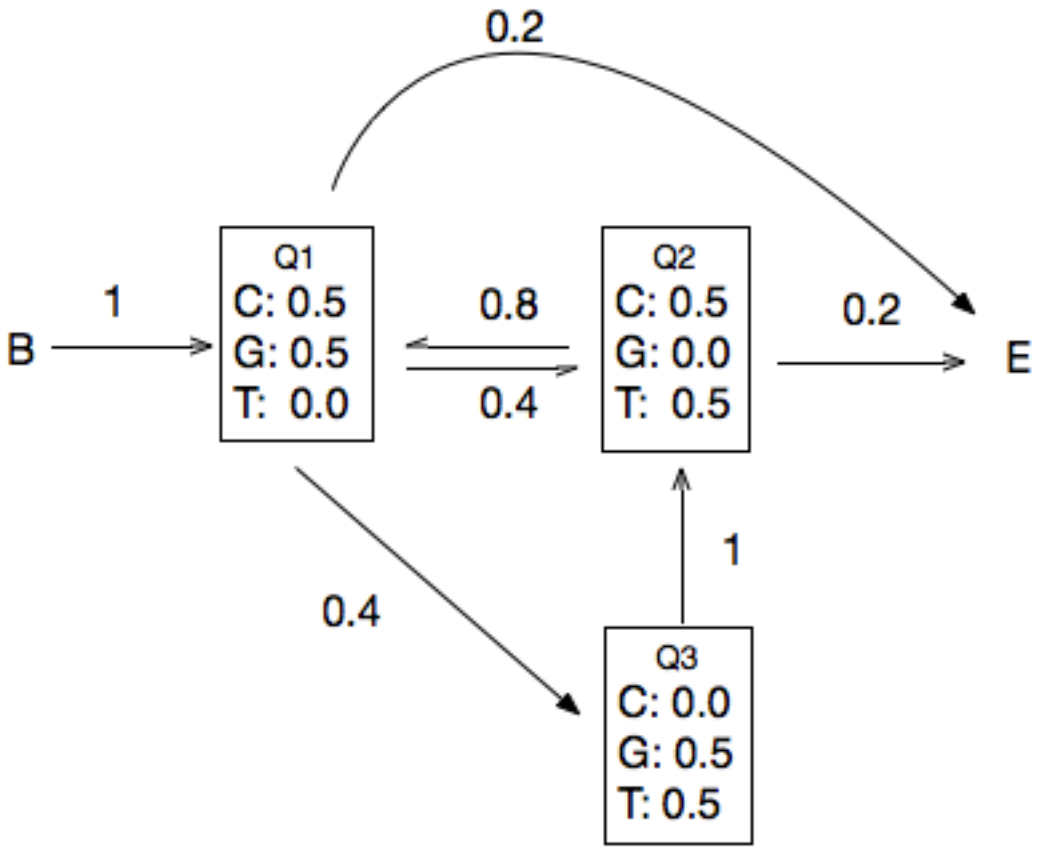
Abbas Rizvi
Assignment 4 – Hidden Markov Models

PART I:



1 a.

1b.
X_1 = CGT = 0.01
X_2 = CTC = 0.0018

2a.
P(X=ATG, Q=BQ1Q2Q1E | HMM) = 0.00003125

2b.

	- 0	A 1	T 2	G 3	- 4
B	1	0	0	0	0
Q1	0	0.13	0.022	0.0038	0
Q2	0	0.5	0.0015	0.001	0
E	0	0	0	0	0.0007
Most probable: Q = B,Q1,Q1,Q1,E					

2c. The Viterbi algorithm calculates the probability for the most likely path. The forward algorithm calculates the probability that a sequence given all the possible paths. The forward will be either higher or equal to the result obtained in b as more information is provided. In this case it is higher.

Abbas Rizvi
Assignment 4 – Hidden Markov Models

PART II.

1.

