Markov Models

Anna Yeaton

Fall 2018

Lab Section

In this lab, we will go over Hidden Markov Models.

Hidden Markov Models: Gene Detection from DNA Sequence

Possible states: * 5' UTR * Exon * Intron * 3' UTR

State Transition Matrix, A, where A_ij is the probability of j given i

```
A <- data.frame("five_prime_UTR" = c(0.4, 0, 0.1, 0), "exon" = c(0.4, 0.4, 0.6,0), "intron" =c(0.2, 0.4
rownames(A) <- c("five_prime_UTR", "exon", "intron", "three_prime_UTR")

A <- data.matrix(A)
A
```

```
##                 five_prime_UTR exon intron three_prime_UTR
## five_prime_UTR             0.4  0.4    0.2             0.0
## exon                       0.0  0.4    0.4             0.2
## intron                     0.1  0.6    0.3             0.0
## three_prime_UTR            0.0  0.0    0.6             0.4
```

Emission matrix, B, where B_ij is the probability of of j given i

```
codons <- names(GENETIC_CODE)
five_prime_UTR <- rep(0.016, 64)
three_prime_UTR <- rep(0.0158, 64)
exon <- rep(0.0164, 64)
intron <- rep(0.0164, 64)
B <- rbind(five_prime_UTR, three_prime_UTR, exon, intron)
colnames(B) <- codons
#B <- B[-1,]

#manually change some probabilities using known information.
#5' UTR probably doesn't have any of the three stop codons
#An exon probably doesn't have any of the three stop codons
#An intron probably doesn't have the start codon or any of the stop codons
#3' UTR probably doesn't have the start codon


#stop codons
B[c(1,3,4),which(colnames(B) == "TAG")] <- c(0.001,0.001,0.001)
B[c(1,3,4),which(colnames(B) == "TGA")] <- c(0.001,0.001,0.001)
B[c(1,3,4),which(colnames(B) == "TAA")] <- c(0.001,0.001,0.001)

#start codon
B[c(2,4), which(colnames(B) == "ATG")] <- c(0.001,0.001)
```

```
B <- data.matrix(B)
B
```

```
##                    TTT    TTC    TTA    TTG    TCT    TCC    TCA    TCG
## five_prime_UTR  0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160
## three_prime_UTR 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158
## exon            0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
## intron          0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
##                    TAT    TAC    TAA    TAG    TGT    TGC    TGA    TGG
## five_prime_UTR  0.0160 0.0160 0.0010 0.0010 0.0160 0.0160 0.0010 0.0160
## three_prime_UTR 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158
## exon            0.0164 0.0164 0.0010 0.0010 0.0164 0.0164 0.0010 0.0164
## intron          0.0164 0.0164 0.0010 0.0010 0.0164 0.0164 0.0010 0.0164
##                    CTT    CTC    CTA    CTG    CCT    CCC    CCA    CCG
## five_prime_UTR  0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160
## three_prime_UTR 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158
## exon            0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
## intron          0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
##                    CAT    CAC    CAA    CAG    CGT    CGC    CGA    CGG
## five_prime_UTR  0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160
## three_prime_UTR 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158
## exon            0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
## intron          0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
##                    ATT    ATC    ATA    ATG    ACT    ACC    ACA    ACG
## five_prime_UTR  0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160
## three_prime_UTR 0.0158 0.0158 0.0158 0.0010 0.0158 0.0158 0.0158 0.0158
## exon            0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
## intron          0.0164 0.0164 0.0164 0.0010 0.0164 0.0164 0.0164 0.0164
##                    AAT    AAC    AAA    AAG    AGT    AGC    AGA    AGG
## five_prime_UTR  0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160
## three_prime_UTR 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158
## exon            0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
## intron          0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
##                    GTT    GTC    GTA    GTG    GCT    GCC    GCA    GCG
## five_prime_UTR  0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160
## three_prime_UTR 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158
## exon            0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
## intron          0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
##                    GAT    GAC    GAA    GAG    GGT    GGC    GGA    GGG
## five_prime_UTR  0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160 0.0160
## three_prime_UTR 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158 0.0158
## exon            0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
## intron          0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164 0.0164
```

We can ask 3 questions using HMM. What is the most likely sequence of states(exon, intron, etc) given the observations(codons)? (Viterbi Algorithm)

What is the probability of the observed sequence(codons)?(Forward Algorithm)

How can we learn the HMM's parameters A and B given some data? (Forward-Backward Algorithm)

We will focus on the first question now: What is the most likely sequence of states given the observed sequence.

We solve this by finding the maximum likelihood state for the observed sequence. The long way to do this

is for every codon position, we find the probability of each state occuring, and keeping the state with the maximum probability. But this can result in near infinite number of calculations, instead, we use the viterbi algorithm.

Given this DNA sequence, what is the the most probable sequence of states?

```
STAT <- c("ATG","AGA","GCT","CCA","GGG", "AGG", "GAC", "CTG", "GGT",
          "AGA","AGG","AGA","AGC","CGG","AAA","CAG","CGG","GCT","GGG",
          "GCA","GCC","ACT","GCT","TAC","ACT","GAA","GAG","GGA","GGA",
          "CGG","GAG","AGG","AGT","GTG","TGT","GTG","TGT","GTG","TGT",
          "GTG","TGT","GTG")
```

Initialize HMM

```
hmm <- initHMM(colnames(A), colnames(B), transProbs = A, emissionProbs = B)

viterbi(hmm, STAT)
```

```
##  [1] "intron" "exon"   "intron" "exon"   "intron" "exon"   "intron"
##  [8] "exon"   "intron" "exon"   "intron" "exon"   "intron" "exon"
## [15] "intron" "exon"   "intron" "exon"   "intron" "exon"   "intron"
## [22] "exon"   "intron" "exon"   "intron" "exon"   "intron" "exon"
## [29] "intron" "exon"   "intron" "exon"   "intron" "exon"   "intron"
## [36] "exon"   "intron" "exon"   "intron" "exon"   "intron" "exon"
```

Homework

1. Create a Hidden Markov model of any system. Explain the Transition matrix and emission matrix. Evaluate your HMM – how well did it predict the sequence of events in your ground truth case?