BCB BCB/GDCB/STAT/COM S 568 Spring 2011

Homework 4 Solution

March 4, 2011

Assume we had some knowledge that transcribed parts of a genome have different nucleotide composition in terms of purines and pyrimidines relative to intergenic regions. In this case, we may want to model the genome as a Markov chain of states E (”expressed”) and I (”intergenic”), each state with prescribed output bias toward purines (”R”) or pyrimidines (”Y”). Thus, the genome is modeled as a 2-state Hidden Markov model with parameters $\lambda = \{\pi, \tau, P\}$, where (assuming specific parameter values for sake of calculation)

π are the initial state probabilities: $\pi = \{\pi_E = 3/4, \pi_I = 1/4\}$,
τ are the state transition probabilities: $\tau = \{\tau_{EE} = 1/4, \tau_{EI} = 3/4, \tau_{IE} = 1/2, \tau_{II} = 1/2\}$, and
P are the output probabilities: $P = \{P(R|E) = 4/5, P(Y|E) = 1/5, P(R|I) = 2/5, P(Y|I) = 3/5\}$.

(1) Calculate $P_\lambda(RYYR)$, using first the ”forward algorithm” of HMM theory and secondly, independently, the ”backward algorithm” of HMM theory.

Solution:
Forward algorithm.
Define:

$$\begin{aligned} E_i &= P(O_1, \ldots, O_i, Q_i = E) \\ I_i &= P(O_1, \ldots, O_i, Q_i = I) \end{aligned}$$

For sequence $RYYR$:

$$\begin{aligned} E_1 &= \pi_E P(R|E) = \frac{3}{4} \cdot \frac{4}{5} = \frac{3}{5} \\ I_1 &= \pi_I P(R|I) = \frac{1}{4} \cdot \frac{2}{5} = \frac{1}{10} \\ E_2 &= (E_1 \tau_{EE} + I_1 \tau_{IE}) P(Y|E) = (\frac{3}{5} \cdot \frac{1}{4} + \frac{1}{10} \cdot \frac{1}{2}) \cdot \frac{1}{5} = \frac{1}{25} \\ I_2 &= (E_1 \tau_{EI} + I_1 \tau_{II}) P(Y|I) = (\frac{3}{5} \cdot \frac{3}{4} + \frac{1}{10} \cdot \frac{1}{2}) \cdot \frac{3}{5} = \frac{3}{10} \\ E_3 &= (E_2 \tau_{EE} + I_2 \tau_{IE}) P(Y|E) = (\frac{1}{25} \cdot \frac{1}{4} + \frac{3}{10} \cdot \frac{1}{2}) \cdot \frac{1}{5} = \frac{4}{125} \\ I_3 &= (E_2 \tau_{EI} + I_2 \tau_{II}) P(Y|I) = (\frac{1}{25} \cdot \frac{3}{4} + \frac{3}{10} \cdot \frac{1}{2}) \cdot \frac{3}{5} = \frac{27}{250} \\ E_4 &= (E_3 \tau_{EE} + I_3 \tau_{IE}) P(R|E) = (\frac{4}{125} \cdot \frac{1}{4} + \frac{27}{250} \cdot \frac{1}{2}) \cdot \frac{4}{5} = \frac{31}{625} \\ I_4 &= (E_3 \tau_{EI} + I_3 \tau_{II}) P(R|I) = (\frac{4}{125} \cdot \frac{3}{4} + \frac{27}{250} \cdot \frac{1}{2}) \cdot \frac{2}{5} = \frac{39}{1250} \\ \Rightarrow P_\lambda(RYYR) &= E_4 + I_4 = \frac{31}{625} + \frac{39}{1250} = \frac{101}{1250} \end{aligned}$$

Backward algorithm:
Define:

$$\begin{aligned} \overline{E}_i &= P(O_{i+1}, \ldots, O_N | Q_i = E) \\ \overline{I}_i &= P(O_{i+1}, \ldots, O_N | Q_i = I) \end{aligned}$$

