# 10: Biological Applications for HMMs

Machine Learning and Real-world Data (MLRD)

Andreas Vlachos
(based on slides by Ann Copestake and Simone Teufel)

Dept. of Computer Science and Technology
University of Cambridge

Lent 2020

# Last session: dice rolls and HMM decoding

- You may by 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 sequence of observations and (hidden) states.
- But the dice rolls are very simple and somewhat artificial.

# 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

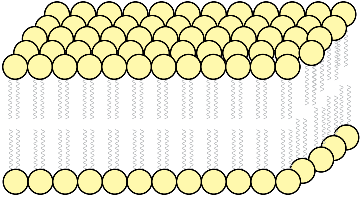
# A biological application: the data

```
#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
 iiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooo
```

- top line records the amino acid sequence representing the protein (one character per amino acid)
- bottom line shows the states:
  - i: inside the cell
  - M: within the cell membrane
  - o: outside the cell
- Ignoring the start and end sequence states/labels for simplicity.

# A few 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)



Jerome Walker - Own work, CC BY 2.5, `http://commons.wikimedia.org/w/index.php?curid=915557`

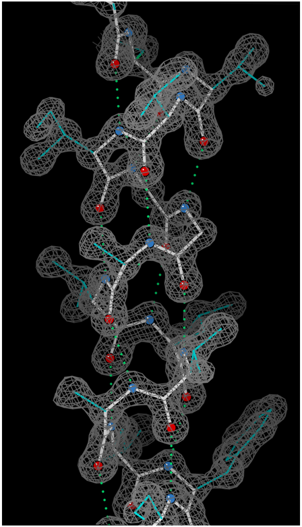
# Proteins

- 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 the same core structure (**amine** and **carboxyl** groups), but they have very different **side chains**
- 20 amino acids are coded for directly by DNA
- as amino acid sequences are constructed in the cell, they fold into very complex 3D protein structure

# 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 the same core structure (**amine** and **carboxyl** groups), but they have very different **side chains**
- 20 amino acids are coded for directly by DNA
- as amino acid sequences are constructed in the cell, they fold into very complex 3D protein structure
- experimental 3D structure determination is very difficult, 3D structure prediction is an important task for machine learning (lecture on Friday).

# Example of protein structure: alpha helix

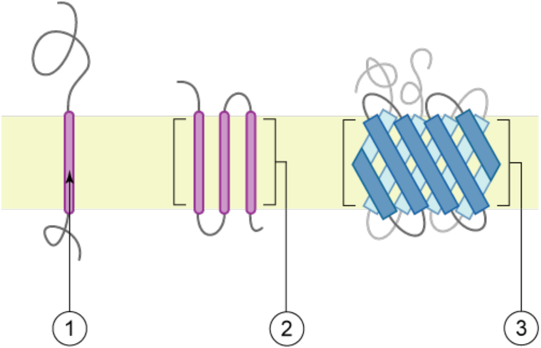


Dcrjsr - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9131613

# Transmembrane proteins

- **transmembrane** proteins go through the cell 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 $\alpha$-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 $\alpha$-helix (bitopic membrane protein)
2. a polytopic transmembrane $\alpha$-helical protein
3. a polytopic transmembrane $\beta$-sheet protein

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

# HMMs for determination of membrane location of proteins

```
#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
 iiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooo
```

- HMM-based modelling: much, much easier and quicker than x-ray crystallography
- distinguish interior of membrane (`M`) from inside(`i`)/outside($\circ$) of cell
- very simple HMM approach in practical, but could be improved: more discussion in practical notes

# HMMs for determination of membrane location of proteins

```
#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
 iiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooo
```

- HMM-based modelling: much, much easier and quicker than x-ray crystallography
- distinguish interior of membrane (`M`) from inside(`i`)/outside(`o`) of cell
- very simple HMM approach in practical, but could be improved: more discussion in practical notes
- think about the properties of the problem that the HMM can model and those it cannot.

# 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.

# Next sessions

- Friday catch-up session: non-examinable mini-lecture on protein structure determination.
- For Task 10 (Monday next week), you will need to download gephi (graph visualization).
  `https://gephi.org/users/download/`
  Please do this in advance of the scheduled session if at all possible.