Homework 3

Michael Li

1. (a) code submitted online

List of high G+C regions: (66, 415), (527, 720), (951, 1000)

(b) code submitted online

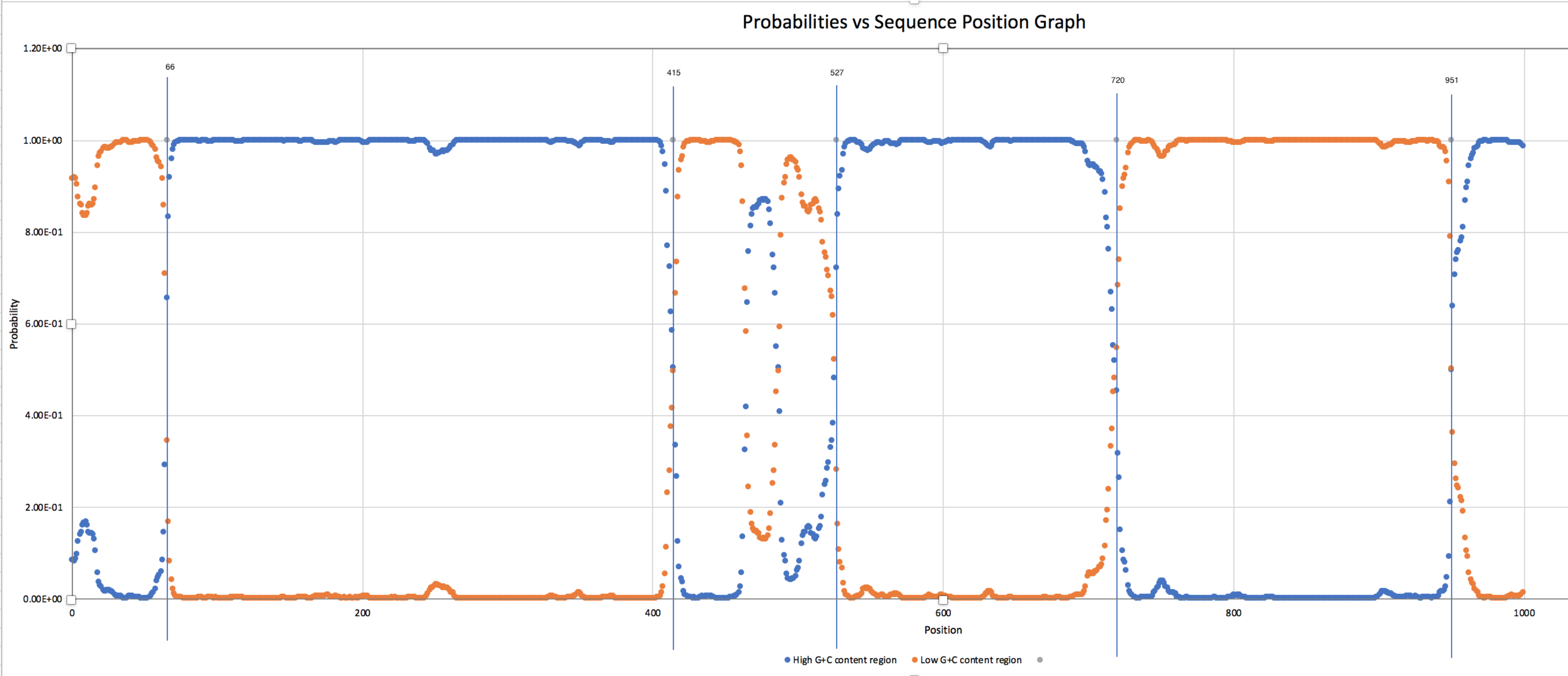
screenshot of graph on next page

(c)

