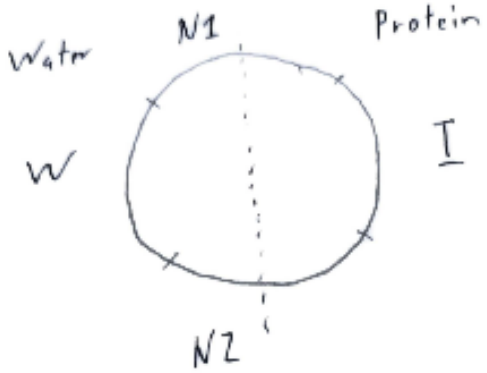
CBMF 4761 Homework 4

Alex Ying

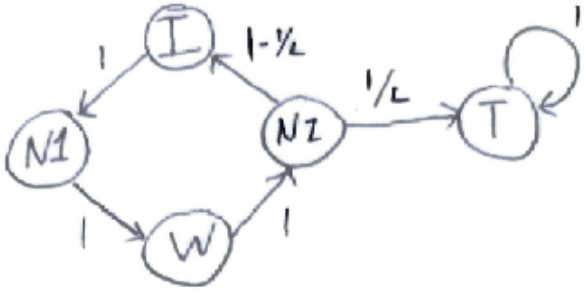
I discussed this problem set with Jason Mohabir.

Problem 1

a. A hidden Markov model that describes the given alpha helix can be constructed using 5 hidden states. These states are water side (W), inner side (I), interface 1 (N1), interface 2 (N2), and the termination state (T). It should be noted that although the exact alignment of the starting amino acid is random within the inner side, the set of these 5 states is invariant to that starting position. This can be proven by noting that, given a circle divided into 4 arcs of equal length, every possible line through the circle's center must pass through two opposite arcs. These two arcs can be labeled as the two interface sectors, N1 and N2, and the other two arcs can be labeled W and I. This construction can also be used to prove that the transition between these states is also invariant to starting position. Every point of an 1 of these arcs corresponds with exactly 2 points 90° away, 1 in each adjacent arc, and these arcs correspond to the connected nodes of the Markov chain.



The probability of transitioning to the termination state is given 0 for all other states except for N2, as the alpha helix is known to complete each rotation. The probability of the chain terminating is known to be $1/L$, as the expected number of rotations is L and the number of rotations is distributed geometrically. This gives graph representing the Markov model:



This gives the transition matrix:

	I	N1	W	N2	T
I	0	1	0	0	0
N1	0	0	1	0	0
W	0	0	0	1	0
N2	$1 - 1/L$	0	0	0	$1/L$
T	0	0	0	0	1

The emissions of the states are given by the following equations (a matrix doesn't fit):

