=1}^{M}\frac{A_{kl}}{\sum_{k'} A_{kk'}}\cdot\log\left(\frac{a_{kl}^{chosen}}{a_{kl}^{other}}\right) =$$

$$= \sum_{k=1}^{M}\sum_{k'} A_{kk'}\sum_{l=1}^{M} a_{kl}^{chosen}\cdot\log\left(\frac{a_{kl}^{chosen}}{a_{kl}^{other}}\right)$$

➔ always positive

CG

Summary: Parameter Estimation in HMM When States are Unknown

Input: $X^1,\ldots,X^n$ indep training sequences

Baum-Welch alg. (1972):

✶ **Expectation:**

- compute expected no. of $k \to l$ state transitions:

  $P(\pi_i=k,\ \pi_{i+1}=l \mid X,\ \Theta) = [1/P(x)] \cdot f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)$

  $A_{kl} = \sum_j [1/P(X^j)] \cdot \sum_i f_k^j(i) \cdot a_{kl} \cdot e_l(x^j_{i+1}) \cdot b_l^j(i+1)$

- compute expected no. of symbol b appearances in state k

  $E_k(b) = \sum_j [1/P(X^j)] \cdot \sum_{\{i \mid x^j_i=b\}} f_k^j(i) \cdot b_k^j(i)$ *(ex.)*

✸ **Maximization:**

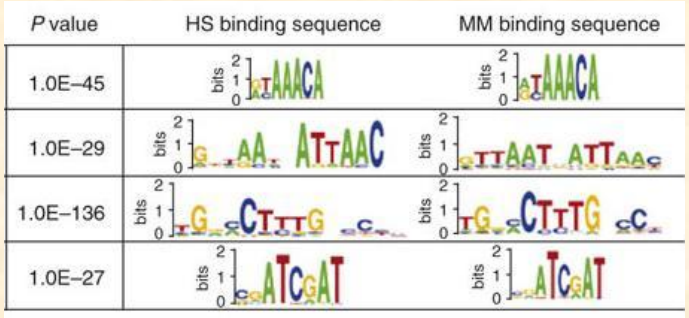
- re-compute new parameters from A, E using max. likelihood.

