

CSE 527

Lecture 7

Relative entropy
Convergence of EM
Weight matrix motif models

Talk Today

COMBI Seminar Today:

Dr. David Baker

“Progress in High-Resolution Modeling
of Protein Structure and Interactions”

Today, October 19, 2005

1:30-2:30

HSB K-069

Relative Entropy

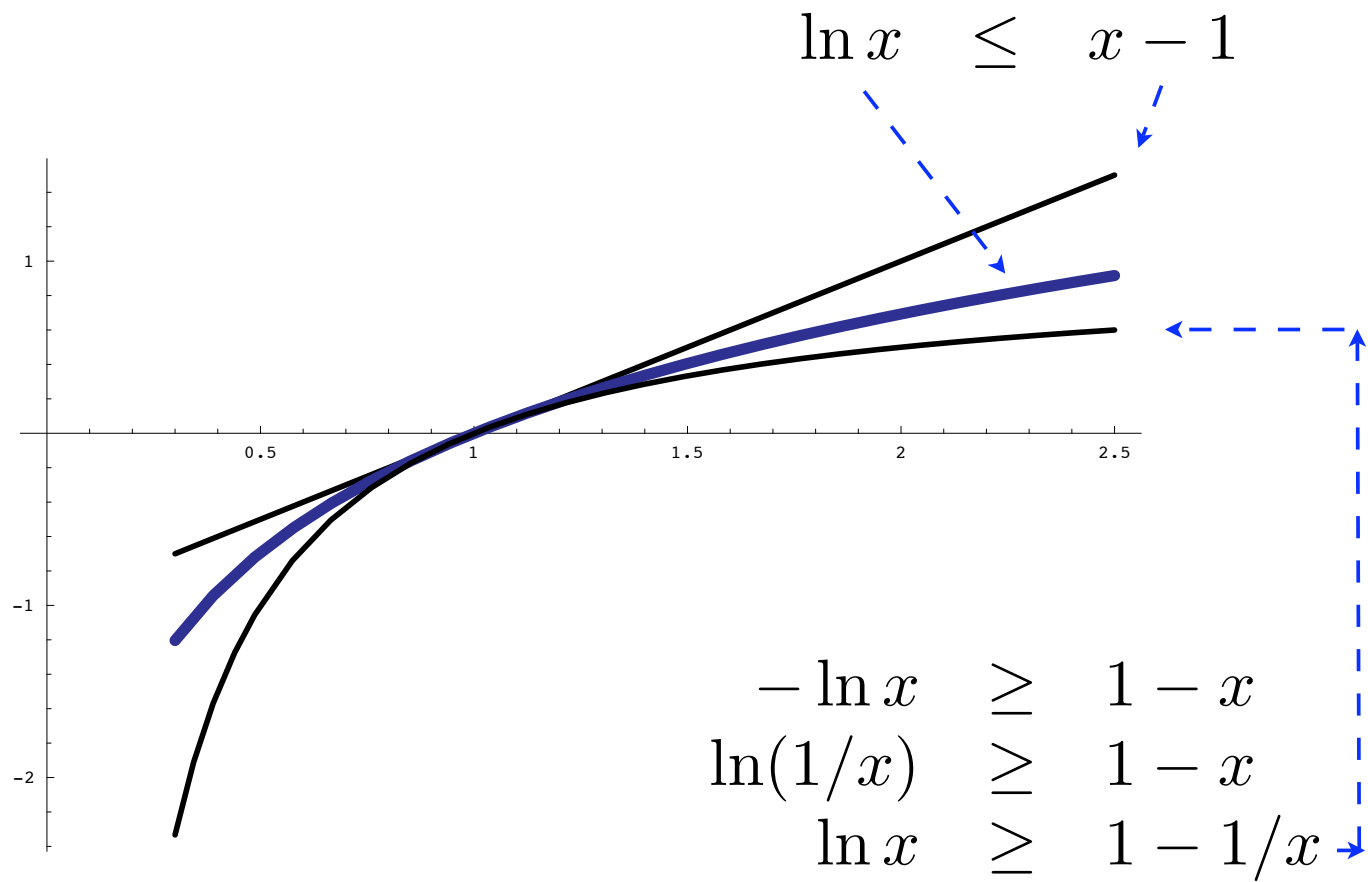
- AKA Kullback-Liebler Distance/Divergence, AKA Information Content
- Given distributions P, Q

$$H(P||Q) = \sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)}$$

Notes:

