10: Biological Applications for HMMs

Machine Learning and Real-world Data

Ann Copestake and Simone Teufel

Computer Laboratory
University of Cambridge

Lent 2017

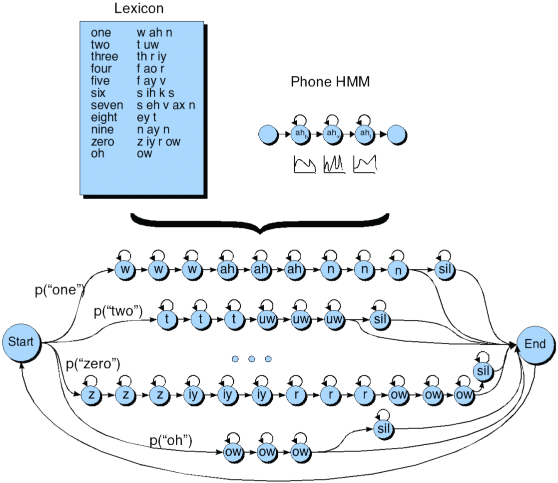
Last session: dice world and HMM decoding

- You now have written a decoder, i.e., an algorithm that can determine the most likely state sequence of an HMM.
- From the task before that, you also have code that can estimate the parameters from a labelled HMM sequence.
- But the dice world is far too simple/artificial to show the potential of HMMs in the real world.

Sequence Learning in the real world

- HMMs for speech recognition
  - Goal: determine from signal which words were said
  - States: words
  - Observations: acoustic inputs from signal
- HMMs for parts of speech tagging
  - Goal: determine the parts of speech for text
  - States: parts of speech
  - Observations: words
- HMM for protein analysis
  - Goal: Find which sections of proteins are in cell membranes
  - States: zones relating to cells
  - Observations: amino acids

HMMs in Automatic Speech Recognition (ASR)



Lexicon

one	w ah n
two	t uw
three	th r iy
four	f ao r
five	f ay v
six	s ih k s
seven	s eh v ax n
eight	ey t
nine	n ay n
zero	z iy r ow
oh	ow

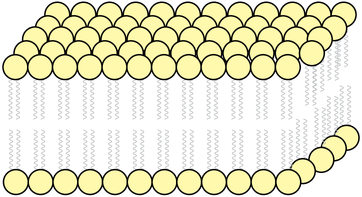
A biological application

```
#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
 iiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooo
```

- top line records the amino acid sequence (one character per amino acid)
- bottom line shows the states:
  - i: inside the cell
  - M: within the cell membrane
  - o: outside the cell

Eight minutes about biology of cells

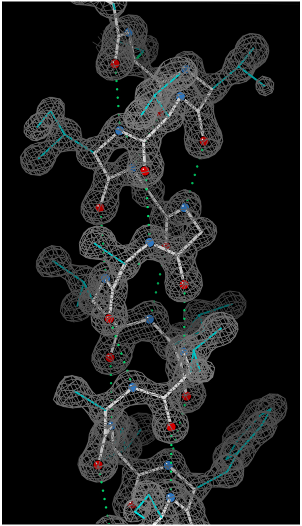
- living organisms are made up of cells
- multicellular organisms have lots of cells
- cells are surrounded by a cell membrane
- cell membranes are lipid bilayers: inside the membrane is hydrophobic (water-hating), the two sides are hydrophilic (water-loving)



Proteins

- in cell metabolism: proteins make sure the right thing happens in the right place at the right time
- proteins are made up of amino acid sequences
- all amino acids have amine and carboxyl groups, but they have very different **side chains**
- 20 amino acids are coded for directly by DNA
- amino acid sequences fold into very complex 3-D protein structure

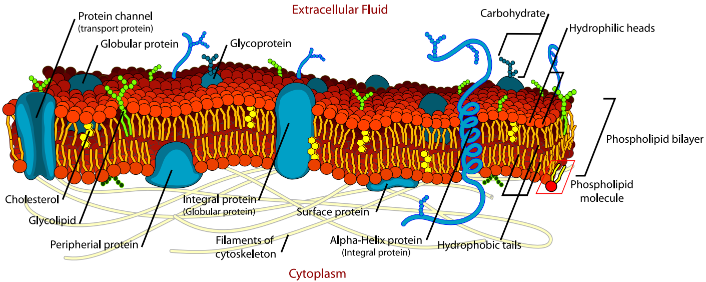
Alpha helix



Dcrjsr - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9131613

Cell membranes and proteins

- cell membranes have to let things in and out of the cell (e.g., water, glucose, sodium ions, calcium ions)
- proteins which are part of the cell membrane allow this (membrane proteins do other things too)

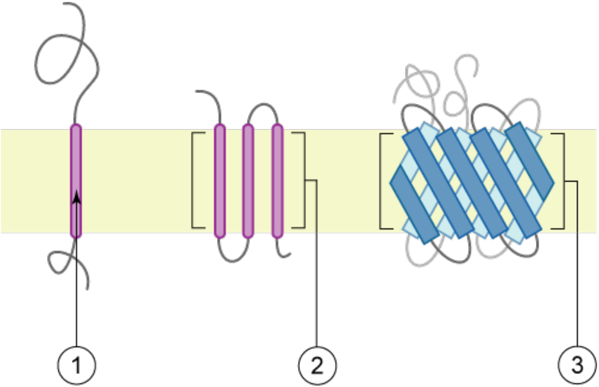


By LadyofHats Mariana Ruiz - Own work. https://commons.wikimedia.org/w/index.php?curid=6027169

Transmembrane proteins

- transmembrane proteins go through the membrane one or more times
- the regions of the protein which lie inside and outside the cell tend to have more hydrophilic amino acids
- the regions inside the membranes tend to have more hydrophobic amino acids
- many transmembrane proteins involve one or more α-helixes in the membrane
- the channels formed by the protein allow ions and molecules through, in a controlled way

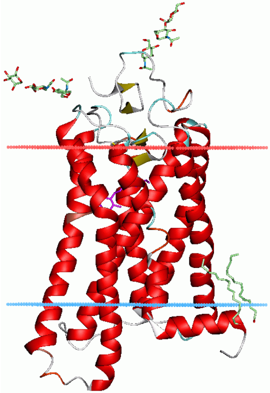
Transmembrane protein: schematic diagram



1. a single transmembrane α-helix (bitopic membrane protein) 2. a polytopic transmembrane α-helical protein 3. a polytopic transmembrane β-sheet protein

By Foobar - self-made by Foobar, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=802476

Transmembrane protein: Bovine rhodopsin



- one of the visual pigments
- accurate structure via x-ray crystallography: difficult and time-consuming, membrane location undetermined

By Andrei Lomize - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=34114850

A biological application

```
#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
 iiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooo
```

- HMM-based modelling: much, much easier and quicker than x-ray crystallography
- distinguish interior of membrane from inside/outside of cell
- simple HMM in practical, but could be improved: more discussion in practical notes

Your Task

Task 9:

- Download the biological dataset and familiarise yourself with it.
- Modify your code so that your HMM parameter estimation from Task 7 and decoder from Task 8 works with this data format.
- Use 10-fold cross validation.
- Evaluate.
- Get Tick 6 today or Friday (Task 8 plus Task 9)