Generative horizontal sequence models

Colin Dewey
2007.02.20
Previously: vertical sequence models

- Model state of individual sequence positions through time
- Observed: state of positions in extant sequences
- Hidden: state of positions in ancestral sequences
- Model is probability distribution over all possible states of sequence positions
Horizontal models

- Models of sequence through **space**
- Different positions in same sequence
- Observed: complete sequences
- Hidden: Class of sequence (e.g., exon, intron, intergenic, etc.)
Generative vs. discriminative classification

• Discriminative:
  • only model $P(Y|X)$
  • probability of different classes given data

• Generative:
  • model $P(X|Y)$
  • probability of data given class
  • can model complete joint distribution $P(X,Y)$ with $P(Y)$
  • Obtain $P(Y|X)$ from Bayes rule for classification
Generative models we’ve seen

• Motifs
• Profile matrices
• Genes
• k-th order Markov models
Profile matrices

- 0th-order inhomogenous Markov model conditioned on class (motif or background)

\[
P(x|y) = \prod_{i=1}^{5} p_i(x_i|y)
\]

Also called weight matrix method (WMM)
Homogenous Markov Chain

- Homogenous: conditional probability distributions independent of sequence position

\[ P[x|y] = p_1(x_1|y) \prod_{i=2}^{5} p_i(x_i|x_{i-1},y) \]

(1st order model)
Inhomogenous Markov models

- Inhomogenous: conditional probability distributions depend on position

\[
P[x|y] = p_1(x_1|y) \prod_{i=2}^{5} p_i(x_i|x_{i-1}, y)
\]

Weight Array Model (WAM) = 1st order inhomogenous

\[
P[x|y] = p_1(x_1|y) \prod_{i=2}^{5} p_i(x_i|x_{i-1}, y)
\]

**Example Table:**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>T</td>
<td>0.1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Example Table (Motif):**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>G</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>T</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**Example Table (Background):**
Coding DNA model

- 5th-order 3 periodic inhomogenous

\[ \mathbb{P}[x|y] = p_1(x_1|y)p_2(x_2|x_1,y)p_3(x_3|x_2,x_1,y)p_1(x_4|x_3,x_2,x_1,y)p_2(x_5|x_4,x_3,x_2,x_1,y)p_3(x_6|x_5,x_4,x_3,x_2,x_1,y) \]

Y: phase 0, phase 1, phase 2, or background
More arbitrary dependencies

- Dependencies need not be linearly Markovian in space

\[
P[x|y] = p_2(x_2|y)p_3(x_3|x_2, y)p_5(x_3|y)p_4(x_4|x_5, x_3, y)p_1(x_1|x_4, y)
\]
Genscan donor site model

<table>
<thead>
<tr>
<th>All sites:</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
<th>+4</th>
<th>+5</th>
<th>+6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>A%</td>
<td>C%</td>
<td>G%</td>
<td>U%</td>
<td>A%</td>
<td>C%</td>
<td>G%</td>
<td>U%</td>
<td>A%</td>
</tr>
<tr>
<td>A%</td>
<td>33</td>
<td>60</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>49</td>
<td>71</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>C%</td>
<td>37</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>G%</td>
<td>18</td>
<td>14</td>
<td>81</td>
<td>100</td>
<td>0</td>
<td>45</td>
<td>12</td>
<td>84</td>
<td>20</td>
</tr>
<tr>
<td>U%</td>
<td>12</td>
<td>13</td>
<td>7</td>
<td>0</td>
<td>100</td>
<td>3</td>
<td>9</td>
<td>5</td>
<td>46</td>
</tr>
</tbody>
</table>

-3 33 36 19 13
-2 56 15 15 15
-1 9 4 78 9
+3 44 3 51 3
+4 75 4 13 9
+6 14 18 19 49
-3 34 37 18 11
-2 59 10 15 16
+3 40 4 53 3
+4 70 4 16 10
+6 17 21 21 42
-3 37 42 18 3
+3 39 5 51 5
+4 62 5 22 11
+6 19 20 25 36
-3 32 40 23 5
+3 27 4 59 10
+4 51 5 25 19

All donor splice sites
(1254)

G5
(1057)

H5
(197)

G5G-1
(823)

G5H-1
(234)

G5G-1A-2
(487)

G5G-1B-2
(336)

G5G-1A-2U6
(177)

G5G-1A-2V6
(310)
Hidden Markov models

- Simultaneous classification of multiple positions

\[
P[x | y] = p(x_1 | y_1) \prod_{i=2}^5 p(y_i | y_{i-1})p(x_i | y_i)
\]
Forward, Backward, Posteriors

- **Forward values** \( f_l(i) = \mathbb{P}[x_1, \ldots, x_i, y_i = l] \)
  \[
  f_l(i) = e_l(x_i) \sum_k f_k(i - 1) a_{kl}
  \]

- **Backward values** \( b_l(i) = \mathbb{P}[x_{i+1}, \ldots, x_L | y_i = l] \)
  \[
  b_l(i) = \sum_k a_{lk} e_k(x_{i+1}) b_k(i + 1)
  \]

- **Posterior probabilities**
  \[
  \mathbb{P}[y_i = l | x] = \frac{f_l(i) b_l(i)}{\mathbb{P}[x]}
  \]
HMM decoding

- Posterior decoding
  - Maximize $P[y_i = l | x]$ separately for each position $i$
  - May produce invalid or unlikely hidden sequences

- Viterbi decoding
  - Choose $y$ such that $P[y | x]$ is maximized
Discriminative classifiers

- Advantage: don’t need to model sequence
  - Can lead to better performance
- Disadvantage: no model of sequence
  - Can’t generate sequences from model
  - Parameters don’t have direct biological meaning
- Examples: support vector machines (SVM), maximum entropy models, many others
Generalized HMMs

• Use smaller models we’ve discussed as submodels

• States in generalized HMMs emit sequences (generated by models like those discussed), instead of single symbols

• Need to model length of emitted sequences