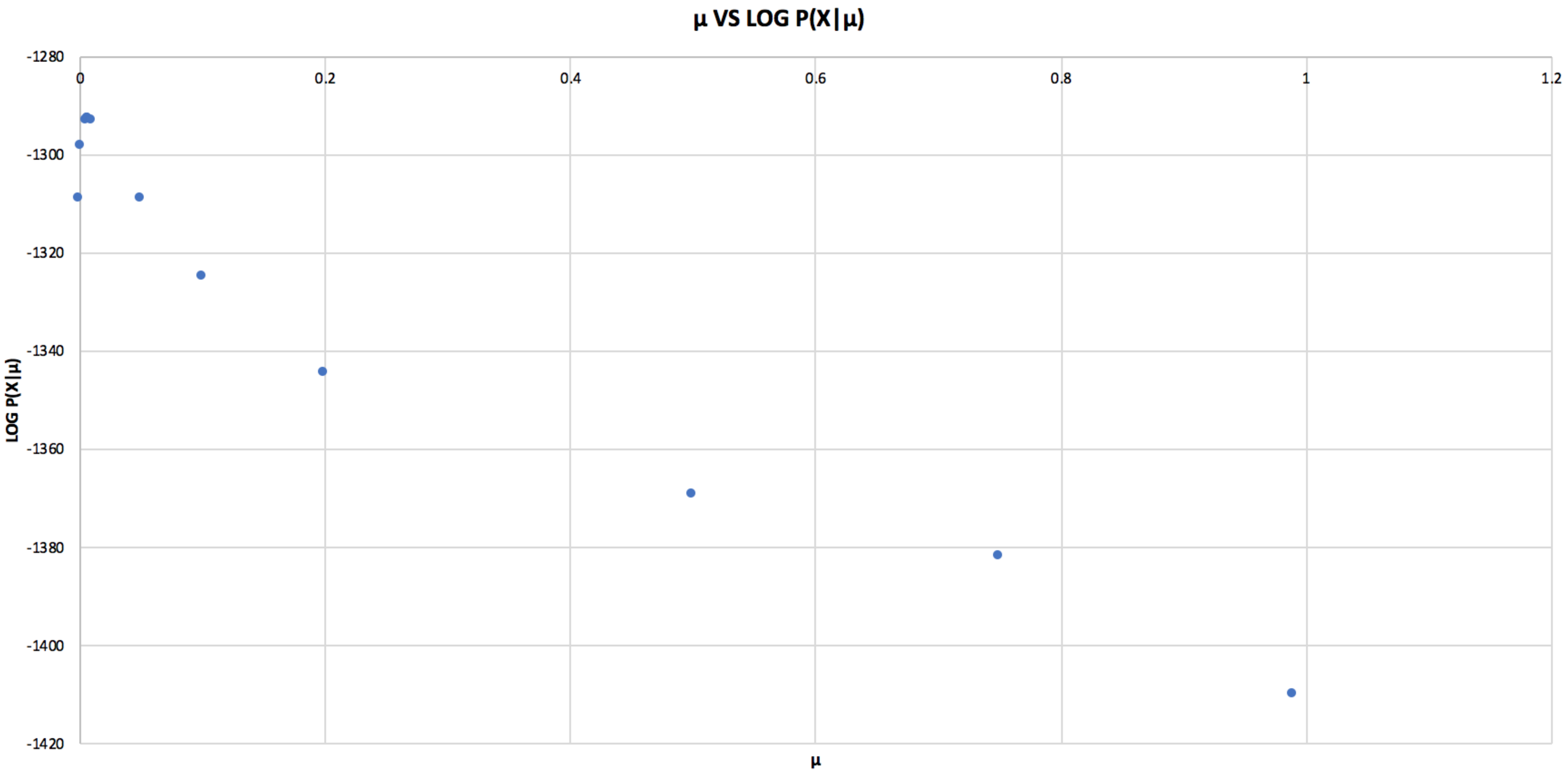
μ	$\ln P(x\|\mu)$	μ	$\ln P(x\|\mu)$
0.0001	-1309.121	0.05	-1308.874
0.001	-1298.423	0.1	-1325.005
0.006	-1292.987	0.2	-1344.422
0.007	-1292.887	0.5	-1369.246
0.008	-1292.896	0.75	-1381.872
0.01	-1293.142	0.99	-1409.911

plot attached as screenshot

It looks like the MLE of μ is around 0.007 to 0.0075.
The graph has a pretty sharp peak around that area before gently tapering off at high values.
If I used smaller values for μ the viterbi path should have fewer switches between the two states, while for a larger μ it should have more switches.


