36-467 Homework 12

Eu Jing Chua
eujingc

December 3, 2018

Question 1

Table 1: Counts of each base

dicty	Freq
A	1886024
C	559670
G	560462
N	607
T	1916833

The sequence has 4923596 bases.

Question 2

Table 2: Counts of all successive pairs of bases

succ.pairs	Freq
AA	876552
AC	190849
AG	163053
AN	6
AT	655564
CA	262777
CC	95410
CG	35031
CN	2
CT	166450
GA	214211
GC	54274
GG	96292
GN	3
GT	195682
NA	2
NC	2
NG	3
NN	594
NT	5
TA	532482
TC	219135
TG	266083
TN	1
TT	899132

Question 3

Table 3: Transition Matrix of First-Order Markov Chain

	A	C	G	N	T
A	0.46476	0.10119	0.08645	0.00000	0.34759
C	0.46952	0.17048	0.06259	0.00000	0.29741
G	0.38220	0.09684	0.17181	0.00001	0.34914
N	0.00330	0.00330	0.00495	0.98020	0.00825
T	0.27779	0.11432	0.13881	0.00000	0.46907

Question 4

The log-likelihood of the estimated first-order Markov chain is -5.939×10^6.

Question 5

The upper and lower bounds of 95% C.I. for each entry in the transition matrix is as follows:

	A	C	G	N	T
A	0.46547	0.10162	0.08685	0.00001	0.34827
C	0.47083	0.17146	0.06323	0.00001	0.29860
G	0.38348	0.09761	0.17280	0.00001	0.35039
N	0.00825	0.00825	0.01155	0.99175	0.01650
T	0.27843	0.11477	0.13930	0.00000	0.46978

	A	C	G	N	T
A	0.46405	0.10076	0.08605	0.0000	0.34691
C	0.46821	0.16949	0.06196	0.0000	0.29621
G	0.38093	0.09606	0.17082	0.0000	0.34790
N	0.00000	0.00000	0.00000	0.9703	0.00165
T	0.27716	0.11387	0.13832	0.0000	0.46837

These intervals seem unusually small, being very tight.

Question 6

Table 6: Distributions of the Markov chain

	Invariant Dist.	Overall Dist.
A	0.38306	0.38306
C	0.11367	0.11367
G	0.11383	0.11383
N	0.00012	0.00012
T	0.38932	0.38932

From the above table, we can see that the distributions are approximately the same.

Question 7

Table 7: Transition Matrix of Second-Order Markov Chain

	AA	AC	AG	AN	AT	CA	CC	CG	CN	CT	GA	GC	GG	GN	GT	NA	NC	NG	NN	NT	TA	TC	TG	TN	TT
AA	0.50	0.09	0.08	0	0.33	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
AC	0.00	0.00	0.00	0	0.00	0.41	0.24	0.06	0	0.29	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
AG	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.42	0.11	0.13	0	0.34	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
AN	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.33	0	0.50	0.17	0.00	0.00	0.00	0	0.00
AT	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.30	0.13	0.12	0	0.45
CA	0.46	0.14	0.10	0	0.30	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
CC	0.00	0.00	0.00	0	0.00	0.60	0.12	0.05	0	0.23	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
CG	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.43	0.11	0.14	0	0.32	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
CN	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.50	0.00	0	0.50	0.00	0.00	0.00	0.00	0	0.00
CT	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.27	0.16	0.15	0	0.42
GA	0.42	0.07	0.12	0	0.39	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
GC	0.00	0.00	0.00	0	0.00	0.44	0.14	0.07	0	0.34	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
GG	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.30	0.08	0.11	0	0.50	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
GN	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.33	0.00	0	0.33	0.33	0.00	0.00	0.00	0	0.00
GT	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.28	0.08	0.19	0	0.44
NA	0.50	0.00	0.00	0	0.50	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
NC	0.00	0.00	0.00	0	0.00	0.00	0.50	0.00	0	0.50	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
NG	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.33	0.00	0.33	0	0.33	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
NN	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.99	0.01	0.00	0.00	0.00	0	0.00
NT	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.40	0.20	0.20	0	0.20
TA	0.43	0.10	0.08	0	0.38	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
TC	0.00	0.00	0.00	0	0.00	0.47	0.14	0.07	0	0.32	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
TG	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.38	0.09	0.22	0	0.31	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00
TN	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	1	0.00	0.00	0.00	0.00	0.00	0	0.00
TT	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.26	0.10	0.14	0	0.50

Question 8

The log-likelihood of the estimated second-order Markov chain is -5.903×10^6.

Question 9

We test:

$$H_0 : \text{The model is a first-order Markov chain} \tag{1}$$
$$H_a : \text{The model is a second-order Markov chain} \tag{2}$$

Our test statistic is 7.2071×10^4, which follows a χ^2 distribution with degrees of freedom 80. This gives us a p-value very close to 0, hence we reject H_0 and conclude that there is sufficient evidence that the model is actually a second-order Markov chain.

Question 10

Table 8: Transition Matrix of First-Order Markov Chain for Second set of Chromosomes

	A	C	G	N	T
A	0.47110	0.09938	0.08336	0.00000	0.34616
C	0.48135	0.17049	0.05785	0.00001	0.29031
G	0.38735	0.09427	0.17050	0.00001	0.34789
N	0.00783	0.00104	0.00157	0.98382	0.00574
T	0.28286	0.11145	0.13628	0.00000	0.46940

The upper and lower bound of the 95% C.I. for each entry in the transition matrix is as follows:

	A	C	G	N	T
A	0.26903	0.05691	0.04776	0.00000	0.19777
C	0.28582	0.10154	0.03462	0.00001	0.17262
G	0.23198	0.05673	0.10236	0.00001	0.20840
N	0.00470	0.00157	0.00157	0.31472	0.00418
T	0.16514	0.06518	0.07967	0.00000	0.27384

	A	C	G	N	T
A	0.26821	0.05642	0.04731	0.00000	0.19700
C	0.28427	0.10038	0.03389	0.00000	0.17121
G	0.23045	0.05581	0.10119	0.00000	0.20692
N	0.00052	0.00000	0.00000	0.30846	0.00000
T	0.16440	0.06466	0.07910	0.00000	0.27302

Question 11

The log-likelihood of the first-order Markvoc chain for the second set of chromosomes is -1.0169×10^7. The total log-likelihood of the first-order Markov chains of both sets of chromosomes is -1.6108×10^7.

Question 12

Table 11: Transition Matrix of First-Order Markov Chain for both chromosomes

	A	C	G	N	T
A	0.4687960	0.1000364	0.0844817	0.0000037	0.3466823
C	0.4769516	0.1704840	0.0596125	0.0000047	0.2929473
G	0.3854238	0.0952270	0.1709880	0.0000053	0.3483559
N	0.0067407	0.0015860	0.0023791	0.9829500	0.0063442
T	0.2809977	0.1125097	0.1372132	0.0000017	0.4692777

Question 13

The log-likelihood for the pooled fisrt-order Markov chain is -1.6108×10^7.

Question 14

We test:

$$H_0 : \text{Both sets of chromosomes are modelled with a shared first-order Markov chain} \quad (3)$$

$$H_a : \text{Both sets of chromosomes are individually modelled with first-order Markov chains} \quad (4)$$

Our test statistic is 881.63, which follows a χ^2 distribution with degrees of freedom 20. This gives us a p-value of 6.3558×10^{-174} which is close to 0, hence we reject H_0 at 0.05 significance and conclude that there is sufficient evidence that both sets of chromosomes can actually be modelled with a shared first-order Markov chain.