Let $P(x) \log \frac{P(x)}{Q(x)} = 0$ if $P(x) = 0$ [since $\lim_{y \rightarrow 0} y \log y = 0$]

Undefined if $0 = Q(x) < P(x)$



Theorem: $H(P||Q) \geq 0$

$$\begin{aligned} H(P||Q) &= \sum_x P(x) \log \frac{P(x)}{Q(x)} \\ &\geq \sum_x P(x) \left(1 - \frac{Q(x)}{P(x)}\right) \\ &= \sum_x (P(x) - Q(x)) \\ &= \sum_x P(x) - \sum_x Q(x) \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

Furthermore: $H(P||Q) = 0$ if and only if $P = Q$

EM Convergence

Visible x
hidden y
Parameters Θ

Goal Maximum likelihood estimate of Θ
i.e. Find Θ maximizing $\Pr(x|\Theta)$ (or $\log P(x|\Theta)$)

$$P(y|x) = P(x,y)/P(x) \text{ so } P(x) = P(x,y)/P(y|x)$$

$\forall y$:

$$\log P(x|\Theta) = \log P(x,y|\Theta) - \log P(y|x,\Theta)$$

$$\log P(x|\Theta) =$$

$$\underbrace{\sum_y P(y|x,\Theta^*) \cdot \log P(x,y|\Theta)}_{Q(\Theta|\Theta^*)} - \sum_y P(y|x,\Theta^*) \cdot \log P(y|x,\Theta)$$

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

A key trick: Q is easier to optimize than whole thing.

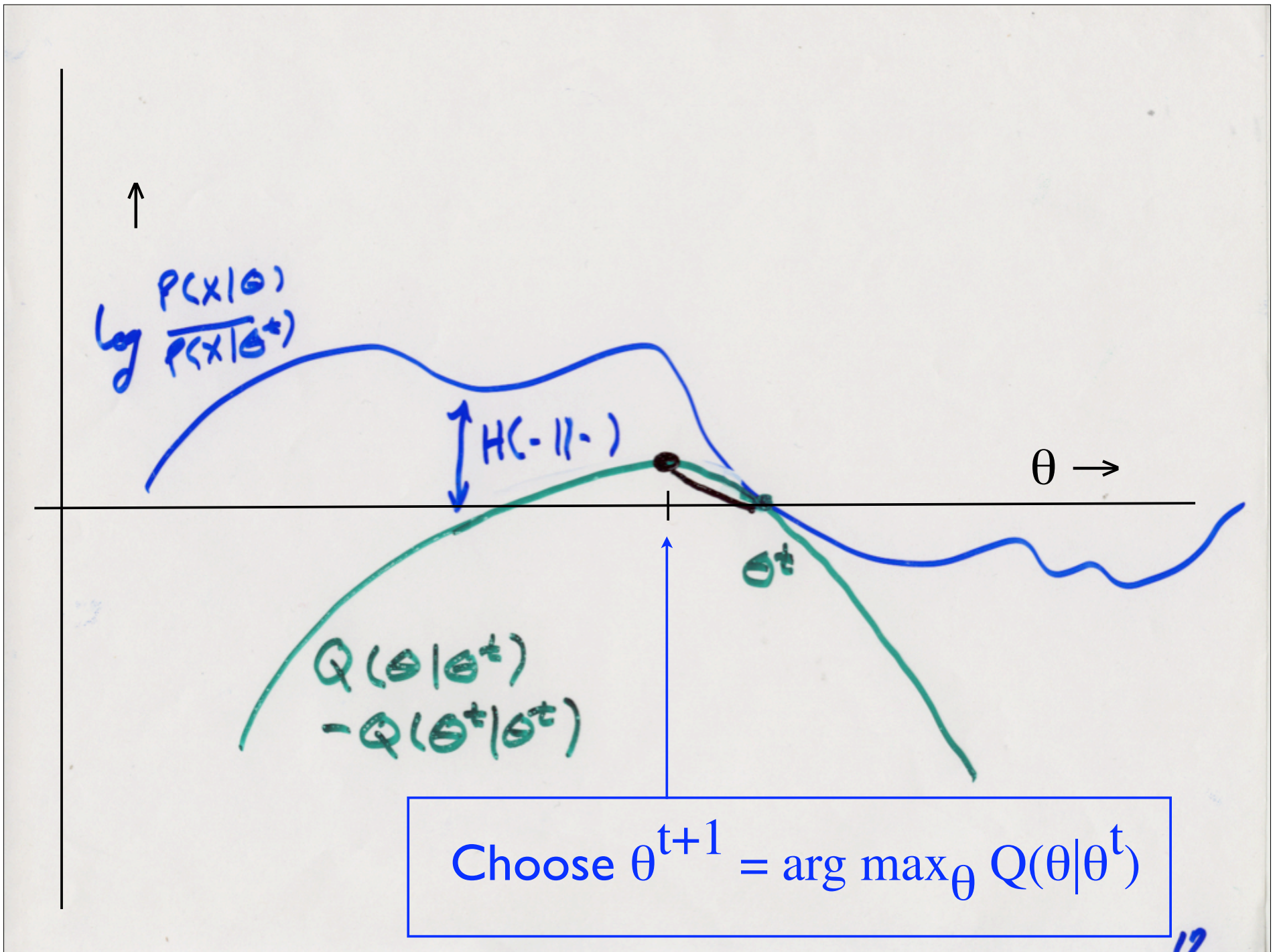
$$\textcircled{1} \quad \log P(x|\theta) - \log P(x|\theta^t) =$$

