BIOINFORMATICS & COMPUTATIONAL GENETICS MSc PROBABILITY AND STATISTICS

MOCK EXAMINATION : SOLUTIONS

1. (i) We have

$$\begin{aligned}
\mathrm{P}(E_A) &= \mathrm{P}(E_A|F_A)\,\mathrm{P}(F_A) + \mathrm{P}(E_A|F_C)\,\mathrm{P}(F_C) + \mathrm{P}(E_A|F_G)\,\mathrm{P}(F_G) + \mathrm{P}(E_A|F_T)\,\mathrm{P}(F_T)\\
&= (0.900 \times 0.30) + (0.025 \times 0.20) + (0.025 \times 0.20) + (0.050 \times 0.30) = 0.295
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}(E_C) &= \mathrm{P}(E_C|F_A)\,\mathrm{P}(F_A) + \mathrm{P}(E_C|F_C)\,\mathrm{P}(F_C) + \mathrm{P}(E_C|F_G)\,\mathrm{P}(F_G) + \mathrm{P}(E_C|F_T)\,\mathrm{P}(F_T)\\
&= (0.025 \times 0.30) + (0.850 \times 0.20) + (0.100 \times 0.20) + (0.025 \times 0.30) = 0.205
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}(E_G) &= \mathrm{P}(E_G|F_A)\,\mathrm{P}(F_A) + \mathrm{P}(E_G|F_C)\,\mathrm{P}(F_C) + \mathrm{P}(E_G|F_G)\,\mathrm{P}(F_G) + \mathrm{P}(E_G|F_T)\,\mathrm{P}(F_T)\\
&= (0.025 \times 0.30) + (0.100 \times 0.20) + (0.850 \times 0.20) + (0.025 \times 0.30) = 0.205
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}(E_T) &= \mathrm{P}(E_T|F_A)\,\mathrm{P}(F_A) + \mathrm{P}(E_T|F_C)\,\mathrm{P}(F_C) + \mathrm{P}(E_T|F_G)\,\mathrm{P}(F_G) + \mathrm{P}(E_T|F_T)\,\mathrm{P}(F_T)\\
&= (0.050 \times 0.30) + (0.025 \times 0.20) + (0.025 \times 0.20) + (0.900 \times 0.30) = 0.295
\end{aligned}$$

or, in matrix form

$$\begin{bmatrix} \mathrm{P}(E_A) \\ \mathrm{P}(E_C) \\ \mathrm{P}(E_G) \\ \mathrm{P}(E_T) \end{bmatrix} = \begin{bmatrix} \mathrm{P}(E_A|F_A) & \mathrm{P}(E_A|F_C) & \mathrm{P}(E_A|F_G) & \mathrm{P}(E_A|F_T) \\ \mathrm{P}(E_C|F_A) & \mathrm{P}(E_C|F_C) & \mathrm{P}(E_C|F_G) & \mathrm{P}(E_C|F_T) \\ \mathrm{P}(E_G|F_A) & \mathrm{P}(E_G|F_C) & \mathrm{P}(E_G|F_G) & \mathrm{P}(E_G|F_T) \\ \mathrm{P}(E_T|F_A) & \mathrm{P}(E_T|F_C) & \mathrm{P}(E_T|F_G) & \mathrm{P}(E_T|F_T) \end{bmatrix} \begin{bmatrix} \mathrm{P}(F_A) \\ \mathrm{P}(F_C) \\ \mathrm{P}(F_G) \\ \mathrm{P}(F_T) \end{bmatrix}$$

so that

$$\begin{bmatrix} 0.295 \\ 0.205 \\ 0.205 \\ 0.295 \end{bmatrix} = \begin{bmatrix} 0.900 & 0.025 & 0.025 & 0.050 \\ 0.025 & 0.850 & 0.100 & 0.025 \\ 0.025 & 0.100 & 0.850 & 0.025 \\ 0.050 & 0.025 & 0.025 & 0.900 \end{bmatrix} \begin{bmatrix} 0.30 \\ 0.20 \\ 0.20 \\ 0.30 \end{bmatrix}$$

(note there are some symmetries in the probability specification that simplify the calculation)