Viterbi Matrix:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	-	C	C	H	H	P	C	C	P	H	H	C	H	-
B	1	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0.25	0.0875	0.030625	0.01071875	0	0.000107188	3.75E-05	0	4.60E-06	1.61E-06	5.63E-07	1.97E-07	0
D	0	0.25	0.0875	0	0	0.001071875	0.000375156	0.000131305	4.60E-05	0	0	1.61E-07	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	1.97E-08

Most probable sequence: ['B', 'L', 'L', 'L', 'L', 'D', 'D', 'D', 'D', 'L', 'L', 'L', 'L', 'E']

Probability: 1.97e-08

2.

Forward:

B	1	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0.25	0.1125	0	0	0.001771875	0.000620156	0.000234773	9.46E-05	0	0	3.31E-07	0	0
L	0	0.25	0.1125	0.050625	0.017718750	0	0.000177187	0.000124031	0	9.46E-06	3.31E-06	1.16E-06	4.39E-07	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	**4.39E-08**

Forward $P(X_1|HMM)$: 4.39e-08

Backward:

B	**4.39E-08**	9.75E-08	2.17E-07	6.19E-07	6.19E-06	2.35E-05	5.22E-05	0.000115937	0.001159375	0.003312	0.01125	0.025	0	0
D	0	6.06E-08	8.66E-08	2.48E-07	8.66E-06	2.48E-05	6.14E-05	0.000162312	0.00046375	0.001325	0.007	0.01	0.1	0
L	0	1.15E-07	3.03E-07	8.66E-07	2.48E-06	1.75E-05	3.25E-05	4.64E-05	0.001623125	0.004637	0.01325	0.035	0.1	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Backward $P(X_1|HMM)$: 4.39e-08

3.

Final transition A

	B	L	D	E
B	0	0.65	0.35	0
L	0	0.70	0.17	0.13
D	0	0.38	0.62	0
E	0	0	0	0

Final emission E

	H	P	C
L	0.66	0	0.34
D	0	0.46	0.54

Abbas Rizvi
Assignment 4 – Hidden Markov Models

Probability P1(X_1|HMM): 1.69016010535e-07
sumLL 24.4425645775

4a.

Prior transition A1					Final transition A1				
	B	**L**	**D**	**E**		**B**	**L**	**D**	**E**
B	0	0.5	0.5	0	**B**	0	0.75	0.25	0
L	0	0.7	0.2	0.1	**L**	0	0.62	0.34	0.04
D	0	0.2	0.7	0.1	**D**	0	0.49	0.48	0.03
E	0	0	0	0	**E**	0	0	0	0

Prior emissions E1				Final emission E1			
	H	**P**	**C**		**H**	**P**	**C**
L	0.5	0	0.5	**L**	0.97	0	0.03
D	0	0.5	0.5	**D**	0	0.5	0.5

These are not biologically relevant because the priors have linker regions with greater hydrophobicity than the domains. Domains transition begin states should be larger than the linker begin states. The sequence has converged at iteration 123 with a log likelihood of: -9430.70. The convergence occurs at a local minimum. The final transition matrix shows that there is a higher probability that the set of sequences are a linker. It does not make much sense for there to be more linkers than domains, as linker connects two domains.

4b.

Prior transition A2					Final transition A2				
	B	**L**	**D**	**E**		**B**	**L**	**D**	**E**
B	0	0.3	0.7	0	**B**	0	0.02	0.98	0
L	0	0.8	0.15	0.05	**L**	0	0.68	0.31	0.01
D	0	0.15	0.8	0.05	**D**	0	0.12	0.84	0.04
E	0	0	0	0	**E**	0	0	0	0

Prior emission E2				Final emission E2			
	H	**P**	**C**		**H**	**P**	**C**
L	0.1	0.3	0.6	**L**	0.18	0.38	0.44
D	0.7	0.2	0.1	**D**	0.72	0.14	0.14

A2 and E2 prior probabilities parameters were adjusted to a more biological sensible. Biologically sensible parameters inherently have more information directing convergence to a more optimal result. The prior probability parameters transition A2 were altered such that the begin state for the domain was larger than the linker. Also, for a linker transition to a linker and a domain to transition into a domain were given a higher probability than the two parameters transitioning into each other. The final transition A2 matrix showed that the begin state has approximately a 98% of transitioning to a domain.

