Notes for CSE 527 11/3/03 by Jon Bleyhl (jbleyhl@amath)

Last Time:
EM Algorithm: Based on Maximum Likelihood Estimation

$L(x_1 \ldots x_n \mid \theta)$ function to be maximized

Mixture => complicated likelihood (hidden parameters):
- Not easy to solve analytically

Likelihood Surface (see example in slides):
- Potentially many bumps observed in distribution surface due to relabeling and more clusters than have been allowed for
- Local maxima are a problem for many optimization techniques
- Relabeling in particular is problematic for some, including Markov Chain Monte Carlo

EM: Estimates z's (hidden parameters) then estimate the hidden parameters repeat (iteration)

How Convergence Works:

Goal: Maximum likelihood estimate of θ i.e. find θ maximizing Pr(x|θ) (or log(Pr(x|θ)).

$P(X|Y) = P(X,Y)/P(X) \qquad P(X) = P(X,Y)/P(Y|X)$

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

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

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

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

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

2) $$Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t) + \sum_y P(y \mid X\theta^t) \cdot \log \frac{P(y \mid X,\theta^t)}{P(y \mid X,\theta)}$$

$$H(P(y \mid X,\theta^t) \,\|\, Q(y \mid X,\theta)) = \sum_y P(y \mid X,\theta^t) \cdot \log \frac{P(y \mid X,\theta^t)}{P(y \mid X,\theta)} \geq 0$$ Relative entropy

1) ≥ 0 if 2) ≥ 0

Find θ that maximizes $Q(\theta \mid \theta^t)$ by maximizing $Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t)$

Relative Entropy:

-Given *P, Q*

$$H(P \,||\, Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Let $P(x)\log\{P(x)/Q(x)\} = 0$ if $P(x) = 0$ (since lim y log y = 0)

And undef. If $0 = Q(x) < P(x)$

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

Upper bound: $\log x \leq x - 1$

Lower Bound: $\log x \geq 1 - 1/x$

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

Furthermore if $H(P||Q) = 0$ iff $P=Q$.

SEQUENCE ANALYSIS

3 billion basepairs, 98% is not traditional "genes"

-"TATA" Box analysis shows consensus sequence TATAAT ~ 10 bp up stream from transcriptional start site
- NOT exact: of 168 studied
  - Nearly all had 2/3 TaxyzT
  - 80-90% had all 3
  - 50% agreed in each of x,y,z
  - no perfect match

-Other common feature at –35 …

Turn frequencies into probabilistic model:

Assign probability to string based on frequencies

Weight matrices, I: Statistics

- Assume:
  - $f_{b,i}$ = frequency of base in position
  - f_b = frequency of base $\square\, i$ in all sequences

-Log likelihood ratio, given $S = B_1, B_2, \ldots B_6$

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

Weight Matrices, II: Chemistry

- Experiments Show ~ 80% correlation of log likelihood weight matrix scores to measure binding energy of RNA polymerase to variation on TATAAT consensus.

Why are Promoters so Fuzzy?

- Restriction enzymes, e.g., recognize very precise sequences?
- Why not polymerase/promoter?
- Likely: Variation gives ~1000 fold variation in expression level
- Possible consequence: rarely expressed genes will be hard to find