repeat (1)+(2) until improvement $\leq \varepsilon$

Leonard Baum, many years after the IDA

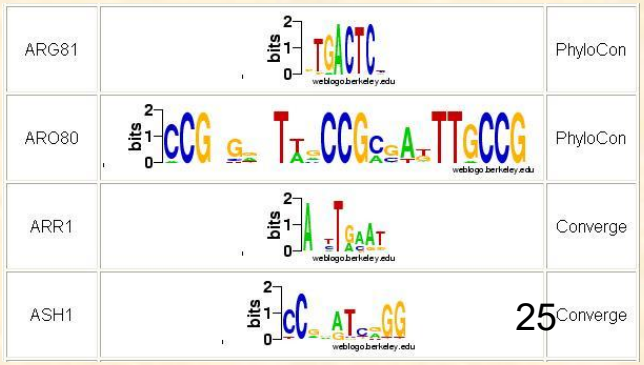
Lloyd Welch, USC Electrical Engineering

CG



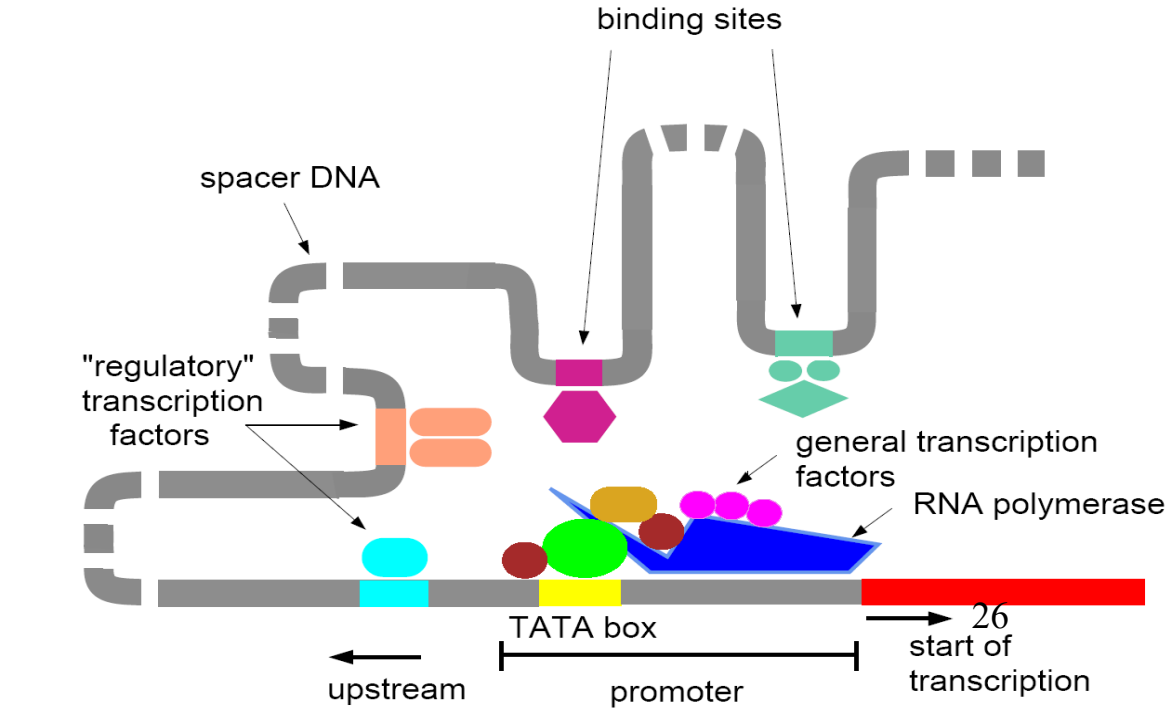
de novo motif discovery using EM

Slides sources:
Chaim Linhart, Danit Wider, Katherina Kechris



Transcription Factors

- A transcription factor (TF) is a protein that regulates a gene by binding to a binding site (BS) in its vicinity, specific to the TF.
- Binding sites vary in their sequences. Their sequence pattern is called a **motif**.



Motif profile

Alignment

a	G	g	t	a	c	T	t
C	c	A	t	a	c	g	t
a	c	g	t	T	A	g	t
a	c	g	t	C	c	A	t
C	c	g	t	a	c	g	G

Profile

A	3	0	1	0	3	1	1	0
C	2	4	0	0	1	4	0	0
G	0	1	4	0	0	0	3	1
T	0	0	0	5	1	0	1	4

- Line up the patterns by their start indexes

$$\mathbf{s} = (s_1, s_2, \ldots, s_t)$$

- Construct matrix profile with frequencies of each nucleotide in columns

<u>Motif finding:</u> Given a set of co-regulated genes, find a recurrent motif in their promoter regions.

An example: Implanting Motif AAAAAAAGGGGGGG

atgaccgggatactgatAAAAAAAAGGGGGGGgcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa

tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag

gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa

Where is the Implanted Motif? (*)

atgaccgggatactgataaaaaaaaggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataaaaaaaaggggggga

tgagtatccctgggatgacttaaaaaaaaggggggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgaaaaaaaaggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaaggggggcttatag

gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtaaaaaaaaggggggcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaaaaggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaaggggggaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga

Implanting Motif **AAAAAAGGGGGGG** with Four Mutations

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

Where is the Motif???

atgaccgggatactgatagaagaaaggttgggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

accccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacaataaaacggcggga

tgagtatccctgggatgacttaaaataatggagtggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcaaaaaaagggattgtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatataataaaggaagggcttatag

gtcaatcatgttcttgtgaatggatttaacaataagggctgggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtataaacaaggagggccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaatagggagccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatactaaaaaggagcggaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttactaaaaaggagcgga

MEME

Multiple EM for Motif Elicitation

[Bailey, Elkan ISMB ’94]

Goal: Given a set of sequences, find a motif (PWM) that maximizes the expected likelihood of the data

Technique: EM (Expectation Maximization)

(based on [Lawrence, Reilly ’90])

The Mixture Model

Data: $X = (X_1,\ldots,X_n)$ **:**

all (overlapping) *l*-mers in the input sequences

Assume X_i's were generated by a two-component mixture model - $\theta = (\theta_1, \theta_2)$:

Model #1: θ_1 = <u>motif model</u>:

$f_{i,b}$ **= prob. of base *b* at pos *i* in motif, $1 \le i \le l$**

Model #2: θ_2 = <u>background (BG) model</u>:

$f_{0,b}$ **= prob. of base *b***

<u>Mixing</u> parameter: $\lambda = (\lambda_1, \lambda_2)$

λ_j **= prob. that model #*j* is used ($\lambda_1+\lambda_2=1$)**

Assume independence between *l*-mers

Log Likelihood

Missing data: $Z = (Z_1,\ldots,Z_n)$:

$Z_i = (Z_{i1}, Z_{i2})$; $Z_{ij} = 1$ if X_i from model #j ; 0 o/w

Complete Likelihood of model given data:

$$L(\theta, \lambda \mid X, Z) = p(X, Z \mid \theta, \lambda)$$

$$= \Pi_{i=1\ldots n}\, p(X_i, Z_i \mid \theta, \lambda)$$

$$p(X_i, Z_i \mid \theta, \lambda) = p(X_i \mid Z_i, \theta, \lambda)\, p(Z_i) =$$

$$= \lambda_1\, p(X_i|\theta_1) \text{ if } Z_{i1}=1;\ \lambda_2\, p(X_i|\theta_2) \text{ if } Z_{i2}=1$$

$$\rightarrow \log L = \Sigma_{i=1\ldots n} \Sigma_{j=1,2} Z_{ij} \log(\lambda_j\, p(X_i|\theta_j))$$

