CSE 527
Lecture 7

Relative entropy
Convergence of EM
Weight matrix motif models

Talk this week

- COMBI/GS Seminar
  Thomas R. Gingeras, Ph.D.
  "Empirical Analysis of Sites of RNA Transcription for 30% of the Human Genome: The Changing Landscape of the Human Genome Annotations"

  Wednesday, October, 20, 2004
  3:30 pm, Hitchcock 132
- Refreshments in lobby at 3:20

Relative Entropy

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

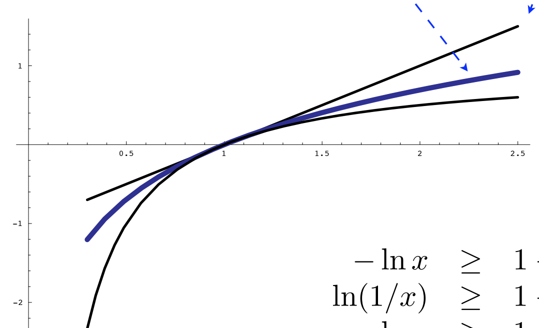
$$H(P||Q) = \sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)}$$

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

$$\ln x \leq x - 1$$



$$\begin{aligned} -\ln x &\geq 1 - x \\ \ln(1/x) &\geq 1 - x \\ \ln x &\geq 1 - 1/x \end{aligned}$$

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

Visible X

hidden y

Parameters θ

Goal Maximum likelihood estimate of θ

i.e. Find θ maximizing $\Pr(X \mid \theta)$ (or $\log P(X|\theta)$)

$P(Y|X) = P(X,Y)/P(X)$ so $P(X) = P(X,Y)/P(Y|X)$

$\forall y$:

$\log P(X|\theta) = \log P(X,Y|\theta) - \log P(Y|X,\theta)$

$\log P(X|\theta) =$

$$\underbrace{\sum_y P(y|X,\theta^t) \cdot \log P(X,y|\theta)}_{Q(\theta|\theta^t)} - \sum_y P(y|X,\theta^t) \cdot \log P(y|X,\theta)$$

15

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

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

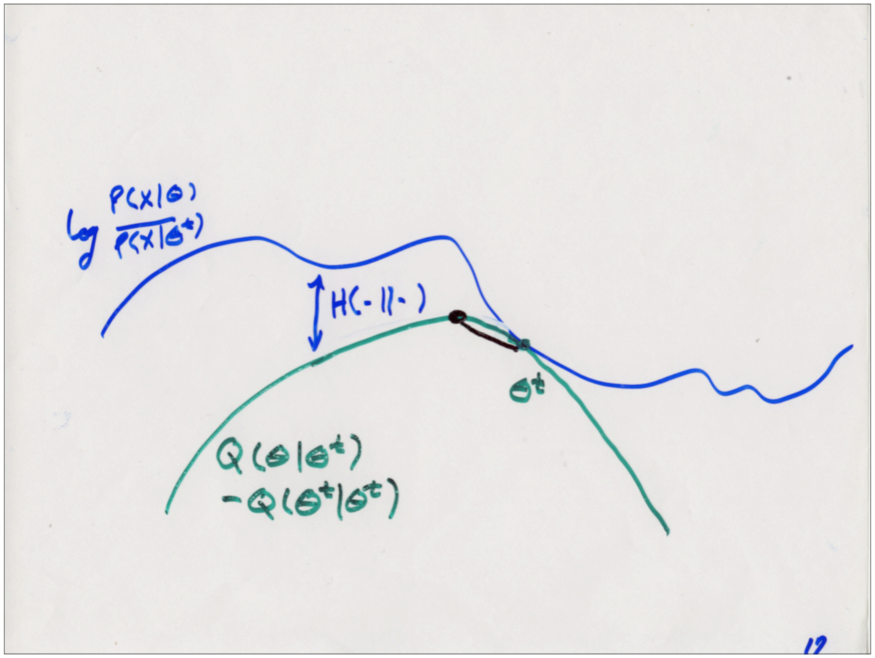
① $\log P(X|\theta) - \log P(X|\theta^t) =$

② $Q(\theta|\theta^t) - Q(\theta^t|\theta^t)$

$$+ \underbrace{\sum_y P(y|X,\theta^t) \log \frac{P(y|X,\theta^t)}{P(y|X,\theta)}}_{H(P(y|X,\theta^t) \,||\, P(y|X,\theta)) \geq 0}$$

$\therefore$ ① ≥ 0 if ② ≥ 0

16



Sequence Motifs

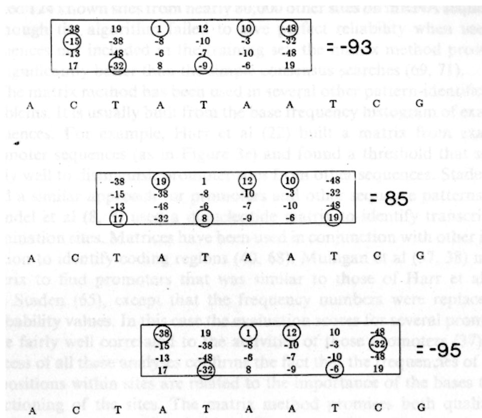
E. coli Promoters

- "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 $S = B_1 B_2 \ldots B_6$:

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

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