Pset3

1a.

```
Log(P)= −1303.642
Viterbi intervals:
(66,415)  (527, 720)  (951, 1000)
```

1b.

```
When tested with hmm.fasta
```

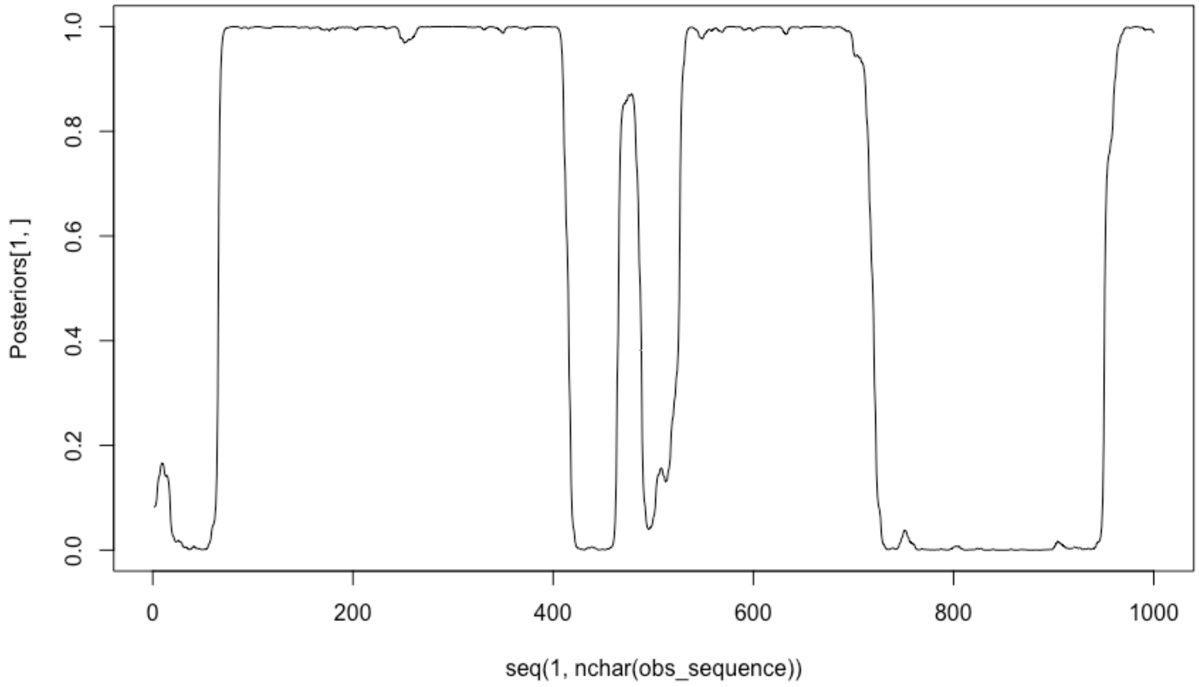


Figure 1 hmm sequence h state (GC rich) posterior likelihood

1c.

Mu = seq(0,1,by=0.05)