For sequence $RYYR$:

$$\begin{aligned}
\overline{E}_4 &= 1 \\
\overline{I}_4 &= 1 \\
\overline{E}_3 &= \tau_{EE}P(R|E) + \tau_{EI}P(R|I) = \frac{1}{4}\cdot\frac{4}{5}+\frac{3}{4}\cdot\frac{2}{5} = \frac{1}{2} \\
\overline{I}_3 &= \tau_{IE}P(R|E) + \tau_{II}P(R|I) = \frac{1}{2}\cdot\frac{4}{5}+\frac{1}{2}\cdot\frac{2}{5} = \frac{3}{5} \\
\overline{E}_2 &= \tau_{EE}P(Y|E)\overline{E}_3 + \tau_{EI}P(Y|I)\overline{I}_3 = \frac{1}{4}\cdot\frac{1}{5}\cdot\frac{1}{2}+\frac{3}{4}\cdot\frac{3}{5}\cdot\frac{3}{5} = \frac{59}{200} \\
\overline{I}_2 &= \tau_{IE}P(Y|E)\overline{E}_3 + \tau_{II}P(Y|I)\overline{I}_3 = \frac{1}{2}\cdot\frac{1}{5}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{3}{5}\cdot\frac{3}{5} = \frac{23}{100} \\
\overline{E}_1 &= \tau_{EE}P(Y|E)\overline{E}_2 + \tau_{EI}P(Y|I)\overline{I}_2 = \frac{1}{4}\cdot\frac{1}{5}\cdot\frac{59}{200}+\frac{3}{4}\cdot\frac{3}{5}\cdot\frac{23}{100} = \frac{473}{4000} \\
\overline{I}_1 &= \tau_{IE}P(Y|E)\overline{E}_2 + \tau_{II}P(Y|I)\overline{I}_2 = \frac{1}{2}\cdot\frac{1}{5}\cdot\frac{59}{200}+\frac{1}{2}\cdot\frac{3}{5}\cdot\frac{23}{100} = \frac{197}{2000} \\
\Rightarrow P_\lambda(RYYR) &= \pi_E P(R|E)\overline{E}_1 + \pi_I P(R|I)\overline{I}_1 = \frac{3}{4}\cdot\frac{4}{5}\cdot\frac{473}{4000}+\frac{1}{4}\cdot\frac{2}{5}\cdot\frac{197}{2000} = \frac{101}{1250}
\end{aligned}$$

(2) Calculate $P_\lambda(Q_2 = E|RYY)$ using the relevant algorithms.

Solution:
Forward algorithm:

$$\begin{aligned}
E_1 &= \pi_E P(R|E) = \frac{3}{4}\cdot\frac{4}{5} = \frac{3}{5} \\
I_1 &= \pi_I P(R|I) = \frac{1}{4}\cdot\frac{2}{5} = \frac{1}{10} \\
E_2 &= (E_1\tau_{EE} + I_1\tau_{IE})P(Y|E) = (\frac{3}{5}\cdot\frac{1}{4}+\frac{1}{10}\cdot\frac{1}{2})\cdot\frac{1}{5} = \frac{1}{25} \\
I_2 &= (E_1\tau_{EI} + I_1\tau_{II})P(Y|I) = (\frac{3}{5}\cdot\frac{3}{4}+\frac{1}{10}\cdot\frac{1}{2})\cdot\frac{3}{5} = \frac{3}{10} \\
E_3 &= (E_2\tau_{EE} + I_2\tau_{IE})P(Y|E) = (\frac{1}{25}\cdot\frac{1}{4}+\frac{3}{10}\cdot\frac{1}{2})\cdot\frac{1}{5} = \frac{4}{125} \\
I_3 &= (E_2\tau_{EI} + I_2\tau_{II})P(Y|I) = (\frac{1}{25}\cdot\frac{3}{4}+\frac{3}{10}\cdot\frac{1}{2})\cdot\frac{3}{5} = \frac{27}{250}
\end{aligned}$$

Backward algorithm:

$$\begin{aligned}
\overline{E}_2 &= \tau_{EE}P(Y|E) + \tau_{EI}P(Y|I) = \frac{1}{4}\cdot\frac{1}{5}+\frac{3}{4}\cdot\frac{3}{5} = \frac{1}{2} \\
\overline{I}_2 &= \tau_{IE}P(Y|E) + \tau_{II}P(Y|I) = \frac{1}{2}\cdot\frac{1}{5}+\frac{1}{2}\cdot\frac{3}{5} = \frac{2}{5}
\end{aligned}$$

$$\begin{aligned}
P_\lambda(Q_2 = E|RYY) &= \frac{P(O_1 = R, O_2 = Y, Q_2 = E)P(O_3 = Y|Q_2 = E)}{P_\lambda(RYY)} \\
&= \frac{E_2\cdot\overline{E}_2}{E_3 + I_3} \\
&= \frac{\frac{1}{25}\cdot\frac{1}{2}}{\frac{4}{125}+\frac{27}{250}} \\
&= \frac{\frac{1}{50}}{\frac{7}{50}} \\
&= \frac{1}{7}
\end{aligned}$$

(3) Check your answer in (2) using the rules of total probability directly.

