10: Biological Applications for HMMs

Machine Learning and Real-world Data (MLRD)

Ann Copestake
(partly based on slides created by Simone Teufel)

Lent 2019

Last session: dice world 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 labelled HMM sequence.
- But the dice world is very simple/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

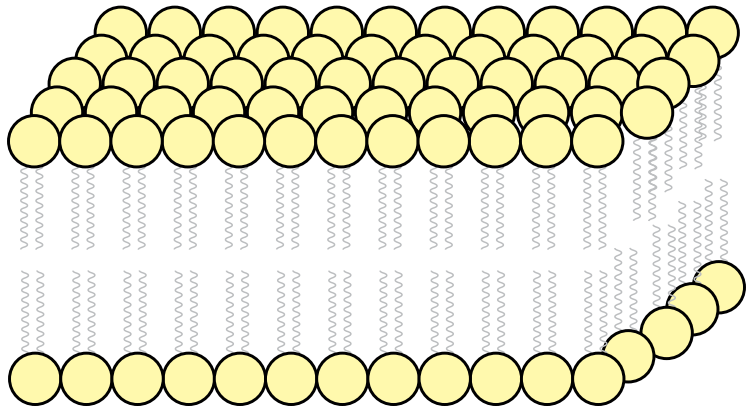
A biological application: the data

```
#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
 iiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooo
```

- 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
- Ignoring the start and end sequence states/labels for simplicity.

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)



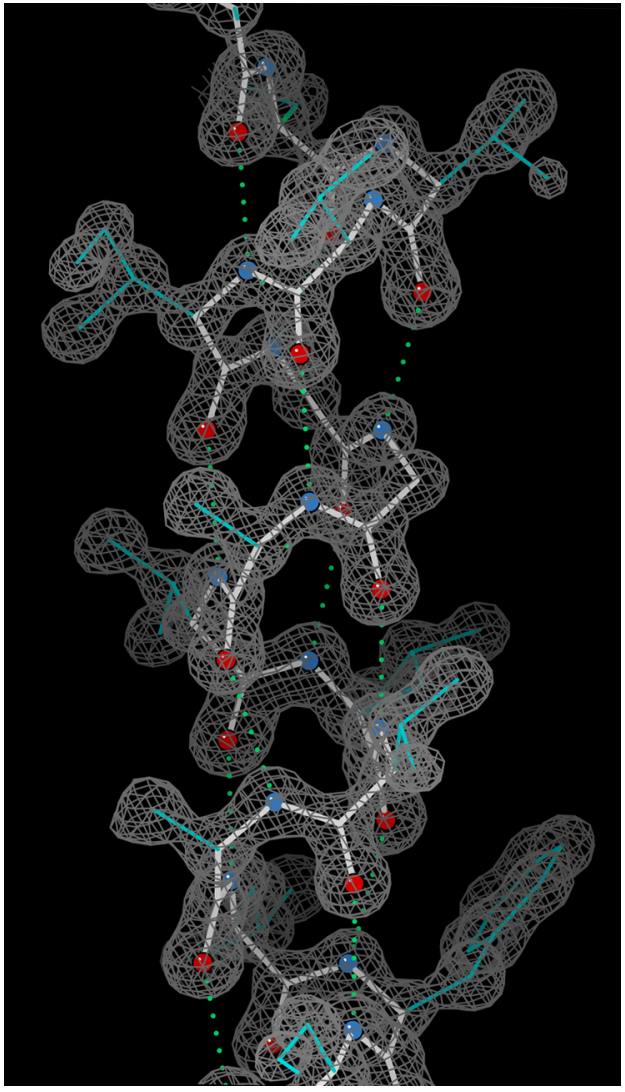
Proteins

- 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 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

