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

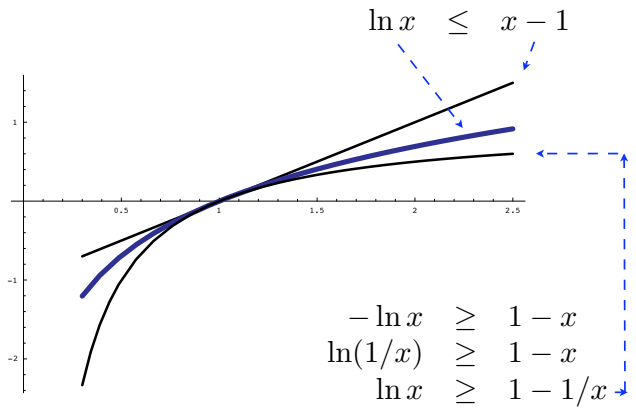
- AKA Kullback-Liebler Distance/Divergence, AKA Information Content
- Given distributions P, Q

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

Notes:

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

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

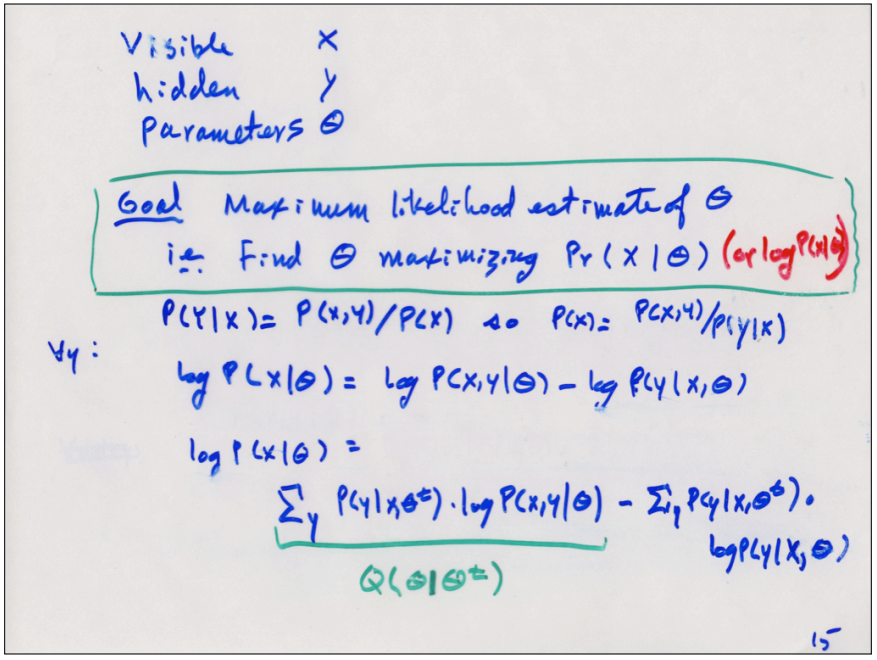


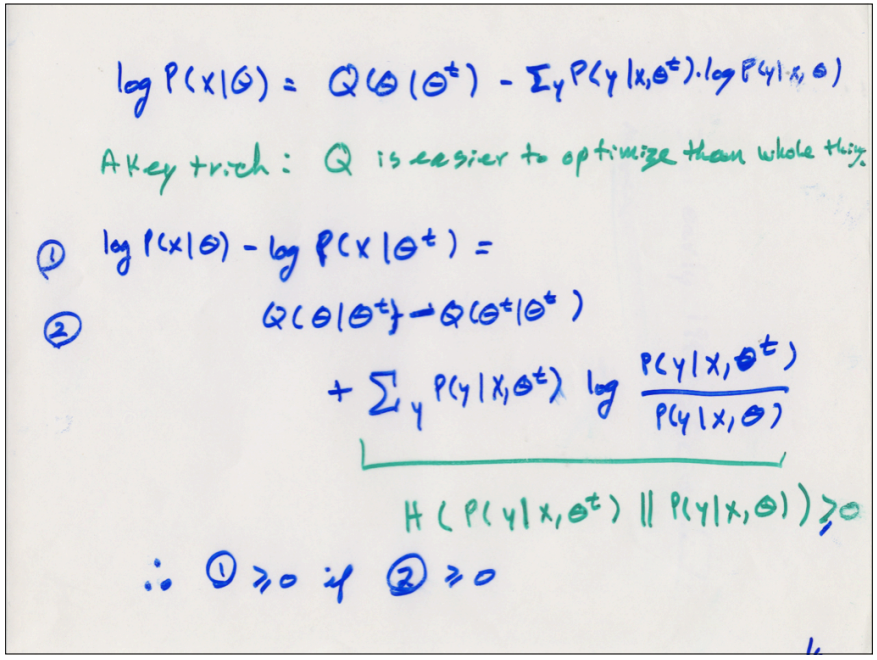
Theorem: $H(P||Q) \geq 0$

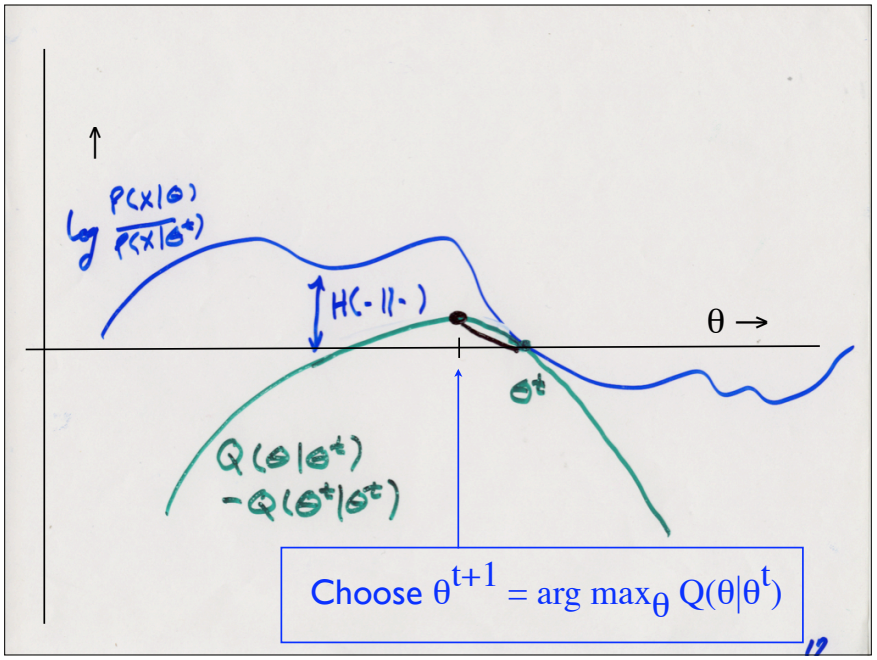
$$
\begin{aligned}
H(P||Q) &= \textstyle\sum_x P(x) \log \frac{P(x)}{Q(x)} \\