Solution:

P($O_1O_2O_3 \mid Q_1Q_2Q_3$)	EEE	EEI	EIE	EII	IEE	IEI	IIE	III
RRR	$\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}$	$\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{2}{5}$	$\frac{4}{5}\cdot\frac{2}{5}\cdot\frac{4}{5}$	$\frac{4}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}$	$\frac{2}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}$	$\frac{2}{5}\cdot\frac{4}{5}\cdot\frac{2}{5}$	$\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{4}{5}$	$\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}$
RRY	$\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{1}{5}$	$\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{3}{5}$	$\frac{4}{5}\cdot\frac{2}{5}\cdot\frac{1}{5}$	$\frac{4}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}$	$\frac{2}{5}\cdot\frac{4}{5}\cdot\frac{1}{5}$	$\frac{2}{5}\cdot\frac{4}{5}\cdot\frac{3}{5}$	$\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{1}{5}$	$\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}$
RYR	$\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}$	$\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{2}{5}$	$\frac{4}{5}\cdot\frac{3}{5}\cdot\frac{4}{5}$	$\frac{4}{5}\cdot\frac{3}{5}\cdot\frac{2}{5}$	$\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}$	$\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{2}{5}$	$\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{4}{5}$	$\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{2}{5}$
RYY	$\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}$	$\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}$	$\frac{4}{5}\cdot\frac{3}{5}\cdot\frac{1}{5}$	$\frac{4}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}$	$\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}$	$\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}$	$\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{1}{5}$	$\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}$
YRR	$\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}$	$\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{2}{5}$	$\frac{1}{5}\cdot\frac{2}{5}\cdot\frac{4}{5}$	$\frac{1}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}$	$\frac{3}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}$	$\frac{3}{5}\cdot\frac{4}{5}\cdot\frac{2}{5}$	$\frac{3}{5}\cdot\frac{2}{5}\cdot\frac{4}{5}$	$\frac{3}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}$
YRY	$\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{1}{5}$	$\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{3}{5}$	$\frac{1}{5}\cdot\frac{2}{5}\cdot\frac{1}{5}$	$\frac{1}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}$	$\frac{3}{5}\cdot\frac{4}{5}\cdot\frac{1}{5}$	$\frac{3}{5}\cdot\frac{4}{5}\cdot\frac{3}{5}$	$\frac{3}{5}\cdot\frac{2}{5}\cdot\frac{1}{5}$	$\frac{3}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}$
YYR	$\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}$	$\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{2}{5}$	$\frac{1}{5}\cdot\frac{3}{5}\cdot\frac{4}{5}$	$\frac{1}{5}\cdot\frac{3}{5}\cdot\frac{2}{5}$	$\frac{3}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}$	$\frac{3}{5}\cdot\frac{1}{5}\cdot\frac{2}{5}$	$\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{4}{5}$	$\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{2}{5}$
YYY	$\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}$	$\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}$	$\frac{1}{5}\cdot\frac{3}{5}\cdot\frac{1}{5}$	$\frac{1}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}$	$\frac{3}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}$	$\frac{3}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}$	$\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{1}{5}$	$\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}$
P($Q_1Q_2Q_3$)	$\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{1}{4} = \mathbf{\frac{3}{64}}$	$\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4} = \mathbf{\frac{9}{64}}$	$\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{2} = \mathbf{\frac{9}{32}}$	$\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{2} = \mathbf{\frac{9}{32}}$	$\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{1}{4} = \mathbf{\frac{1}{32}}$	$\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{3}{4} = \mathbf{\frac{3}{32}}$	$\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{1}{2} = \mathbf{\frac{1}{16}}$	$\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{1}{2} = \mathbf{\frac{1}{16}}$