8 MARKS

(ii) Using the Bayes Theorem formula

$$\mathrm{P}(F_A|E_A) = \frac{\mathrm{P}(E_A|F_A)\,\mathrm{P}(F_A)}{\mathrm{P}(E_A)} = \frac{0.900 \times 0.30}{0.295} = 0.9153$$

$$\mathrm{P}(F_A|E_C) = \frac{\mathrm{P}(E_C|F_A)\,\mathrm{P}(F_A)}{\mathrm{P}(E_C)} = \frac{0.025 \times 0.20}{0.205} = 0.0244$$

$$\mathrm{P}(F_A|E_G) = \frac{\mathrm{P}(E_G|F_A)\,\mathrm{P}(F_A)}{\mathrm{P}(E_C)} = \frac{0.025 \times 0.20}{0.205} = 0.0244$$

$$\mathrm{P}(F_A|E_T) = \frac{\mathrm{P}(E_T|F_A)\,\mathrm{P}(F_A)}{\mathrm{P}(E_T)} = \frac{0.050 \times 0.30}{0.295} = 0.0508$$

4 MARKS

(iii) We have, for event M,

$$M \equiv (E_A' \cap F_A) \cup (E_C' \cap F_C) \cup (E_G' \cap F_G) \cup (E_T' \cap F_T)$$

that is, the union of mutually exclusive events, so that, by axiom (III),

$$\begin{aligned}
\mathrm{P}(M) &= \mathrm{P}(E_A' \cap F_A) + \mathrm{P}(E_C' \cap F_C) + \mathrm{P}(E_G' \cap F_G) + \mathrm{P}(E_T' \cap F_T) \\
&= \mathrm{P}(E_A'|F_A)\,\mathrm{P}(F_A) + \mathrm{P}(E_C'|F_C)\,\mathrm{P}(F_C) + \mathrm{P}(E_G'|F_G)\,\mathrm{P}(F_G) + \mathrm{P}(E_T'|F_T)\,\mathrm{P}(F_T) \\
&= [1 - \mathrm{P}(E_A|F_A)]\,\mathrm{P}(F_A) + [1 - \mathrm{P}(E_C|F_C)]\,\mathrm{P}(F_C) + [1 - \mathrm{P}(E_G|F_G)]\,\mathrm{P}(F_G) + [1 - \mathrm{P}(E_T|F_T)]\,\mathrm{P}(F_T) \\
&= [1 - 0.900] \times 0.30 + [1 - 0.850] \times 0.20 + [1 - 0.850] \times 0.20 + [1 - 0.900] \times 0.30 \\
&= 0.12
\end{aligned}$$

Note also, by the total probability formula

$$\mathrm{P}(M) = \mathrm{P}(M|F_A)\,\mathrm{P}(F_A) + \mathrm{P}(M|F_C)\,\mathrm{P}(F_C) + \mathrm{P}(M|F_G)\,\mathrm{P}(F_G) + \mathrm{P}(M|F_T)\,\mathrm{P}(F_T)$$

that gives the same result. Note also that the probability of correct classification for any base is

$$\mathrm{P}(M') = 1 - \mathrm{P}(M) = 1 - 0.12 = 0.88$$

.

4 MARKS

(iv) For a sequence of k bases, need each base to be correctly classified, and as the appearance of bases and the base classifications are mutually independent, the required probability is

$$\overbrace{\mathrm{P}(M')}^{\text{Position 1}} \times \overbrace{\mathrm{P}(M')}^{\text{Position 2}} \times \quad \ldots \quad \times \overbrace{\mathrm{P}(M')}^{\text{Position } k} = \{\mathrm{P}(M')\}^k = \{0.88\}^k$$

2 MARKS

(v) The probability required is

$$\overbrace{\mathrm{P}(M')}^{\text{Correct at position 1}} \times \overbrace{\mathrm{P}(M')}^{\text{Correct at position 2}} \times \ldots \times \overbrace{\mathrm{P}(M')}^{\text{Correct at position } x-1} \times \overbrace{\mathrm{P}(M)}^{\text{Misclassification at position } x}$$

$$= \{\mathrm{P}(M')\}^{x-1}\,\mathrm{P}(M) = \{0.88\}^{x-1} \times 0.12$$

that is the *Geometric* (0.12) distribution

2 MARKS

(v) For a sequence of length L, under the independence assumptions above, we have that the total number of misclassifications in the sequence is a discrete random variable, X say, where $X \sim Binomial(L,\mathrm{P}(M))$ (think of a corresponding binary sequence of length L with 1s representing misclassifications and 0s representing correct classifications, where the 1s and 0s appear independently with P("1") =$\mathrm{P}(M) = 0.12$), so that

