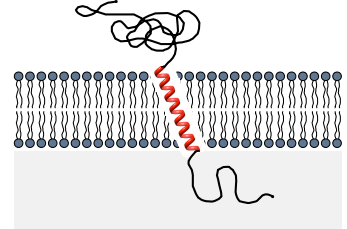# Question 3

A transmembrane protein is a protein that has one or more domains going from one side of the plasmic membrane through to the other side. As a consequence, they have extracellular, intracellular and transmembrane domains. Very often the transmembrane domains are structures known as "alpha-helices". We want to find them using a Hidden Markov Model.

The propensity of different amino-acids to be part of an alpha-helix is well-known and scored in the table below. For this exercise we assume that:

- All transmembrane domains are alpha-helices of appropriate length: in average 20 amino-acids.

- In other domains of a protein, amino-acids are equally distributed.

- Intracellular domains are in average 200 amino-acids long and extracellular domains, 500.

Table 1: Helical propensity of amino-acids, as the fraction of each type found in helices.

A	R	N	D	C	E	Q	G	H	I
8%	7%	4%	4%	4%	6%	6%	3%	5%	6%

L	K	M	F	P	S	T	W	Y	V
7%	6%	7%	5%	**0.3%**	5%	4%	5%	5%	5%

1. Propose the structure of a simple Hidden Markov Model to predict transmembrane domains from a given amino-acids sequence. Draw a diagram.

2. What are the emission probabilities?

3. Based on the length of the respective domains, give an estimate of the transition probabilities. Write the transition matrix.

4. Here is the sequence of a dummy transmembrane protein.

   NGAKTTL

   How woud you calculate the probability of observing this sequence, given the model? (Do not calculate it).

5. Here is the sequence of a dummy transmembrane protein, where the transmembrane domain is known and underlined.

   NGA<u>KT</u>TL

   What is the probability of observing this sequence?

6. *Bonus:* how would you adapt the model (only the diagram) if moreover you know that an alpha-helix is defined as hydrogen bounds linking groups of 4 amino-acids?