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

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

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

EM Convergence
Choose $\theta^{t+1} = \arg \max_{\theta} Q(\theta | \theta^t)$

### 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

<table>
<thead>
<tr>
<th>pos base</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>95</td>
<td>26</td>
<td>59</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
<td>2</td>
<td>14</td>
<td>13</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>1</td>
<td>16</td>
<td>15</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>79</td>
<td>3</td>
<td>44</td>
<td>13</td>
<td>17</td>
<td>96</td>
</tr>
</tbody>
</table>
Scanning for TATA

Weight Matrices: Statistics

- Assume:
  \[ f_{b,i} = \text{frequency of base } b \text{ in position } i \]
  \[ f_b = \text{frequency of base } b \text{ in all sequences} \]
- Log likelihood ratio, given \( S = B_1B_2...B_6 \):
  \[
  \log \frac{P(S|\text{"promoter"})}{P(S|\text{"nonpromoter"})} = \log \left( \frac{\prod_{i=1}^{6} f_{B_i,i}}{\prod_{i=1}^{6} f_{B_i}} \right) = \sum_{i=1}^{6} \log \left( \frac{f_{B_i,i}}{f_{B_i}} \right)
  \]

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]