$$\mathrm{P}[X = 0] = \binom{L}{0}(0.12)^0(1 - 0.12)^L = (0.88)^L$$

$$\mathrm{P}[X = 1] = \binom{L}{1}(0.12)^1(1 - 0.12)^{L-1} = L\,(0.12)\,(0.88)^{L-1}$$

and thus the probability that there is at most one misclassification is

$$\mathrm{P}[X = 0] + \mathrm{P}[X = 1] = (0.88)^L + L\,(0.12)\,(0.88)^{L-1} = (0.88)^{L-1}(0.88 + 0.12 \times L)$$

which we can compute for different values of L

L	1	2	3	4	5	6
Prob	1.0000	0.9856	0.9603	0.9268	0.8875	0.8444

which dips below 0.95 for the first time when $L = 4$, so the longest sequence that can be analyzed is of length 3.

4 MARKS

2. (a) (i) If $t_1 = 5000$ then $\lambda t_1 = 0.001 \times 5000 = 5$. Hence, using the Poisson mass function formula

$$\begin{aligned}
\mathrm{P}[X_1 = 2] &= \frac{e^{-5}5^2}{2!} = 0.0842 \\
\mathrm{P}[X_1 \geq 0] &= 1 \\
\mathrm{P}[X_1 < 4] &= P[X_1 = 0] + P[X_1 = 1] + P[X_1 = 2] + P[X_1 = 3] \\
&= \frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^1}{1!} + \frac{e^{-5}5^2}{2!} + \frac{e^{-5}5^3}{3!} \\
&= 0.0067 + 0.0337 + 0.0842 + 0.1404 = 0.2650
\end{aligned}$$

4 MARKS

(ii) The assumptions underlying the Poisson process mean that we can consider the entire 20000 base segment in its entirety and forget that we were asked to consider ten subsections. Thus, from the first Poisson process distribution result, as $0.001 \times 20000 = 20$, we must have

$$S_{10} \sim Poisson(20)$$

so that

$$\begin{aligned}
\mathrm{P}[S_{10} \geq 3] &= 1 - \mathrm{P}[S_{10} < 3] = 1 - \{P[S_{10} = 0] + P[S_{10} = 1] + P[S_{10} = 2]\} \\
&= 1 - \left\{\frac{e^{-20}20^0}{0!} + \frac{e^{-20}20^1}{1!} + \frac{e^{-20}20^2}{2!}\right\} \\
&= 0.9999995
\end{aligned}$$

4 MARKS

(iii) Need to use the second distributional result that $T_1 \sim Exponential(\lambda)$ so that the pdf of T_1 is given (from the formula sheet) by

$$f_{T_1}(t) = \lambda e^{-\lambda t} \qquad t > 0$$

Now, we require

$$\mathrm{P}[T_1 > 7500] = 1 - \mathrm{P}[T_1 \leq 7500]$$

which cannot be computed from the pdf; the cdf is needed. Can quote the cdf from notes

$$F_{T_1}(t) = 1 - e^{-\lambda t} \qquad t > 0$$

or deduce it by integration of the pdf

$$F_{T_1}(t) = \int_0^t f_{T_1}(x)\, dx = \int_0^t \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_0^t = 1 - e^{-\lambda t}.$$

Hence

$$\mathrm{P}[T_1 > 7500] = 1 - \mathrm{P}[T_1 \leq 7500] = 1 - \left(1 - e^{-0.001 \times 7500}\right) = e^{-0.001 \times 7500} = 0.00055$$

Alternately, could set $t_1 = 7500$, and use the Poisson mass function to compute $\mathrm{P}[X_1 = 0]$ (that is, the probability of zero occurrences of the pattern in the first 7500 bases), where

$$X_1 \sim Poisson(\lambda t_1) \equiv Poisson(0.001 \times 7500) \equiv Poisson(7.5)$$

which gives

$$\mathrm{P}[X_1 = 0] = \frac{e^{-7.5}7.5^0}{0!} = e^{-7.5}$$

as above

4 MARKS

(iv) Need the third distributional result to deduce that the position at which the 10th occurrence of the pattern is observed, Y_{10}, has a Gamma distribution

$$Y_{10} \sim Gamma\,(n,\lambda) = Gamma\,(10, 0.001) \approx Normal\left(\frac{n}{\lambda},\frac{n}{\lambda^2}\right) = Normal\left(1\times 10^4, 1\times 10^7\right)$$

and thus we approximate the required probability by noting that

$$P\left[Y_{10} \leq 10000\right] \approx \Phi\left(\frac{0.001\times 10000-10}{\sqrt{10}}\right) = \Phi\left(\frac{10-10}{\sqrt{10}}\right) = \Phi\,(0) = 0.5$$

The exact probability, using SPLUS, is 0.5421,

```
> pgamma(10000,10,0.001)
[1] 0.5420703
```

so, in fact, the approximation is not that good.