Probabilities vs Sequence Position Graph
1.20E+00
1.00E+00
8.00E-01
6.00E-01
4.00E-01
2.00E-01
0.00E+00
Probability
66
415
527
720
951
0
200
400
600
800
1000
Position
High G+C content region
Low G+C content region

mu	ln(P)
0.0001	-1309.121
0.001	-1298.423
0.006	-1292.987
0.007	-1292.887
0.008	-1292.896
0.01	-1293.142
0.05	-1308.874
0.1	-1325.005
0.2	-1344.422
0.5	-1369.246
0.75	-1381.872
0.99	-1409.911



2. $P(x_i \mid z_i = h) = \begin{cases} \frac{\theta_h}{2} & x_i = G \text{ or } C \\ \frac{1-2\theta_h}{2} & x_i = A \text{ or } T \end{cases}$ $\qquad P(x_i \mid z_i = l) = \begin{cases} \theta_l/2 & x_i = G \text{ or } C \\ \frac{1-2\theta_l}{2} & x_i = A \text{ or } T \end{cases}$

(a) $\log P(X, z \mid \mu, \theta_h, \theta_l) = ?$

$$P(x, z) = a_{0 z_1} \cdot e_{z_1}(x_1) \cdot \prod_{i=2}^{n} e_{z_i}(x_i) \cdot a_{z_{i-1}, z_i}$$

initial prob — initial emission (is a T) — emission at i — transition prob

$$P(x, z \mid \mu, \theta_h, \theta_l) = 0.5 \cdot \left(1 - \frac{\theta_l}{2}\right) \cdot \prod_{i=2}^{1000} P(x_i \mid z_i) \cdot P(z_i \mid z_{i-1})$$

$P(z_i \mid z_{i-1}) = \begin{cases} \mu & z_i \neq z_{i-1} \\ 1-\mu & z_i = z_{i-1} \end{cases}$

$$= 0.5 \cdot e_l(x_1) \cdot \prod_{i=2}^{65} e_l(x_i) \cdot (1-\mu) \cdot e_h(x_{66}) \cdot \mu$$
$$\cdot \prod_{i=67}^{417} e_h(x_i) \cdot (1-\mu) \cdot e_l(x_{418}) \cdot \mu$$
$$\cdot \prod_{i=419}^{467} e_l(x_i) \cdot (1-\mu) \cdot e_h(x_{468}) \cdot \mu$$
$$\cdot \ldots \cdot \prod_{i=947}^{1000} e_h(x_i) \cdot (1-\mu)$$

across all intervals

Since all products, we can move variables around

$$= 0.5 \cdot \mu^{c_b} \cdot (1-\mu)^{c_s} \cdot \prod_{i=1}^{65} e_l(x_i) \cdot \prod_{i=66}^{417} e_h(x_i) \cdot \ldots \cdot \prod_{i=946}^{1000} e_h(x_i)$$

$$= 0.5 \cdot \mu^{c_b} \cdot (1-\mu)^{c_s} \cdot \prod_{i \text{ in state } l} e_l(x_i) \cdot \prod_{i \text{ in state } h} e_h(x_i)$$

$$= 0.5 \cdot \mu^{c_b} \cdot (1-\mu)^{c_s} \cdot \left(\frac{\theta_l}{2}\right)^{d_{lG}} \cdot \left(\frac{1-2\theta_l}{2}\right)^{d_{lA}} \cdot \left(\frac{\theta_h}{2}\right)^{d_{hG}} \cdot \left(\frac{1-2\theta_h}{2}\right)^{d_{hA}}$$

$$\log P(x, z \mid \mu, \theta_h, \theta_l) = \log(\text{'' \quad ''})$$
$$= \log(0.5) + c_b \cdot \log(\mu) + c_s \cdot \log(1-\mu) + d_{lG} \cdot \log\left(\frac{\theta_l}{2}\right)$$
$$+ d_{lA} \cdot \log\left(\frac{1-2\theta_l}{2}\right) + d_{hG} \cdot \log\left(\frac{\theta_h}{2}\right) + d_{hA} \cdot \log\left(\frac{1-2\theta_h}{2}\right)$$

