TIME-DEPENDENT HMMS
IN BIOINFORMATIC APPLICATIONS
A General Definition of HMM

\[ \text{HMM} = (S, V, B, A, \Pi) \]

**N states**
\[ S = \{s_1, ..., s_N\} \]

**M symbols**
\[ V = \{v_1, ..., v_M\} \]

- **Initial state probability**:
  \[ \Pi = \{\pi_1, ..., \pi_N\} \quad \sum_{i=1}^{N} \pi_i = 1 \]
  \[ \pi_i : \text{prob of starting at state } s_i \]

- **State transition probability**:
  \[ A = \{a_{ij}\} \quad 1 \leq i, j \leq N \quad \sum_{j=1}^{N} a_{ij} = 1 \]
  \[ a_{ij} : \text{prob of going } s_i \rightarrow s_j \]

- **Output probability**:
  \[ B = \{b_i(v_k)\} \quad 1 \leq i \leq N, 1 \leq k \leq M \quad \sum_{k=1}^{M} b_i(v_k) = 1 \]
  \[ b_i(v_k) : \text{prob of "generating" } v_k \text{ at } s_i \]
THREE QUESTIONS OF INTEREST

• What is the probability of a particular sequence?
• What is the most probable sequence of a given length?
• What are the best model parameters given a set of examples?
What’s the most likely path?

\[ P(x|B) \]

\[ \begin{align*}
P(a|B) &= 0.25 \\
P(t|B) &= 0.40 \\
P(c|B) &= 0.10 \\
P(g|B) &= 0.25 \\
P(B) &= 0.5 \\
P(I) &= 0.5 \\
P(x|I) \end{align*} \]

\[ S_1^* S_2^* ... S_T^* = \arg \max_{S_1 S_2 ... S_T} p(S_1 S_2 ... S_T, O) = \arg \max_{S_1 S_2 ... S_T} \pi(S_1) b_{S_1}(v_{o_1}) \prod_{t=2}^{T} a_{S_{t-1}S_t} b_{S_t}(v_{o_t}) \]
Viterbi Algorithm: An Example

P(x|B)
P(a|B)=0.251
P(t|B)=0.40
P(c|B)=0.098
P(g|B)=0.251

P(x|I)
P(a|I)=0.25
P(t|I)=0.25
P(c|I)=0.25
P(g|I)=0.25

P(B)=0.5
P(I)=0.5

Remember the best paths so far
FINDING CPG ISLANDS ON CHR 21

- Naïve initial parameter setting:
- Emission Probabilities:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>.251</td>
<td>.098</td>
<td>.251</td>
<td>.40</td>
</tr>
<tr>
<td>I</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
</tr>
</tbody>
</table>

- Transition Probabilities:

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>.7</td>
<td>.3</td>
</tr>
<tr>
<td>I</td>
<td>.5</td>
<td>.5</td>
</tr>
</tbody>
</table>
FINDING CPG ISLANDS ON CHR 21

- Naïve initial parameter setting:
- Emission Probabilities:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>.251</td>
<td>.098</td>
<td>.251</td>
<td>.40</td>
</tr>
<tr>
<td>I</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
</tr>
</tbody>
</table>

- Transition Probabilities:

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>.7</td>
<td>.3</td>
</tr>
<tr>
<td>I</td>
<td>.5</td>
<td>.5</td>
</tr>
</tbody>
</table>
THREE QUESTIONS OF INTEREST

• What is the probability of a particular sequence?
• What is the most probable sequence of a given length?
• What are the best model parameters given a set of examples?
Supervised Training

Given:
1. N – the number of states, e.g., 2, (s1 and s2)
2. V – the vocabulary, e.g., V={a,b}
3. O – observations, e.g., O=aaaaabbbbb
4. State transitions, e.g., S=1121122222

Task: Estimate the following parameters

1. $\pi_1, \pi_2$
2. $a_{11}, a_{12}, a_{22}, a_{21}$
3. $b_1(a), b_1(b), b_2(a), b_2(b)$
FINDING CPG ISLANDS ON CHR 21

- Naïve initial parameter setting:
  - Emission Probabilities:
    | A | C | G | T |
    |---|---|---|---|
    | B | .251 | .098 | .251 | .40 |
    | I | .25 | .25 | .25 | .25 |
  - Transition Probabilities:
    | B | I |
    |---|---|
    | B | .7 | .3 |
    | I | .5 | .5 |

- Parameters tuned w/ supervised training:
  - Emission Probabilities:
    | A | C | G | T |
    |---|---|---|---|
    | B | .282 | .203 | .217 | .298 |
    | I | .155 | .340 | .338 | .167 |
  - Transition Probabilities:
    | B | I |
    |---|---|
    | B | .99 | .1 |
    | I | .1 | .99 |
BAUM-WELCH TRAINING

- Expectation-Maximization
- Unsupervised training
- Requires good estimations of input parameters
- Calculate alpha and beta vectors to find parameter updates

\[
\pi'_i = \alpha_1(i) \cdot \beta_1(i) / \sum_i \alpha_1(i) \cdot \beta_1(i)
\]

\[
a'_{ij} = \frac{\sum_t \alpha_t(i) \cdot a_{ij} \cdot b_j(O_{t+1}) \cdot \beta_{t+1}(j)}{\sum_t \alpha_t(i) \cdot \beta_t(i)}
\]

\[
b'_{jk} = \frac{\sum_t \alpha_t(j) \cdot b_j(O_t) \cdot \beta_t(j)}{\sum_t \alpha_t(j) \cdot \beta_t(j)}
\]

- Online vs batch updates
FINDING CPG ISLANDS ON CHR 21

- Supervised training parameters:
  - Emission Probabilities:
    - Transition Probabilities:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>.282</td>
<td>.203</td>
<td>.217</td>
<td>.298</td>
</tr>
<tr>
<td>I</td>
<td>.155</td>
<td>.340</td>
<td>.338</td>
<td>.167</td>
</tr>
</tbody>
</table>

- Parameters tuned w/ Baum-Welch:
  - Emission Probabilities:
    - Transition Probabilities:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>.304</td>
<td>.219</td>
<td>.209</td>
<td>.268</td>
</tr>
<tr>
<td>I</td>
<td>.194</td>
<td>.319</td>
<td>.320</td>
<td>.167</td>
</tr>
</tbody>
</table>

- Transition Probabilities:
**REVERSE TRANSLATION**

- Given an amino acid sequence, can we infer the codon sequence from which it derived?

MKKMQSIVLALSLLVAPMAAQAAEITLVPSVKLQIGDRDNRGYYWDGGHWRDHGWWKQ...
TIME-DEPENDENT HMM

- Probabilities vary depending on the position in the sequence
- One transition probability matrix for the first $n$ codons
- A different probability matrix for the rest
CONCLUSION

• HMMs are sensitive to the parameters you use
• Baum-Welch training works well, especially if you have good initial parameters
• Setbacks: obtaining data & data formatting
• Other “might have been” projects
  • Music chord generation
  • Protein secondary structure
  • Speech recognition
CONCLUSION

• HMMs are sensitive to the parameters you use
• Baum-Welch training works well, especially if you have good initial parameters
• Setbacks: obtaining data & data formatting
• Other “might have been” projects
  • Music chord generation
  • Protein secondary structure
  • Speech recognition

• QuEsTiOnS?