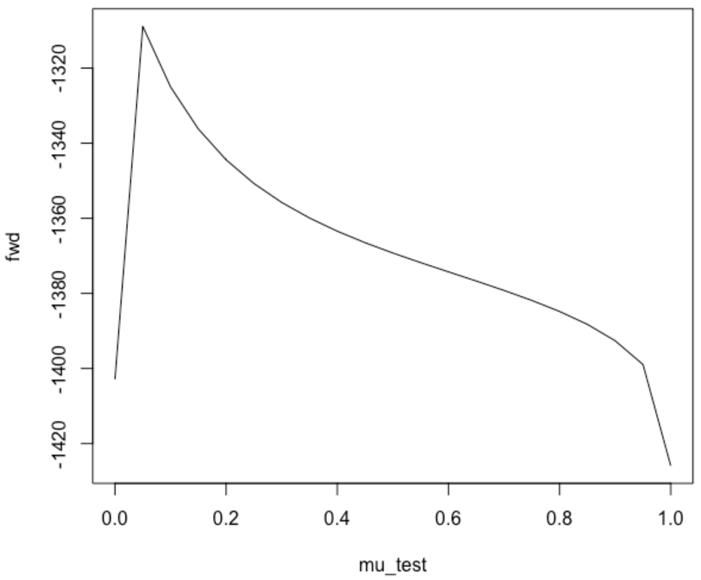
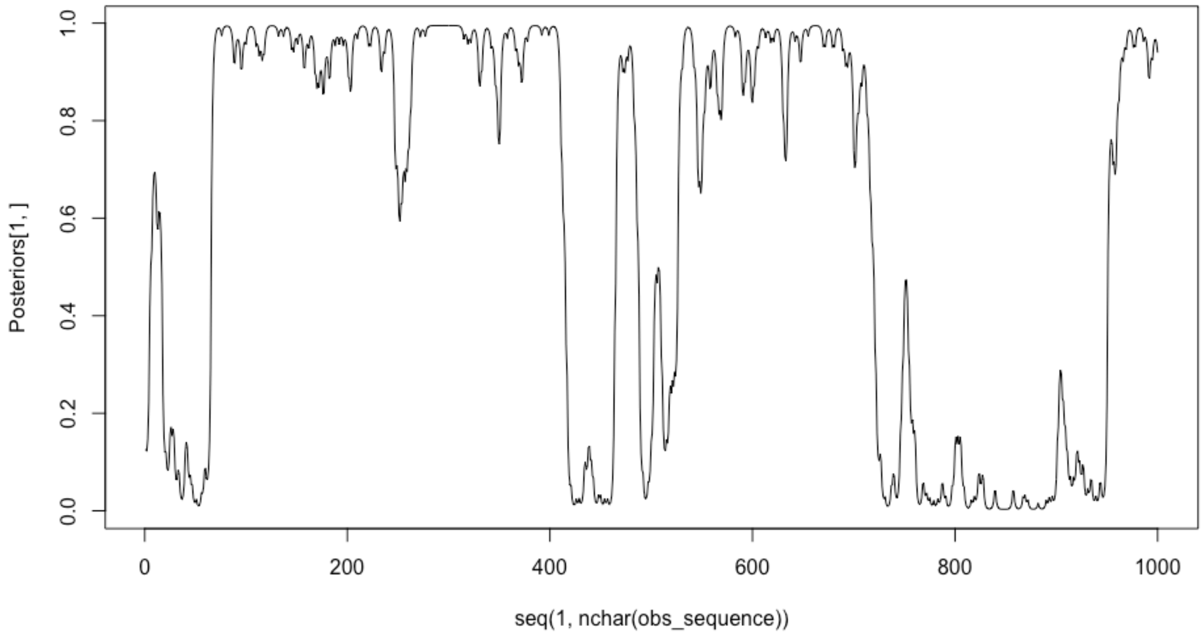