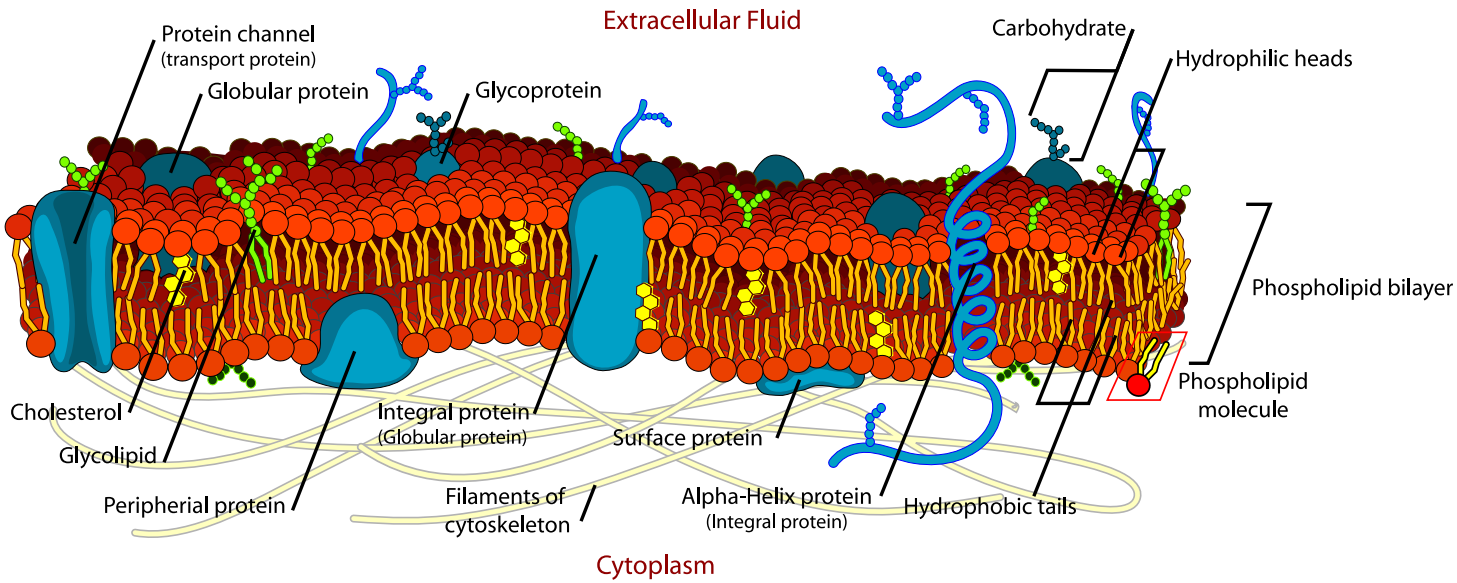
- in cell metabolism: proteins make sure the right thing happens in the right place at the right time
- proteins are made up of **amino acid** sequences
- all amino acids have the same core structure (**amine** and **carboxyl** groups), but they have very different **side chains**
- 20 amino acids are 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).

Alpha helix



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)