(b) To maximize $\log P(X, Z \mid \mu, \theta_h, \theta_\ell)$ take the derivative and set equal to 0. Since we have 3 variables, take the partial derivative of each

$$\log P(X,Z \mid \mu, \theta_h, \theta_\ell) = \log(0.5) + C_b \cdot \log(\mu) + C_s \cdot \log(1-\mu) + d_{\ell G} \cdot \log\left(\frac{\theta_\ell}{2}\right) + d_{\ell A} \cdot \log\left(\frac{1}{2} - \theta_\ell\right) + d_{hG} \cdot \log\left(\frac{\theta_h}{2}\right) + d_{hG} \cdot \log\left(\frac{1}{2} - \theta_h\right)$$

$$\frac{d}{d\mu} \log P(X,Z \mid \mu, \theta_h, \theta_\ell) = \frac{C_b}{\mu} - \frac{C_s}{1-\mu} = 0 \Rightarrow C_b/\mu = C_s/(1-\mu)$$

$$C_b(1-\mu) = C_s \cdot \mu$$

$$C_b - C_b \cdot \mu = C_s \cdot \mu$$

$$C_b = C_s \cdot \mu + C_b \cdot \mu$$

$$\mu = \frac{C_b}{C_s + C_b}$$

$$\mu = \frac{7}{993+7} = \frac{7}{1000} = 0.007$$

$$\frac{d}{d\theta_\ell} \log P(X,Z \mid \mu, \theta_h, \theta_\ell) = \frac{d_{\ell G}}{\frac{\theta_\ell}{2}} - \frac{d_{\ell A}}{\frac{1}{2} - \theta_\ell} = 0 \Rightarrow \frac{d_{\ell G} \cdot 2}{\theta_\ell} = \frac{d_{\ell A}}{\frac{1}{2} - \theta_\ell}$$

$$d_{\ell G}(1 - 2\theta_\ell) = d_{\ell A} \cdot \theta_\ell$$

$$d_{\ell G} = d_{\ell A}\theta_\ell + 2 d_{\ell G} \theta_\ell$$

$$\theta_\ell = \frac{d_{\ell G}}{d_{\ell A} + 2 d_{\ell G}}$$

$$\theta_\ell = \frac{115}{234+230} = \frac{115}{464} \approx 0.2478$$

$$\frac{d}{d\theta_h} \log P(X,Z \mid \mu, \theta_h, \theta_\ell) = \frac{d_{hG}}{\frac{\theta_h}{2}} - \frac{d_{hA}}{\frac{1}{2} - \theta_h} = 0 \Rightarrow d_{hG}(1 - 2\theta_h) = d_{hA} \cdot \theta_h$$

$$d_{hG} = d_{hA}\theta_h + 2 d_{hG} \theta_h$$

$$\theta_h = \frac{d_{hG}}{d_{hA} + 2 d_{hG}}$$

$$\theta_h = \frac{495}{156+990} = \frac{495}{1146} = 0.4319$$

We decomposed the parameters by taking partial derivatives which led to 3 equations contaning one parameter variable each, allowing us to maximize the parameters.

$\hat{\mu} = 0.007$, $\hat{\theta}_\ell = 0.2478$, $\hat{\theta}_h = 0.4319$

3(a) Code submitted online

Sequence outputs also submitted separately

The log likelihoods generated are pretty variable, with the range being about 150. This may not be too much when considering the mean in the -1500s, but since a difference of 2.3 corresponds to a 10 fold difference in likelihood, there is a big difference.

The transition mean and variance were 9.86 and 10.0004. I ran the code a few more times and the mean + variance was consistently around those numbers

This is very close to the mean and variance from the binomial model for 999 possible positions

