1. **Transition Probability Matrix:**

Soluble Transmembrane

Soluble 0.98463 0.01275

Transmembrane 0.04733 0.95241

**Emission Matrix Probability:**

A C D E F G H

Soluble 0.05355 0.01211 0.05814 0.06335 0.04650 0.05124 0.02269

Transmembrane 0.08369 0.02655 0.00911 0.00954 0.10724 0.07635 0.00772

I K L M N P Q

Soluble 0.05864 0.07149 0.08894 0.02125 0.05939 0.04329 0.03707

Transmembrane 0.12875 0.00921 0.15801 0.03163 0.02093 0.02884 0.01394

R S T V W Y

Soluble 0.04996 0.09558 0.06138 0.05255 0.01366 0.03923

Transmembrane 0.00594 0.06796 0.05234 0.09824 0.02257 0.04144

**Viterbi Matrix:**

K N S F F

Soluble 0.070443 0.004119 0.000388 1.774977e-05 8.126784e-07

Transmembrane 0.000135 0.000019 0.000004 5.300706e-07 5.413953e-08

F F F F

Soluble 3.720872e-08 1.703612e-09 7.800039e-11 3.571271e-12

Transmembrane 5.529619e-09 5.647756e-10 5.768418e-11 5.891657e-12

F L I I

Soluble 1.635117e-13 1.431921e-14 8.267725e-16 4.773677e-17

Transmembrane 6.017529e-13 9.055798e-14 1.110447e-14 1.361661e-15

I

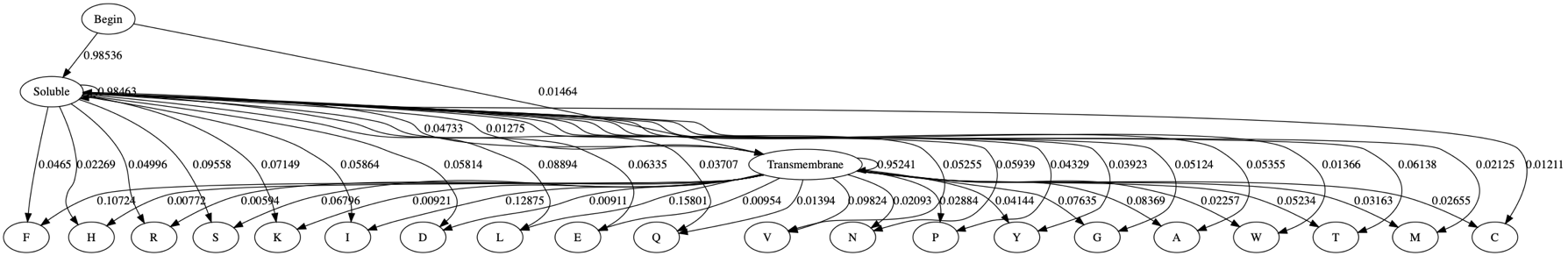
Soluble 3.779197e-18

Transmembrane 1.669707e-16

**Most likely State Sequence (Viterbi):**

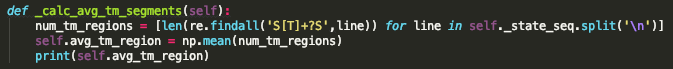
Observed Sequence: KNSFFFFFFFLIII

Viterbi State Prediction: SSSTTTTTTTTTTT

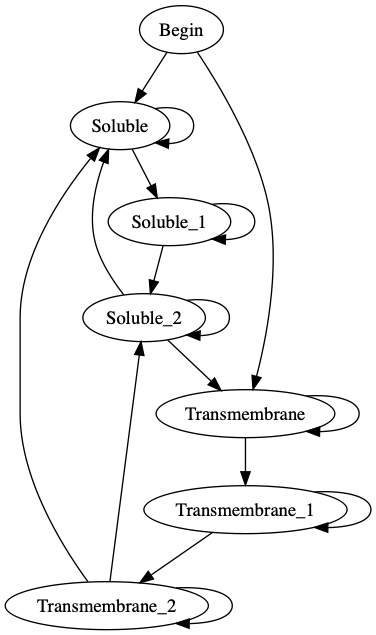


1. Average number of integral membrane segments per protein:

4.643 per protein

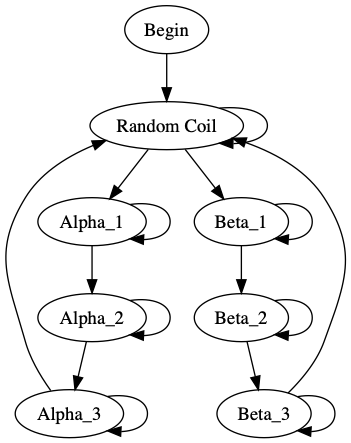
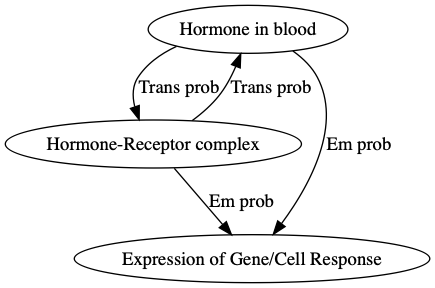


1. In this alternative architecture, it is assumed that each state (besides begin) has 20 observable emissions (to each of the 20 amino acids). They were not included to simplify the drawing.



I the alternative architecture in comparison to the simpler one done in question 1. This architecture incorporates several features that you expect from transitions between soluble and transmembrane regions. For 1, a soluble region and a transmembrane region will have a minimum amount of amino acids before a transition. This is represented by the transitions of soluble to soluble\_2 and transmembrane to transmembrane\_2.

Also, at the end of a transmembrane region (TM\_2) we can jump to a long soluble region or a short soluble region (S\_2). Both of these scenarios are common in in transmembrane proteins and are represented within the architecture.

1. Each state besides the begin state has an emission probability for each amino acid that is not being displayed.
2. An application for a HMM requires a sequence of events or a time series. In biology, all cell signaling and hormonal responses (gene expression/cell response) are time series. The ligand must bind to the receptor before the receptor can be considered activated and commence a phosphorylation cascade. That cascade involves an array of proteins and eventually makes it to the nucleus and initiates transcriptions. This time series could in theory be modelled with HMM where the underlying states are if the hormone is currently bound to the extracellular receptor.