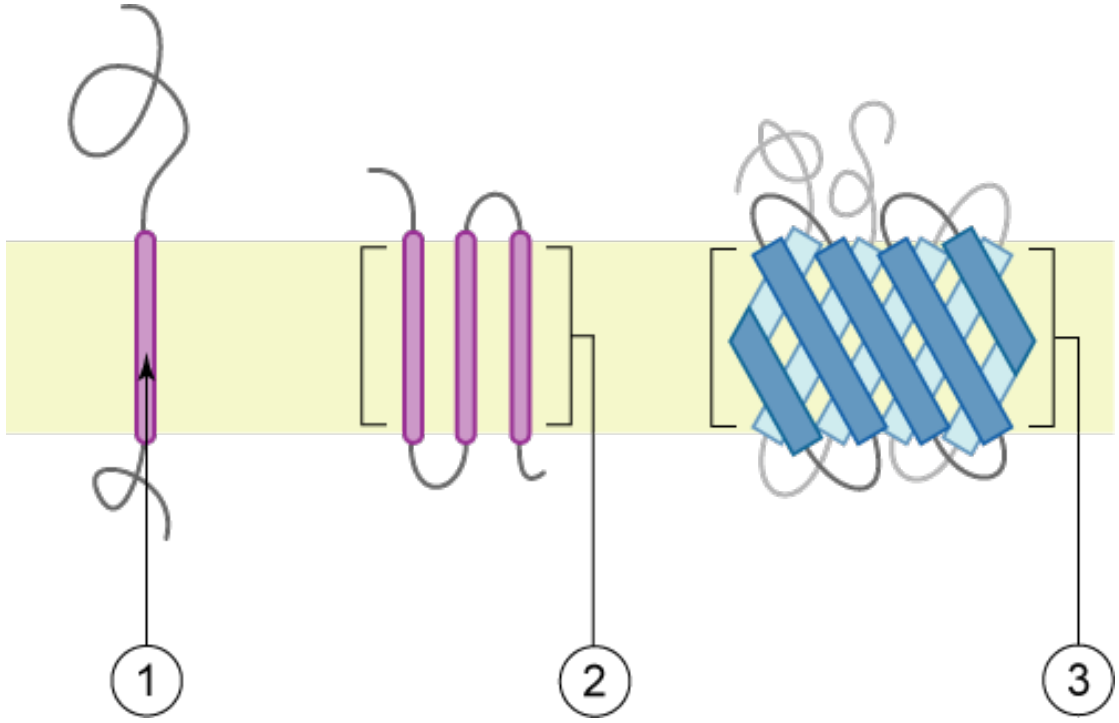


By LadyofHats Mariana Ruiz - Own work. https://commons.wikimedia.org/w/index.php?curid=6027169

Transmembrane proteins

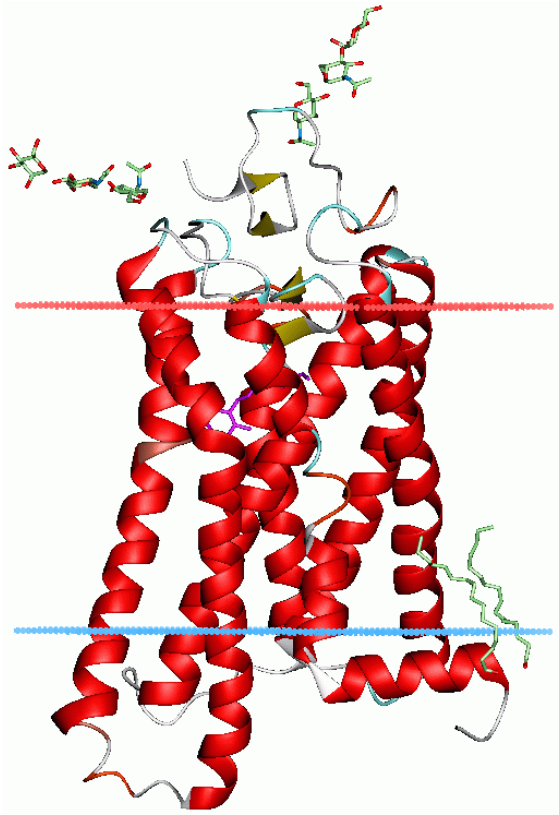
- **transmembrane** proteins go through the cell membrane one or more times
- the regions of the protein which lie inside and outside the cell tend to have more hydrophilic amino acids
- the regions inside the membranes tend to have more hydrophobic amino acids
- many transmembrane proteins involve one or more 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

Transmembrane protein example: (bovine) rhodopsin



- rhodopsin: one of the visual pigments
- accurate structure via x-ray crystallography: difficult and time-consuming, membrane location not determined

HMMs for determination of membrane location of proteins

```
#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
 iiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooo
```

- 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

HMMs for determination of membrane location of proteins

```
#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
 iiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooo
```

- 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.