$$
\begin{aligned}
P_\lambda(Q_2 = E \mid RYY) &= \frac{P_\lambda(Q_2 = E, RYY)}{P_\lambda(RYY)} \\
P_\lambda(RYY) = (\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}) &\times \frac{3}{64} + \\
(\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}) &\times \frac{9}{64} + \\
(\frac{4}{5}\cdot\frac{3}{5}\cdot\frac{1}{5}) &\times \frac{9}{32} + \\
(\frac{4}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}) &\times \frac{9}{32} + \\
(\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}) &\times \frac{1}{32} + \\
(\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}) &\times \frac{3}{32} + \\
(\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{1}{5}) &\times \frac{1}{16} + \\
(\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}) &\times \frac{1}{16} \\
= \frac{7}{50} & \\
P_\lambda(Q_2 = E, RYY) = (\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}) &\times \frac{3}{64} + \\
(\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}) &\times \frac{9}{64} + \\
(\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}) &\times \frac{1}{32} + \\
(\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{3}{5}) &\times \frac{3}{32} \\
= \frac{1}{50} & \\
P_\lambda(Q_2 = E \mid RYY) &= \frac{\frac{1}{50}}{\frac{7}{50}} \\
&= \frac{1}{7}
\end{aligned}
$$

(4) Calculate $P_\lambda(Q_2 = E, Q_3 = I \mid RYYR)$, using the relevant algorithms.

Solution:

$$
\begin{aligned}
P_\lambda(Q_2 = E, Q_3 = I \mid RYYR) &= \frac{E_2 \times \tau_{EI} \times P(Y \mid I) \times \overline{I}_3}{E_4 + I_4} \\
&= \frac{\frac{1}{25} \times \frac{3}{4} \times \frac{3}{5} \times \frac{3}{5}}{\frac{101}{1250}} \\
&= \frac{27}{202}
\end{aligned}
$$

(5) Calculate the most probable state sequence given the observation $RYYR$ using the Viterbi algorithm.

Solution:
Define:

$$\begin{aligned}
e_i &= \max_{Q_i}(O_1, \ldots, O_i, Q_1, \ldots, Q_i = E) \\
m_i &= \max_{Q_i}(O_1, \ldots, O_i, Q_1, \ldots, Q_i = I) \\
\mathrm{ptr}_i(Q) &= \text{the traceback function from position } i \text{ of state } Q
\end{aligned}$$

For sequence $RYYR$:

$$\begin{aligned}
e_1 &= \pi_E P(R|E) = \frac{3}{4} \times \frac{4}{5} = \frac{3}{5} \\
m_1 &= \pi_I P(R|I) = \frac{1}{4} \times \frac{2}{5} = \frac{1}{10} \\
e_2 &= \max(e_1\tau_{EE}, m_1\tau_{IE})P(Y|E) = \max(\frac{3}{20}, \frac{1}{20}) \times \frac{1}{5} = \frac{3}{100},\ \mathrm{ptr}_2(E) = E \\
m_2 &= \max(e_1\tau_{EI}, m_1\tau_{II})P(Y|I) = \max(\frac{9}{20}, \frac{1}{20}) \times \frac{3}{5} = \frac{27}{100},\ \mathrm{ptr}_2(I) = E \\
e_3 &= \max(e_2\tau_{EE}, m_2\tau_{IE})P(Y|E) = \max(\frac{3}{400}, \frac{27}{200}) \times \frac{1}{5} = \frac{27}{1000},\ \mathrm{ptr}_3(E) = I \\
m_3 &= \max(e_2\tau_{EI}, m_2\tau_{II})P(Y|I) = \max(\frac{9}{400}, \frac{27}{200}) \times \frac{3}{5} = \frac{81}{1000},\ \mathrm{ptr}_3(I) = I \\
e_4 &= \max(e_3\tau_{EE}, m_3\tau_{IE})P(R|E) = \max(\frac{27}{4000}, \frac{81}{2000}) \times \frac{4}{5} = \frac{81}{2500},\ \mathrm{ptr}_4(E) = I \\
m_4 &= \max(e_3\tau_{EI}, m_3\tau_{II})P(R|I) = \max(\frac{81}{4000}, \frac{81}{2000}) \times \frac{2}{5} = \frac{81}{5000},\ \mathrm{ptr}_4(I) = I
\end{aligned}$$

The most probable state sequence S is determined as follows:

$S_4 = \mathrm{argmax}_Q(Q_4) = E$
$S_3 = \mathrm{ptr}_4(E) = I$
$S_2 = \mathrm{ptr}_3(I) = I$
$S_1 = \mathrm{ptr}_2(I) = E$

Thus, S is $EIIE$.