$$P(X|S = W) = W_X$$

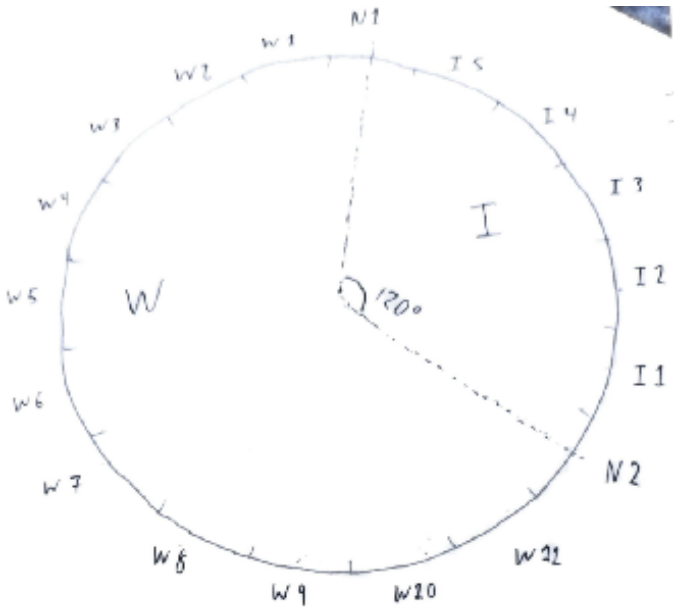
$$P(X|S = I) = I_X$$

$$P(X|S = N1) = P(X|S = N2) = 1/20$$

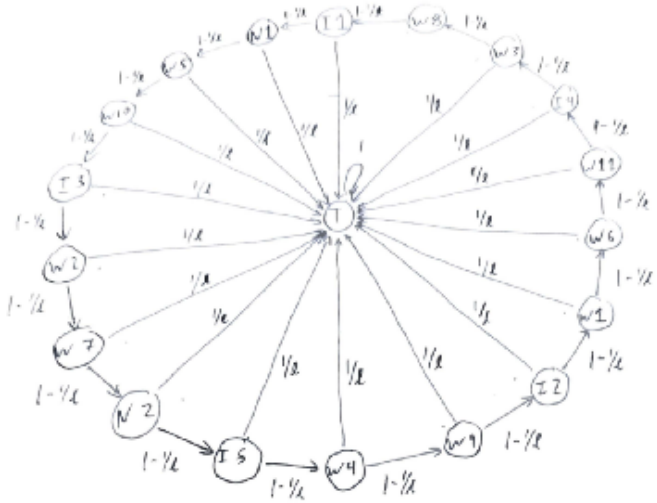
$$P(X|S = T) = 0$$

where X is one of 20 amino acids and S is the state.

b. A hidden Markov model that describes the given alpha helix can be constructed using 19 hidden states. The number of states is derived from the number of 100° steps required to return to one's original position along a circle, which is 18 (the first integer multiple of $\frac{360}{100} = 3.6$). The 19th state is the termination state (T). Of the 18 amino acid states, 11 correspond with the water side (W1-11), 5 correspond with the inner side (I1-5), and 2 correspond with the interface (N1 and N2). Again, the set of these 19 states can be proven to be invariant. Note that for circle divided into 18 arcs of 20°, two radii that have an inner angle of 120° will always intersect 2 such arcs with 5 arcs in between. The 2 arcs intersecting the raddii can be labeled N1 and N2, the 5 arcs between them can be labeled I1-5, the rest of the arcs can be labeled W1-11. This construction can also be used to prove that the transitions between these states are also invariant to starting position. Every point within an arbitrary arc of this circle corresponds with exactly 2 points 100° away, each point belonging to a single separate arc, each of which corresponds to the connected nodes of the Markov chain.



Because the length of the alpha helix is geometrically distributed and the average length is l, the probability of transitioning to the termination state is $1/l$ for all other states. This gives the graph representing the Markov model:



The transition matrix is given by:

	I1	N1	W5	W10	I3	W2	W7	N2	I5	W4	W9	I2	W1	W6	W11	I4	W3	W8	T
I1	0	1-1/l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/l
N1	0	0	1-1/l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/l
W5	0	0	0	1-1/l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/l
W10	0	0	0	0	1-1/l	0	0	0	0	0	0	0	0	0	0	0	0	0	1/l
I3	0	0	0	0	0	1-1/l	0	0	0	0	0	0	0	0	0	0	0	0	1/l
W2	0	0	0	0	0	0	1-1/l	0	0	0	0	0	0	0	0	0	0	0	1/l
W7	0	0	0	0	0	0	0	1-1/l	0	0	0	0	0	0	0	0	0	0	1/l
N2	0	0	0	0	0	0	0	0	1-1/l	0	0	0	0	0	0	0	0	0	1/l
I5	0	0	0	0	0	0	0	0	0	1-1/l	0	0	0	0	0	0	0	0	1/l
W4	0	0	0	0	0	0	0	0	0	0	1-1/l	0	0	0	0	0	0	0	1/l
W9	0	0	0	0	0	0	0	0	0	0	0	1-1/l	0	0	0	0	0	0	1/l
I2	0	0	0	0	0	0	0	0	0	0	0	0	1-1/l	0	0	0	0	0	1/l
W1	0	0	0	0	0	0	0	0	0	0	0	0	0	1-1/l	0	0	0	0	1/l
W6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1-1/l	0	0	0	1/l
W11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1-1/l	0	0	1/l
I4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1-1/l	0	1/l
W3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1-1/l	1/l
W8	1-1/l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/l
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/l