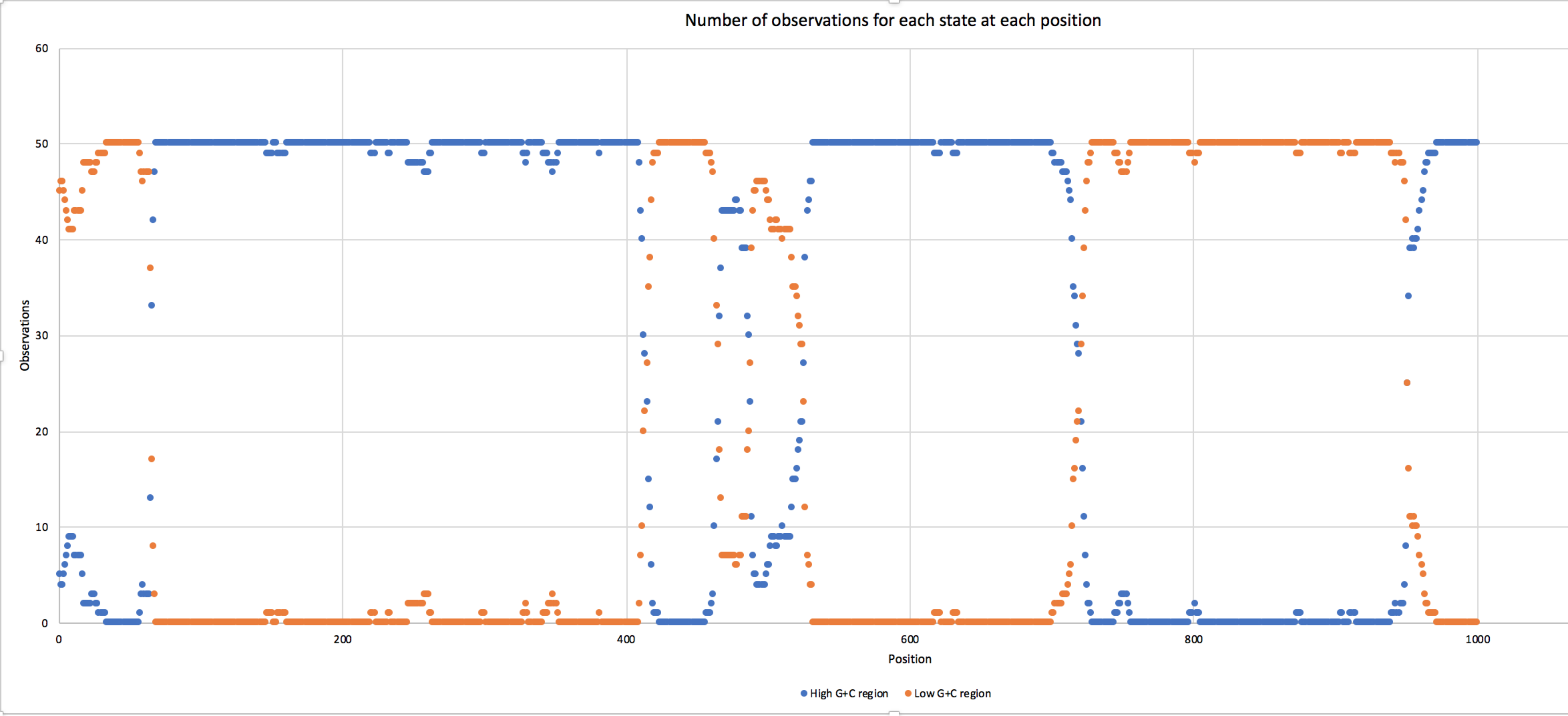
$\mu = n \cdot p = 999 \cdot 0.01 = 9.99$ close to 9.86

$\sigma^2 = n \cdot p(1-p) = 999 \cdot 0.01 \cdot 0.99 = 9.8901$ 10.0004

Sometimes the generated mean and variance were off by much more but this is reasonable from a small sample size of 50.

more but this is reasonable from a small sample size of 50.

(b) Code submitted online
sequence outputs also submitted separately
The log likelihoods for the 50 sampled state paths are all close, but not quite as large as the likelihood given in 1a by viterbi. This make sense since viterbi maximizes the likelihood.
The locations of the high and low G+C regions are all very similar, with some variation around the transition areas.
Plot on next page
Note: the plot was generated by a different runthrough of the code than the rest of the outputs were, so values are slightly different.
The plot of # states at each position is very similar to the posterior state probabilities plot from 1(b)


Number of observations for each state at each position
Observations
Position
0
10
20
30
40
50
60
0
200
400
600
800
1000
High G+C region
Low G+C region