Visible x

hidden y

parameters θ

Goal Maximum likelihood estimate of θ

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

$P(y|x) = P(x,y)/P(x)$ so $P(x) = P(x,y)/P(y|x)$

$\forall y$:

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

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

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

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

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

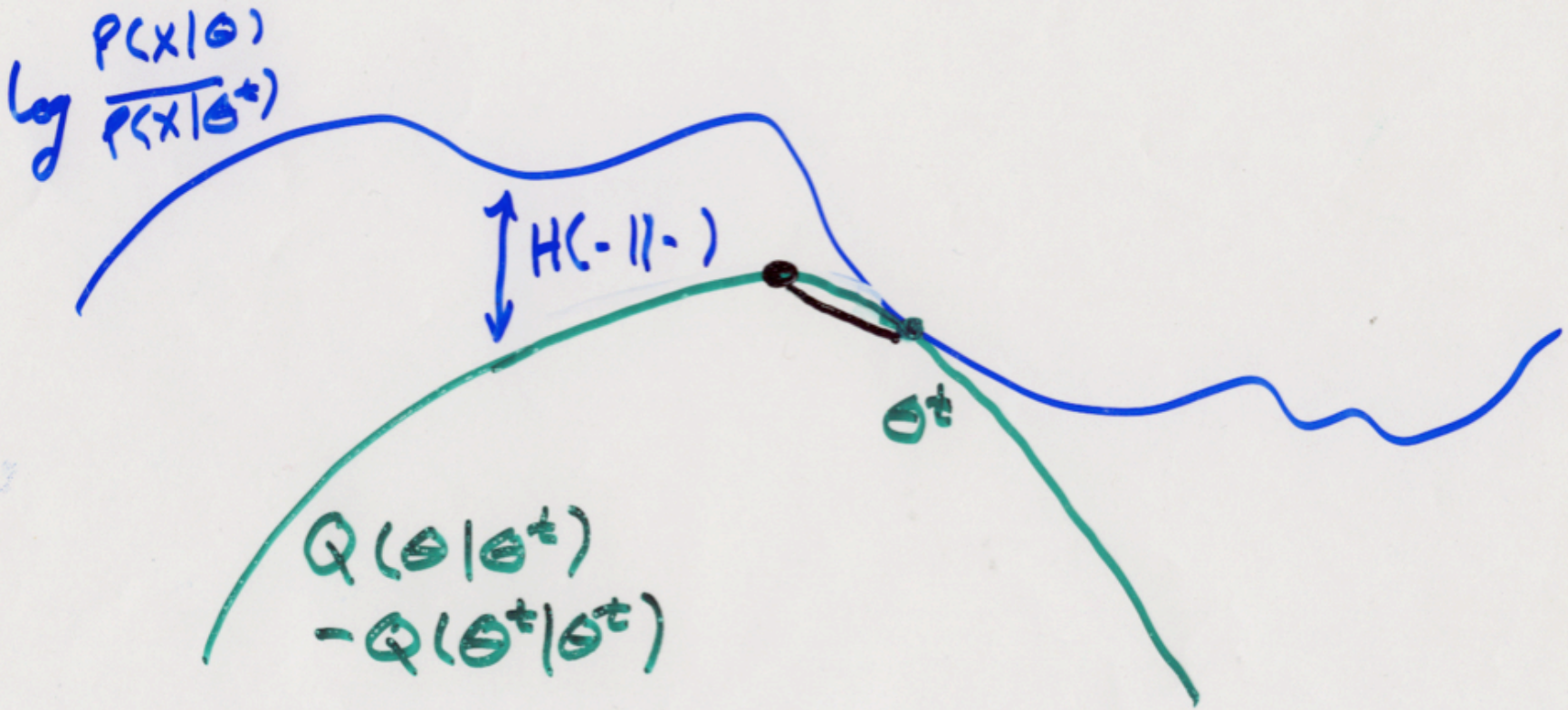
① $\log P(x|\theta) - \log P(x|\theta^t) =$

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

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

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

$\therefore$ ① ≥ 0 if ② ≥ 0


log P(x|θ) / P(x|θ^t)
H(·||·)
θ^t
Q(θ|θ^t)
-Q(θ^t|θ^t)

Class Monday 11/10

Zizhen Yao

Functional Prediction in E. Coli Based on Heterogenous Data

Weight Matrix Example

8 Sequences

A T G
A T G
A T G
A T G
A T G
G T G
G T G
T T G

Profile

	1	2	3
A	.625	0	0
C	0	0	0
G	.250	0	1
T	.125	1	0

Log Likelihood Ratio

	1	2	3
A	1.32	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	0	$-\infty$	2
T	-1	2	$-\infty$

$$\log_2 \frac{f_{x_i,i}}{f_{x_i}}, f_{x_i} = 1/4$$

Nonuniform Background

E.coli - DNA approximately 1/4 A, C, G, T

M. jannaschi - 68% A-T, 32% GC

LLR from Previous Example w $f_A = f_T = 3/8$
$f_C = f_G = 1/8$

	1	2	3
A	.937	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	1.00	$-\infty$	3
T	-1.58	1.42	$-\infty$

E.g. "G" in position 3 is $2^3 = 8\times$ more likely than background

How "Informative" is a WMM?