Abbas Rizvi
Assignment 4 – Hidden Markov Models

A domain transitioning into a domain had a higher probability than a domain transitioning to a linker. A domain to domain also had a greater probability for going to the end state. This means that the sequences are most likely going to be a domain and have a higher probability that its a domain at their end state.

In E2 prior emission probability parameters were altered for E2 such that the domain parameter was given a larger probability for hydrophobic amino acid regions. Also, the linker parameter was given greater probabilities for polar and charged amino acid regions. The final E2 emission probabilities suggest that the set of sequences is most likely a domain with higher probability of hydrophobic amino acids.

The final trained set of sequences converged at of A2 and E2 iteration 24 with a log likelihood of: -9398.24. The convergence at the local minimum is much quicker than a biologically non-sensible emission and transition probability priors in the previous question. The log likelihood is a maximum likelihood estimator and in A2 and E2 it is closer to 0 when compared to the random prior probabilities in A1 and E1. This suggests that the altered, more biological prior trained sequences have a state path that is closer to the global minimum.

4c.

Prior transition A2					**Final transition A2**				
	B	**L**	**D**	**E**		**B**	**L**	**D**	**E**
B	0	0.3	0.7	0	**B**	0	0.98	0.02	0
L	0	0.8	0.15	0.05	**L**	0	0.84	0.12	0.04
D	0	0.15	0.8	0.05	**D**	0	0.12	0.68	0.01
E	0	0	0	0	**E**	0	0	0	0

Prior emission E3				**Final emission E2**			
	H	**P**	**C**		**H**	**P**	**C**
L	0.7	0.2	0.1	L	0.72	0.14	0.14
D	0.1	0.3	0.6	D	0.19	0.43	0.38

The prior A2 emission (L and D) probabilities were swapped for A3 prior emission probabilities. The A2 transition priors were kept the same as A2. The sequence has converged at iteration 26 with a log likelihood of: -9398.27. This log likelihood is very similar to the output produced from the A2 matrices and in fact produces a matrix that mirrors the A2 matrix. This is also closer to the global minimum than random priors.

4d.

Prior transition A2					**Final transition A4**				
	B	**L**	**D**	**E**		**B**	**L**	**D**	**E**
B	0	0.3	0.7	0	**B**	0	0.02	0.98	0
L	0	0.8	0.15	0.05	**L**	0	0.69	0.30	0.01
D	0	0.15	0.8	0.05	**D**	0	0.13	0.82	0.05
E	0	0	0	0	**E**	0	0	0	0

Prior emission E4				**Final emission E4**			
	H	**P**	**C**		**H**	**P**	**C**
L	0.2	0.2	0.4	L	0.22	0.37	0.41
D	0.2	0.2	0.4	D	0.73	0.14	0.13

The A2 transition prior probabilities and equal prior emission probabilities (E4) were used kept for this run. The sequence has converged at iteration 141 with a log likelihood of: -9398.6. Convergence took quite a bit longer than the previous runs. The log likelihood was similar to the previous two runs. When the prior emissions are all equal, the trained set of sequences shows that there is a greater probability that is it a domain with hydrophobic amino acids. These parameters may suggest that the A2 conclusions were that if the set of sequences were domains, the domains have a greater probability of having hydrophobic amino acids. This information could be used to set up priors for a next optimization run.

5.
The Baum Welch algorithm should be used for a small set of training sequences with state annotation should be used in order to yield an accurate result of how the set of sequences behaves. These "learned" behaviors could be implemented as prior parameter probabilities. Heuristically using these prior probabilities with the Baum Welch algorithm for the large set of training sequences without state annotation should then quickly converge and yield relatively optimal results on how well the set of sequences fits the model.

If there is not enough time, the Viterbi algorithm can be used, as it is a quick, but not as accurate way to calculate the probabilities. The results could still provide useful information that can allow one to learn more about their model.