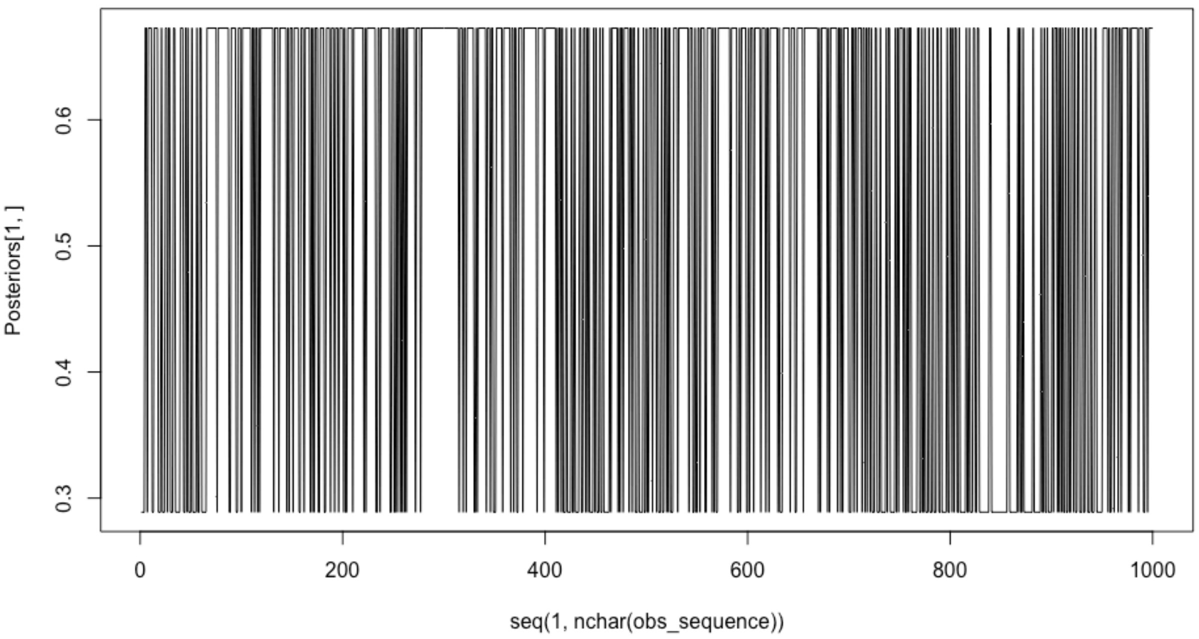
Figure 2 forward probability vs. mu

It peaked around mu = 0.1
Posterior probability in response to mu:

a. When mu = 0.05



b. When mu =0.5



The interval gets shorter and more transitions happen.

2a.

$$P(x,z|\mu,\theta_h,\theta_l) = \mu^{c_b} * (1-\mu)^{c_s} * {\theta_h}^{d_{hG}} * (1-\theta_h)^{d_{hA}} * {\theta_l}^{d_{lG}} * (1-\theta_l)^{d_{lA}}$$

$$\log\big(P(x,z|\mu,\theta_h,\theta_l)\big) = c_b * \log(\mu) + c_s * \log(1-\mu) + d_{hG} * \log(\theta_h) + d_{hA} * \log(1-\theta_h) + d_{lG} * \log(\theta_l) + d_{lA} * \log(1-\theta_l)$$

2b.

Based on MLE,

$$\frac{\partial logP}{\partial\mu} = \frac{cb}{\mu} - \frac{cs}{1-\mu} = 0, \qquad \hat{\mu} = \frac{cb}{cb+cs} = \frac{7}{999} = 0.007$$

$$\frac{\partial logP}{\partial\theta_h} = \frac{dhG}{\theta_h} - \frac{dhA}{1-\theta_h} = 0, \ \widehat{\theta_h} = \frac{dhG}{dhG+dhA} = \frac{495}{495+156} = 0.76$$

$$\widehat{\theta_l} = \frac{dlG}{dlG+dlA} = \frac{115}{115+234} = 0.3295$$

Given:

$c_b = 7, c_s = 999 - 7 = 992, d_{hG} = 495, d_{hA} = 156, d_{lG} = 115, d_{lA} = 234$

3a.

Log forward likelihood of the 50 simulated data. They are very different, with extremity exp(80) times difference.

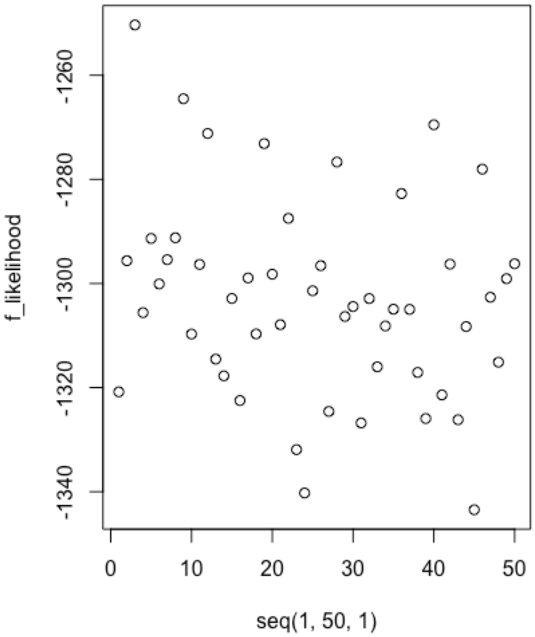


Figure 3 Log likelihood of 50 HMM emiss and states

Sample both x and z at length 1000 for 50 HMM.
Mean(Transition) = 9.93
Var(transition) = 8.69

In the binomial model:
Mu = 0.01
Mean(cb) = 999*mu = 9.99
Var= 999*mu*(1-mu) = 9.89

The mean is mostly the same. The difference between the expected value may be due to the limited sample size (n=50 too small)

3b.
The log likelihood of the sampled states:

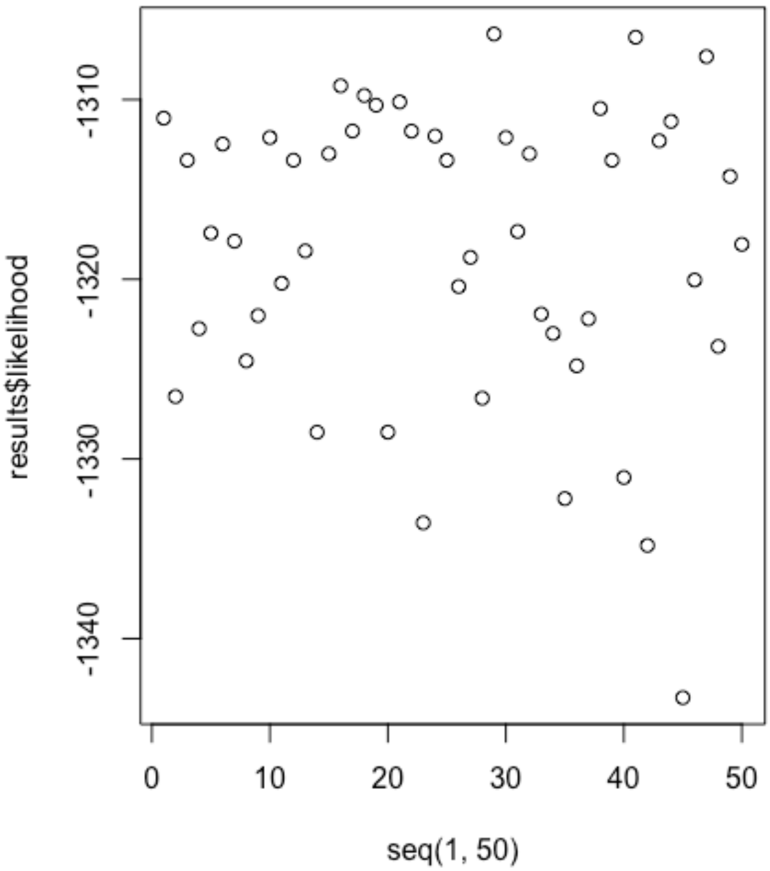


Figure 4 log likelihood of the 50 sampled states

Compare to the log likelihood of Viterbi, all of the sampled states show a lower loglikelihood. But some of the paths are close to Viterbi path.

How different?
Some of the sampled path are close to Viterbi path while others exhibit more GC rich intervals and show lower log likelihood.

Counts of ‘h’ state at every site for 50 simulations:

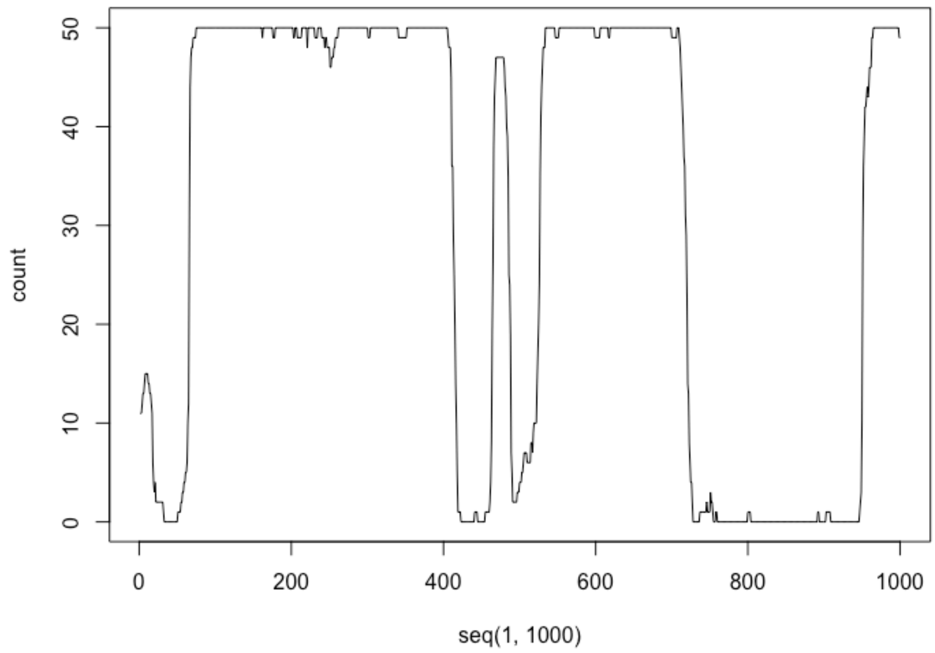


Figure 5 Count of GC rich labels over 1000sites from sampled paths

Over the 50 sampled state path, they show a similar pattern of the posterior likelihood.