3 MARKS

(b) For the data sample provided

$$\bar{x} = 0.188 \qquad s^2 = 3.48924 \qquad s = 1.867951 \qquad n = 10$$

2 MARKS

(i) First test

$$t = \frac{\bar{x}-c}{s/\sqrt{n}} = \frac{0.188-0.00}{1.867951/\sqrt{10}} = \frac{0.188}{0.5906979} = 0.3182676$$

The critical values in this **two-sided** test are given below (from tables of the *Student*(9) distribution)

α	Critical Values
0.05	± 2.2622
0.01	± 3.2498

Test statistic **not more extreme** than critical values $\Longrightarrow$ CANNOT REJECT H_0

4 MARKS

The p-value cannot be computed from tables, but from SPLUS, we have a complete calculation as follows

```
> x <- c(1.9,-1.1,1.2,-0.11,-1.90,2.2,-0.87,0.76,-2.90,2.7)
> n <- length(x)
> t <- mean(x)/(sqrt(var(x))/sqrt(n))
> t
[1] 0.3182676
> pvalue <- pt(-abs(t),n-1)+1-pt(abs(t),n-1)
> pvalue
[1] 0.7575432
```

(ii) Note that the second test is **one-sided,** so using the observed test statistic

$$q = \frac{(n-1)s^2}{\sigma^2} = \frac{9\times 3.48924}{1} = 31.40316$$

we can test the hypothesis by looking up the critical values in this **one-sided** test, that are given below (from tables of the χ^2_9 distribution)

α	Critical Value
0.05	16.919
0.01	21.666

(that is, we look up the 0.950 and 0.990 points of the χ^2_9 distribution).

The test statistic is **more extreme** than critical value $\Longrightarrow$ REJECT H_0

4 MARKS

Using SPLUS, we have the pvalue as 0.0002523893

```
> q <- (n-1)*var(x)/1
> 1-pchisq(q,9)
[1] 0.0002523893
```

3. (a) From notes, we have that the probability of a match at any given position is

$$p_{MATCH} = p_A^2 + p_C^2 + p_G^2 + p_T^2$$

This formula assumes that the nucleotides in each sequence are sampled **independently** from the same multinomial distribution within each sequence and between sequences, and uses the Total Probability result

$$p_{MATCH} = \Pr(\text{Match }) = \sum_{i \in \{A,C,G,T\}} \mathrm{P}(\text{Match } \cap \text{Character is } i) = \sum_{i \in \{A,C,G,T\}} (p_i \times p_i) = \sum_{i \in \{A,C,G,T\}} p_i^2$$

(i) For the probabilities given,

$$p_{MATCH} = 0.30^2 + 0.20^2 + 0.20^2 + 0.30^2 = 0.26$$

2 MARKS

(ii) Under H_0, the test statistic X_{MATCH} has a $Binomial\,(N, p_{MATCH})$ distribution. To see this, consider the Match/Non-Match sequence - it is a binary sequence of length N where the 1s correspond to the Matches and the 0s to the non-Matches, where 1s and 0s appear in the N positions independently with probabilities p_{MATCH} and $1 - p_{MATCH}$ respectively. This is the experimental situation that gives rise to the *Binomial* distribution.