The emissions of the states are given by the following equations:

$$P(X|S \in \{W1, W2, W3, W4, W5, W6, W7, W8, W9, W10, W11\}) = W_X$$

$$P(X|S \in \{I1, I2, I3, I4, I5\}) = I_X$$

$$P(X|S \in \{N1, N2\}) = 1/20$$

$$P(X|S = T) = 0$$

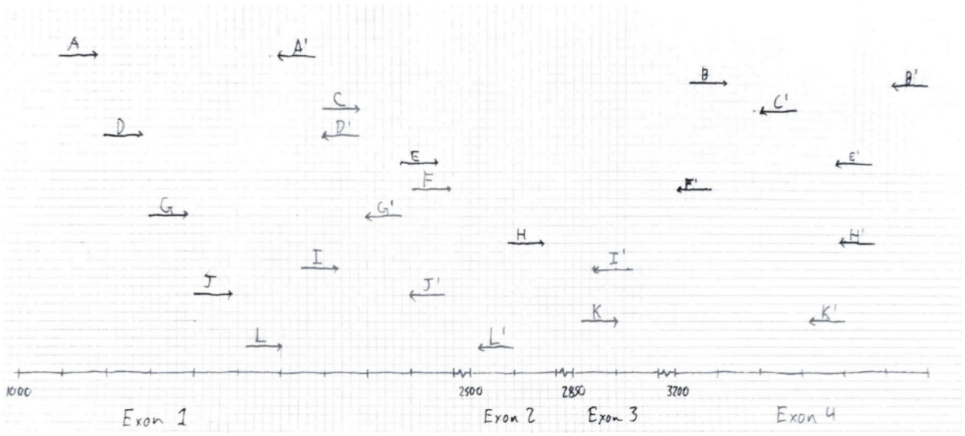
where X is one of 20 amino acids and S is the state.

Problem 2

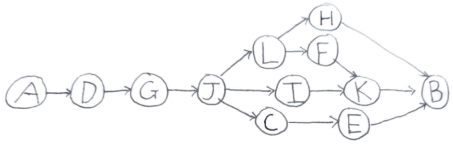
a. Because the length of each fragment is known, reads that span multiple exons can be used to determine which exons are spliced out or have not been spliced out between them. The read pairs are given labels and the exons that are known to be present or absent are listed below:

```
A: ([78001100 -> 78001180], [78001600 <- 78001680]), {1}
B: ([78003210 -> 78003290], [78003710 <- 78003790]), {4}
C: ([78001700 -> 78001780], [78003400 <- 78003480]), {1, !2, !3, 4}
D: ([78001200 -> 78001280], [78001700 <- 78001780]), {1}
E: ([78001880 -> 78001960], [78003580 <- 78003660]), {1, !2, !3, 4}
F: ([78001905 -> 78001985], [78003205 <- 78003285]), {1, 2, 3, 4}
G: ([78001300 -> 78001380], [78001800 <- 78001880]), {1}
H: ([78002590 -> 78002670], [78003590 <- 78003670]), {2, !3, 4}
I: ([78001550 -> 78001630], [78002900 <- 78002980]), {1, !2, 3}
J: ([78001400 -> 78001480], [78001900 <- 78001980]), {1}
K: ([78002870 -> 78002950], [78003520 <- 78003600]), {3, 4}
L: ([78001520 -> 78001600], [78002520 <- 78002600]), {1, 2}
```

The read mapping is can be seen as:



b. The following incompatible pairs are formed: (C, F), (C, H), (C, I), (C, K), (C, L), (E, F), (E, H), (E, I), (E, K), (E, L), (F, H), (F, I), (H, I), and (H, K). This forms the following graph:



Where compatible pairs are indicated by the existence of a path between the pairs' nodes in the graph, and the direction of these edges indicating left to right.

c. The maximum size anti-chain is 4. The set of transcripts that explains the data is {Exons 1+2+3+4, Exons 1+2+4, Exons 1+3+4, Exons 1+4}.

d. Read pairs A, B, D, G, J, K, and L are ambiguous. Read pairs C and E are known to come from the transcript of exons 1+4. Read pairs F is known to come from the transcript of exons 1+2+3+4. Read pair I is known to come from the transcript of exons 1+3+4. Read pair H is known to come from the transcript of exons 1+2+4.

e. The number of transcripts can be reduced if the end read of read pair I was changed to [78002550 <- 78002630], which would make read pair I known to come from Exons 1 and 2. This would reduce the minimal set of transcripts that explain the data to {Exons 1+2+3+4, Exons 1+2+4, Exons 1+4}.