&\geq \textstyle\sum_x P(x) \left(1 - \frac{Q(x)}{P(x)}\right) \\
&= \textstyle\sum_x (P(x) - Q(x)) \\
&= \textstyle\sum_x P(x) - \sum_x Q(x) \\
&= 1 - 1 \\
&= 0
\end{aligned}
$$

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

EM Convergence







Sequence Motifs

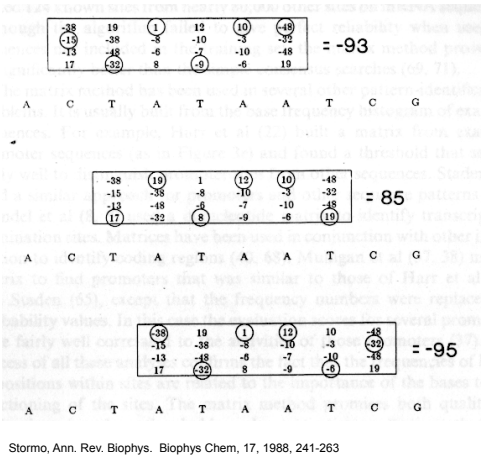
E. coli Promoters

- "TATA Box" - consensus TATAAT ~ 10bp upstream of transcription start
- *Not* exact: of 168 studied
  - nearly all had 2/3 of TAxyzT
  - 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



Stormo, Ann. Rev. Biophys. Biophys Chem, 17, 1988, 241-263

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]