3 MARKS

(iii) If we observe x_{MATCH} matched positions, then we may use x_{MATCH} as a test statistic in a test of alignment as follows. If H_0 is true, then, from (ii) we have that

$$X_{MATCH} \sim Binomial(100, 0.26)$$

so that we would expect the observed value of x_{MATCH} to be a plausible observation from this distribution; if it is not, that is, if it is in the tails of the distribution, then we may have evidence to reject H_0 in favour of the alternative H_1 that admits the possibility that the sequences are evolutionarily related. Specifically, if x_{MATCH} is too large/too small, then we may reject H_0; the only decision left to make is choose appropriate critical values to quantify precisely what "large" or "small" constitutes. From first principles, for a hypothesis test at significance level α, we wish to find constants c_1 and c_2 such that

$$\mathrm{P}\left[c_1 \leq X_{MATCH} \leq c_2\right] = 1 - \alpha$$

So for $\alpha = 0.05$, we could merely choose the 0.025 and 0.975 quantiles of the $Binomial(100, 0.26)$ distributions; these are given in the question as

$$c_1 = 18 \qquad\qquad c_2 = 35$$

5 MARKS

that can be obtained from SPLUS

```
> qbinom(0.025,100,0.26)
[1] 18
> qbinom(0.975,100,0.26)
[1] 35
```

(Note: these critical values may be obtained - approximately - by using the distributional approximation

$$Binomial(n, \theta) \approx Normal(n\theta, n\theta(1-\theta)) \qquad \therefore \qquad Binomial(100, 0.26) \approx Normal(26, 19.24)$$

and from tables the 0.025 and 0.975 quantiles of the $Normal(26, 19.24)$ distribution are

$$26 \pm 1.96 \times \sqrt{19.24} = (17.40277 : 34.59723)$$

that are approximately correct)

(b) For an alignment of five sequences, we can again derive a binary Match/non-Match sequence by recording positions at which **all five** sequences are identical. Using the same approach to that in (a), we have a new match probability

$$p_{MATCH} = \Pr(\text{Match}) = \sum_{i\in\{A,C,G,T\}} \mathrm{P}(\text{Match} \cap \text{Character is } i) = \sum_{i\in\{A,C,G,T\}} (p_i \times p_i \times p_i \times p_i \times p_i) = \sum_{i\in\{A,C,G,T\}} p_i^5$$

so that, here,

$$p_{MATCH} = 0.30^5 + 0.20^5 + 0.20^5 + 0.30^5 = 0.0055$$

Again, we have a test statistic X_{MATCH} where

$$X_{MATCH} \sim Binomial(100, 0.0055)$$

if H_0 is true; again the critical values are computed by studying

$$\mathrm{P}[c_1 \leq X_{MATCH} \leq c_2] = 1 - \alpha$$

in the null distribution. We can perform an exact calculation for the critical values using the Binomial distribution, or an approximation using the Poisson distribution as

$$Binomial(n,\theta) \approx Poisson(n\theta) \qquad \therefore \qquad Binomial(100, 0.0055) \approx Poisson(0.55)$$

or an approximation using the Normal distribution as above

$$Binomial(n,\theta) \approx Normal(n\theta, n\theta(1-\theta)) \qquad \therefore \qquad Binomial(100, 0.0055) \approx Normal(0.55, 0.547)$$

5 MARKS

SPLUS gives critical values as follows:

```
> qbinom(c(0.025,0.975),100,0.0055)
[1] 0 2
> qpois(c(0.025,0.975),100*0.0055)
[1] 0 2
> qnorm(c(0.025,0.975),100*0.0055,sqrt(100*0.0055*(1-0.0055)))
[1] -0.8995454  1.9995454
```

So, if we observe more than two match positions then we can reject H_0.
(c) (i) If the longest run of matches is recorded for a given pair of sequences of equal length, then, if H_0 is true and the sequences are evolutionarily unrelated, then,

$$F_{Y_n}(y) = \mathrm{P}[Y_n \leq y] = \mathrm{P}[Y_n < y+1] = 1 - \mathrm{P}[Y_n \geq y+1] \qquad y = 0, 1, 2, 3, ...$$

so that, from notes (p 64) where $\mathrm{P}[Y_n \geq x] = 1 - (1-p^x)^n$ for $x = 1, 2, 3, ...$we have by substituting in $x = y+1$

$$F_{Y_n}(y) = 1 - \left\{1 - \left(1 - p_{MATCH}^{y+1}\right)^n\right\} = \left(1 - p_{MATCH}^{y+1}\right)^n \qquad y = 0, 1, 2, 3, ...$$

Thus we can assess the plausibility of H_0 by comparing the observed test statistic with this null distribution; if the longest run of matches has length y_n, then the probability of observing a **more extreme** test statistic than the one that **was actually** observed (that is, the p-value) is

$$\mathrm{P}[Y_n > y_n] = 1 - F_{Y_n}(y_n) = 1 - \left(1 - p_{MATCH}^{y_n+1}\right)^n$$

