Question 1

$P(\pi_3 = R \mid ATAAA) = P(ATAAA, \pi_3 = R)/P(ATAAA) = P(ATA, \pi_3 = R) * P(AA, \pi_3 = R)/P(ATAAA) = f_R(3) * g_V(3)/P(ATAAA)$

(1) Use forward algorithm to calculate $f_R(ATA)$ and $P(ATAAA)$ as follows:

$f_V(1) = P(A, \pi_1 = V) = e_V(A) * P(\pi_1 = V) = 0.25 * 0.5 = 0.125$

$f_R(1) = P(A, \pi_1 = R) = e_R(A) * P(\pi_1 = R) = 0.91 * 0.5 = 0.455$

$f_V(2) = e_V(T)t_{VV}f_V(1) + e_V(T)t_{RV}f_R(1) = 0.25 * 0.75 * 0.125 + 0.25 * 0.10 * 0.455 = 0.0348125$

$f_R(2) = e_R(T)t_{VR}f_V(1) + e_R(T)t_{RR}f_R(1) = 0.03 * 0.25 * 0.125 + 0.03 * 0.90 * 0.455 = 0.0132225$

$f_V(3) = e_V(A)t_{VV}f_V(2) + e_V(A)t_{RV}f_R(2) = 0.25 * 0.75 * 0.0348125 + 0.25 * 0.10 * 0.0132225 = 0.006857906$

$f_R(3) = e_R(A)t_{VR}f_V(2) + e_R(A)t_{RR}f_R(2) = 0.91 * 0.25 * 0.0348125 + 0.91 * 0.90 * 0.0132225 = 0.01874907$

$f_V(4) = e_V(A)t_{VV}f_V(3) + e_V(A)t_{RV}f_R(3) = 0.25 * 0.75 * 0.006857906 + 0.25 * 0.10 * 0.01874907 = 0.001754584$

$f_R(4) = e_R(A)t_{VR}f_V(3) + e_R(A)t_{RR}f_R(3) = 0.91 * 0.25 * 0.006857906 + 0.91 * 0.90 * 0.01874907 = 0.01691566$

$f_V(5) = e_V(A)t_{VV}f_V(4) + e_V(A)t_{RV}f_R(4) = 0.25 * 0.75 * 0.001754584 + 0.25 * 0.10 * 0.01691566 = 0.000751876$

$f_R(5) = e_R(A)t_{VR}f_V(4) + e_R(A)t_{RR}f_R(4) = 0.91 * 0.25 * 0.001754584 + 0.91 * 0.90 * 0.01691566 = 0.01425309$

$P(ATAAA) = 0.000751876 + 0.01425309 = 0.01500497$

(2) use backward algorithm to calculate $g_V(AAA)$ as follows:

$g_V(4) = t_{VR}e_R(A)g_R(5) + t_{VV}e_V(A)g_V(5) = 0.25 * 0.91 + 0.75 * 0.25 = 0.415$

$g_R(4) = t_{RR}e_R(A)g_R(5) + t_{RV}e_V(A)g_V(5) = 0.90 * 0.91 + 0.10 * 0.25 = 0.844$

$g_V(3) = t_{VR}e_R(A)g_R(4) + t_{VV}e_V(A)g_V(4) = 0.25 * 0.91 * 0.844 + 0.75 * 0.25 * 0.415 = 0.2698225$

$g_R(3) = t_{RR}e_R(A)g_R(4) + t_{RV}e_V(A)g_V(4) = 0.90 * 0.91 * 0.844 + 0.10 * 0.25 * 0.415 = 0.701611$

Therefore, $P(\pi_3 = R \mid ATAAA) = f_R(3) * g_V(3)/P(ATAAA) = 0.01874907 * 0.701611 / 0.01500497 =$ **0.8766798**

Question 2

Use $W = [w_1, w_2, \dots w_{10}]$ to represent the n=10 observed samples, we know that $W \sim Bin(1000, p)$

Estimate the parameter p using the following four methods:

1. Method of Moments

The p is estimated as $E[W] = 1000 * p = 1/n * \Sigma w_i$

$1000*p= 1/10*(24+33+42+30+44+38+27+39+47+51) = 37.5$

Therefore, **p = 0.00375**

2. Least Squares

This methods minimize the $\Sigma(w_i-E[W])^2 = n*(E[W])^2 - 2E[W]*\Sigma w_i + \Sigma w_i^2$

Take the derivative: $[\Sigma(w_i-E[W])^2]' = 2nE[W] - 2*\Sigma w_i = 0$

$E[W] = 1000p = 1/n * \Sigma w_i$

Therefore, **p = 0.00375**

3. Maximum likelihood

$P(\text{data} \mid p) = \pi[1000!/(w_i)!(1000-w_i)! * p^{w_i}(1-p)^{1000-w_i}] = C * p^{\Sigma w_i}(1-p)^{10000-\Sigma w_i}$ (C is a constant that does not involve p)
$\log[P(\text{data} \mid p)] = C + \Sigma w_i \log p + (10000 - \Sigma w_i) \log (1-p)$
The aim is to maximize $\log[P(\text{data} \mid p)]$. Take the derivative $\log'[P(\text{data} \mid p)] = \Sigma w_i/p - (10000 - \Sigma w_i)/(1-p) = 0$
$\Sigma w_i/p = (10000 - \Sigma w_i)/(1-p)$
Therefore, **$p = \Sigma w_i/10000 = 0.00375$**

4. Bayesian Method:
P(p | data) = P(data | p)P(p)/Integral[P(data | p)P(p)]
When assuming that the distribution for P(p) is uniform, P(p | data) = P(data | p)/Integral[P(data | p)]. Since Integral[P(data | p)] is a constant, the task is to minimize P(data | p), which is the same task as Maximum likelihood estimation described in 3. Therefore, **p = 0.00375**

Question 3
After two generations, the transition matrix is T^2 shown below (T is the Jukes-Cantor transition matrix):

	A	C	G	T
a	0.73	0.09	0.09	0.09
c	0.09	0.73	0.73	0.73
g	0.09	0.09	0.73	0.09
t	0.09	0.09	0.09	0.73

Use lower letter to represent the ancestral states at the position while use upper letter to represent states in S1 and S2 after two generations, and assume P(a) = P(c) = P(t) = P(g) = 0.25

P(1) = P(A and A at first position in S1 and S2) = P(a -> A)P(a -> A)P(a) + P(c -> A)P(c -> A)P(c) + P(g -> A)P(g -> A)P(g) + P(t -> A)P (t -> A)P(t) = 0.73 * 0.73 * 0.25 + 0.09 * 0.09 * 0.25 * 3 = 0.1393

P(2) = P(G and C at second position in S1 and S2) = P(a -> G)P(a -> C)P(a) + P(c -> G)P(c -> C)P(c) + P(g -> G)P(g -> C)P(g) + P(t -> G)P(t -> C)P(t) = 0.09 * 0.09 * 0.25 * 2 + 0.09 * 0.73 * 0.25 * 2 = 0.0369

Since the transition matrix is symmetric, and each state is uniformly distributed at each position in the ancestral string, the probability getting the same state and third and fifth position is the same as P(1). And the probability getting two different states at the fourth position is the same as P(2)

P(S1 and S2 after two generations) = P(1) * P(2) * P(3) * P(4) * P(5) = 0.1393 * 0.0369 * 0.1393 * 0.0369 * 0.1393 = 3.68e-6