MEME: Algorithm

Goal: **Maximize** ***E[log L]***

Outline of EM algorithm:

- **Choose starting** θ, λ
- **Repeat until convergence of** θ**:**
  - **E-step: Re-estimate** Z **from** θ, λ, X
  - **M-step: Re-estimate** θ, λ **from** X, Z
- **Repeat all of the above for various** θ, λ **…**

E-step

Compute expectation of *log L* over Z:

$$E[\log L] = \Sigma_{i=1\ldots n} \Sigma_{j=1,2} Z'_{ij} \log (\lambda_j\, p(X_i|\theta_j))$$

where:

$$Z'_{ij} = p(Z_{ij}=1|\, \theta,\lambda,X_i) =$$

$$= p(Z_{ij}=1,\, X_i|\, \theta,\lambda) \,/\, p(X_i|\, \theta,\lambda) =$$

$$= p(Z_{ij}=1,\, X_i|\, \theta,\lambda) \,/\, \Sigma_{k=1,2}\, p(Z_{ik}=1,\, X_i|\, \theta,\lambda) =$$

$$= \lambda_j\, p(X_i|\theta_j) \,/\, \Sigma_{k=1,2}\, \lambda_k\, p(X_i|\theta_k)$$

M-step

Find θ,λ that maximize E[$\log L$]=$Q(\theta,\lambda|\theta^t,\lambda^t)$:

$$E[\log L] = \Sigma_{i=1\ldots n}\Sigma_{j=1,2} Z'_{ij} \log(\lambda_j\, p(X_i|\theta_j))$$

Finding λ:

Suffices to maximize $L_1 = \Sigma_{i=1\ldots n}\Sigma_{j=1,2} Z'_{ij} \log \lambda_j$

$$\lambda_1+\lambda_2=1 \rightarrow L_1 = \Sigma_{i=1\ldots n}(Z'_{i1}\log\lambda_1 + Z'_{i2}\log(1-\lambda_1))$$

$$dL_1/d\lambda_1 = \Sigma_{i=1\ldots n}(Z'_{i1}/\lambda_1 - Z'_{i2}/(1-\lambda_1))$$

MEME: Algorithm

M-step (cont.):

$$dL_1/d\lambda_1 = \Sigma_{i=1\ldots n} (Z'_{i1} / \lambda_1 - Z'_{i2} / (1-\lambda_1)) = 0$$

$$\rightarrow \lambda_1 \Sigma_{i=1\ldots n} Z'_{i2} = (1-\lambda_1) \Sigma_{i=1\ldots n} Z'_{i1}$$

$$\rightarrow \lambda_1 (\Sigma_{i=1\ldots n} (Z'_{i1}+Z'_{i2})) = \Sigma_{i=1\ldots n} Z'_{i1}$$

$$\rightarrow \lambda_1 = (\Sigma_{i=1\ldots n} Z'_{i1}) / n$$

$$\lambda_2 = 1- \lambda_1 = (\Sigma_{i=1\ldots n} Z'_{i2}) / n$$

Finding θ: …

MEME Suite Menu
- Submit A Job
- Documentation
- Downloads
- User Support
- Alternate Servers
- Authors
- Citing
- New! Postdoc Available

MEME
Multiple Em for Motif Elicitation

Version 4.3.0

Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (**motif**) for each pattern it discovers.

Data Submission Form

Required

Your **e-mail address:**

Re-enter **e-mail address:**

Please enter the **sequences** which you believe share one or more
motifs. The sequences may contain no more than **60000 characters**
total in any of a large number of **formats**.

Enter the **name of a file** containing the sequences here:

Browse... Clear

or
the **actual sequences** here (**Sample Protein Input Sequences**):

How do you think the occurrences of a single motif are **distributed** among the sequences?

- ○ **One per sequence**
- ◉ **Zero or one** per sequence
- ○ **Any number** of repetitions

MEME will find the optimum **width** of each motif within the limits you specify here:

6 **Minimum** width (>= 2)

50 **Maximum** width (<= 300)

3 Maximum **number of motifs** to find

Optional

Description of your sequences:

MEME will find the optimum **number of sites** for each motif within the limits you specify here:

Minimum sites (>= 2)

Maximum sites (<= 300)

☐ **Shuffle** sequence letters

Enter the name of a file containing a **background Markov model**:

Browse... Clear

DNA-ONLY OPTIONS
(Ignored for protein searches)

☐ Search given **strand** only

☐ Look for **palindromes** only

Start search Clear Input

Version 4.3.0 **Please send comments and questions to: meme@nbcr.net** **Powered by Opal**

Home Submit a Job Documentation Downloads User Support Alternate Servers Authors Citing

Tim Bailey, Charles Elkan

- **Senior Research Fellow Institute for Molecular Bioscience , University of Queensland, Brisbane, Australia**

- **Professor Department of Computer Science and Engineering University of California, San Diego**

FIN