Recall Relative Entropy

$$H(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$

If x are sequences (fixed length)

$P(x)$ = Prob. of x according to WMM (or other model)

$Q(x)$ = ... Background model

$H(P||Q)$ is <u>expected log likelihood score</u>

of a randomly chosen site (random according to site model)

For WMM, can show

$$H(P \| Q) = \sum_{i=1}^{n} H(P_i \| Q_i)$$

where P_i, Q_i are distributions of i^{th} position

[Follows from assumption of independence]

Example (cont.)

	1	2	3
A	.625	0	0
C	0	0	0
G	.250	0	1
T	.125	1	0

	1	2	3
A	1.32	-∞	-∞
C	-∞	-∞	-∞
G	0	-∞	2
T	-1	2	-∞
Rel. Ent.	.201	2	2

uniform

	1	2	3
A	.737	-∞	-∞
C	-∞	-∞	-∞
G	1.00	-∞	3.00
T	-1.58	1.42	-∞
rel. Ent.	.512	1.42	3.0

non-uniform

Pseudo counts

Are the $-\infty$'s a problem?

- if you are certain a given residue never occurs in a given position, then $-\infty$ is just right
- If not, then its probably an artifact of small sample

Typical fix:

add a small constant (eg .5, 1, 2) to all observed counts – a pseudocount

Questions

- Given aligned instances of Motifs,
  How do you build model ?

  Frequency counts, as above

- Given model, how do you find
  (probable) instances ?

  Scanning

- Given unaligned strings thought
  to contain a motif, how do you find it?

  Eg. upstream regions from μarray cluster

Motif Discovery
Three Approaches

1. Greedy Search
2. Expectation Maximization
3. Gibb Sampler

P.S. Finding the site of max relative entropy in a set of unaligned sequences is NP-hard (AKutsu)

GREEDY Algorithm [Hertz & Stormo]

Input:
Sequences $S_1 .. S_K$, motif length l, "breadth" d,
& Background

Algorithm

1. Create a singleton set with each length l subsequence of each of $S_1 .. S_K$
2. For each set retained add each possible length l subseq not already present
3. Compute relative entropy of each
   Retain d best.
4. Repeat until each set has K strings.

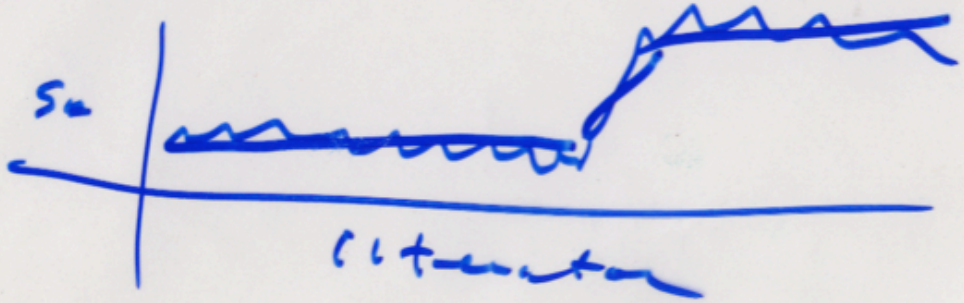
NB: Usual Greedy problems

Expectation Maximization

MEME [Bailey & Elkan]

Input: again sequences $S_1 .. S_k$ & motif length l
& background model
Again assume 1 instance per sequence
(variants possible)

Observed data: the sequences
Parameters Θ: the WMM
Hidden Data: where's motif? $Y_{ij} = \begin{cases} 1 & \text{if start at pos } j \text{ of seq } i \\ 0 & \text{otherwise} \end{cases}$



Expectation Step

$$\hat{Y}_{ik} = E(Y_{ik} \mid S_i, \Theta)$$

$E = 1 \cdot P(1) + 0 \cdot P(0)$

$$= P(Y_{ik}=1 \mid S_i, \Theta)$$

Bayes again

$$= P(S_i \mid Y_{ik}=1, \Theta) \frac{P(Y_{ik}=1 \mid \Theta)}{P(S_i \mid \Theta)}$$

$$= c\, P(S_i \mid Y_{ik}=1, \Theta)$$

$$= c' \prod_{j=1}^{L} P(S_{i,k+j-1} \mid \Theta)$$

Fix c' so $\sum_k \hat{Y}_{ik} = 1$

Maximization Step

given parameter Θ^t @ t^{th} iteration

Find Θ maximizing Expected value

$$Q(\Theta \mid \Theta^t) = E_{Y \sim \Theta^t}\left[\log P(S, Y \mid \Theta)\right]$$

$$= E\left[\log \prod_{i=1}^{k} P(S_i, Y_i \mid \Theta)\right]$$

$$\vdots$$

$$= \sum_i \sum_j E(Y_{ij}) \log P(S_i \mid \Theta, Y_{ij} = 1)$$

Θ maximized by "counting" frequencies in alignment, where counts are $\hat{Y}_{ij}$.

Initialization

1. Buy a supercomputer; call it SDSC
2. Try every motif-length substring as initial & use WMM with, say, 80% of mass on that sequence, rest uniform
3. Run a couple of iterations of each;
4. Run best few to convergence