$$\textcircled{2} \quad Q(\theta|\theta^t) - Q(\theta^t|\theta^t)$$

$$+ \sum_y P(y|x, \theta^t) \log \frac{P(y|x, \theta^t)}{P(y|x, \theta)}$$

$$H(P(y|x, \theta^t) \parallel P(y|x, \theta)) \geq 0$$

$$\therefore \textcircled{1} \geq 0 \text{ if } \textcircled{2} \geq 0$$



Sequence Motifs

E. coli Promoters

- “**TATA Box**” - consensus TATAAT ~ 10bp upstream of transcription start
- *Not exact*: of 168 studied
 - nearly all had 2/3 of TAx_{yz}T
 - 80-90% had all 3
 - 50% agreed in each of x,y,z
 - **no** perfect match
- Other common features at -35, etc.

TATA Box Frequencies

pos base	1	2	3	4	5	6
A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

Scanning for TATA

-38	19	1	12	10	-48
-15	-38	-8	-10	-3	-32
-13	-48	-6	-7	-10	-48
17	-32	8	-9	-6	19

= -93

A C T A T A A T C G

-38	19	1	12	10	-48
-15	-38	-8	-10	-3	-32
-13	-48	-6	-7	-10	-48
17	-32	8	-9	-6	19

= 85

A C T A T A A T C G

-38	19	1	12	10	-48
-15	-38	-8	-10	-3	-32
-13	-48	-6	-7	-10	-48
17	-32	8	-9	-6	19

= -95

A C T A T A A T C G

Weight Matrices: Statistics

- Assume:

$f_{b,i}$ = frequency of base b in position i

f_b = frequency of base b in all sequences

- Log likelihood ratio, given $S = B_1B_2...B_6$:

$$\log\left(\frac{P(S | \text{“promoter”})}{P(S | \text{“nonpromoter”})}\right) = \log\left(\frac{\prod_{i=1}^6 f_{B_i,i}}{\prod_{i=1}^6 f_{B_i}}\right) = \sum_{i=1}^6 \log\left(\frac{f_{B_i,i}}{f_{B_i}}\right)$$

Weight Matrices: Chemistry

- Experiments show ~80% correlation of log likelihood weight matrix scores to measured binding energy of RNA polymerase to variations on TATAAT consensus [Stormo & Fields]