2 MARKS

(ii) The p-value formula given is

$$p \approx 1 - \exp\{-(1-p_{MATCH})Np_{MATCH}^{y_n}\}$$

and for $N = 1000$ and $p_{MATCH} = 0.26$ so that $(1 - p_{MATCH})N = 740$, this reduces to

$$p \approx 1 - \exp\{-740 \times (0.26)^{y_n}\}$$

so that for different values of y_n we observe the following p-values.

y_n	0	1	2	3	4	5	6	7	8	9	10
p	1.0000	1.0000	1.0000	1.0000	0.9660	0.5849	0.2044	0.0577	0.0153	0.0040	0.0010

Thus an appropriate critical value (for significance level $\alpha = 0.05$) is $c_R = 8$; if the maximum run length is 8 or greater then there is evidence to reject H_0. For significance level $\alpha = 0.01, c_R = 9$

4 MARKS

(iii) For $N = 1000$ and a five sequence match, so that $p_{MATCH} = 0.0055$ and $(1 - p_{MATCH})N = 994.5$, this reduces to

$$p \approx 1 - \exp\{-994.5 \times (0.0055)^{y_n}\}$$

To find the critical values we must solve

$$\alpha \approx 1 - \exp\{-994.5 \times (0.0055)^{y_n}\}$$

for y_n in terms of α; re-arranging this equation we have

$$y_n \approx \frac{\log\left(-\dfrac{1}{994.5}\log(1-\alpha)\right)}{\log 0.0055}$$

Using the original approximation formula we have

y_n	0	1	2	3	4
p	1.0000	0.9960	0.030	0.0000	0.0000

giving critical values of 2 and 3 for $\alpha = 0.05$ and $\alpha = 0.01$ respectively

4 MARKS

4. (a)(i) Estimates are given by

$$\hat{p}_i = \frac{n_i}{n} \qquad i = 1, 2, 3, 4$$

that is

$$\hat{p}_1 = \hat{p}_A = \frac{1821}{6853} = 0.2657 \qquad \hat{p}_3 = \hat{p}_G = \frac{1304}{6853} = 0.1903$$

$$\hat{p}_2 = \hat{p}_C = \frac{1348}{6853} = 0.1967 \qquad \hat{p}_4 = \hat{p}_T = \frac{2380}{6853} = 0.3473$$

4 MARKS

(ii) To test the hypothesis use the following table of fitted values

	Nucleotide				
	1	2	3	4	Total
$\hat{n}_i$	1713.25	1713.25	1713.25	1713.25	6853

- if H_0 is true then the fitted values for each nucleotide are **equal** as the hypothesized probabilities are equal.

Hence

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \frac{(1821 - 1713.25)^2}{1713.25} + \frac{(1348 - 1713.25)^2}{1713.25} + \frac{(1304 - 1713.25)^2}{1713.25} + \frac{(2380 - 1713.25)^2}{1713.25} = 441.8846$$

$$LR = 2\sum_{i=1}^{4} n_i \log\frac{n_i}{\hat{n}_i} = 2\left(1821\log\frac{1821}{1713.25} + 1348\log\frac{1348}{1713.25} + 1304\log\frac{1304}{1713.25} + 2380\log\frac{2380}{1713.25}\right) = 428.5018$$

Both of these test statistics give strong evidence for **rejecting** H_0

6 MARKS

(b) (i) For the test, the fitted values are;

	Nucleotide				
	A	C	G	T	Total
Sequence 1	228	164	196	212	800
Sequence 2	342	246	294	318	1200
Total	570	410	490	530	2000

where, for example

$$n_{11} = n_{1.}\hat{p}_1 = \frac{n_{1.}n_{.1}}{n} = \frac{800 \times 570}{2000} = 228 \qquad n_{12} = n_{1.}\hat{p}_2 = \frac{n_{1.}n_{.2}}{n} = \frac{800 \times 410}{2000} = 164$$

and so on.

8 MARKS

(ii) Test statistics

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = 14.11571 \qquad LR = 2\sum_{i=1}^{2}\sum_{j=1}^{4} n_{ij}\log\frac{n_{ij}}{\hat{n}_{ij}} = 14.19433$$

4 MARKS

(iii) For both tests, compare with the 0.95 quantile of the χ^2_3 distribution, that is 7.81. Clearly, **both** tests indicate that there is **evidence to reject** the hypothesis that the nucleotide probabilities are identical for the two sequences.

3 MARKS