10: Biological Applications for HMMs
Machine Learning and Real-world Data (MLRD)

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

Department of Computer Science and Technology
University of Cambridge
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.

Let’s look at some sequence learning in the real world.
HMMs for parts of speech tagging

- **Goal**: determine the parts of speech for text
- **States**: parts of speech
- **Observations**: words

![Diagram of HMMs for parts of speech tagging]

Copyright ©2020 Daniel Jurafsky & James H. Martin. From: Speech and Language Processing.
There are many hidden states in POS tagging.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coord. conj.</td>
<td>and, but, or</td>
<td>NNP</td>
<td>proper noun, sing.</td>
<td>IBM</td>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>one, two</td>
<td>NNPS</td>
<td>proper noun, plu.</td>
<td>Carolinas</td>
<td>UH</td>
<td>interjection</td>
<td>ah, oops</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>a, the</td>
<td>NNS</td>
<td>noun, plural</td>
<td>llamas</td>
<td>VB</td>
<td>verb base</td>
<td>eat</td>
</tr>
<tr>
<td>EX</td>
<td>existential ‘there’</td>
<td>there</td>
<td>PDT</td>
<td>predeterminer</td>
<td>all, both</td>
<td>VBD</td>
<td>verb past tense</td>
<td>ate</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>mea culpa</td>
<td>POS</td>
<td>possessive ending</td>
<td>’s</td>
<td>VBG</td>
<td>verb gerund</td>
<td>eating</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/</td>
<td>of, in, by</td>
<td>PRP</td>
<td>personal pronoun</td>
<td>I, you, he</td>
<td>VBN</td>
<td>verb past participle</td>
<td>eaten</td>
</tr>
<tr>
<td></td>
<td>subordin-conj</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>yellow</td>
<td>PRP$</td>
<td>possess. pronoun</td>
<td>your, one’s</td>
<td>VBP</td>
<td>verb non-3sg-pr</td>
<td>eat</td>
</tr>
<tr>
<td>JJR</td>
<td>comparative adj</td>
<td>bigger</td>
<td>RB</td>
<td>adverb</td>
<td>quickly</td>
<td>VBZ</td>
<td>verb 3sg pres</td>
<td>eats</td>
</tr>
<tr>
<td>JJS</td>
<td>superlative adj</td>
<td>wildest</td>
<td>RBR</td>
<td>comparative adv</td>
<td>faster</td>
<td>WDT</td>
<td>wh-determ.</td>
<td>which, that</td>
</tr>
<tr>
<td>LS</td>
<td>list item marker</td>
<td>1, 2, One</td>
<td>RBS</td>
<td>superlat. adv</td>
<td>fastest</td>
<td>WP</td>
<td>wh-pronoun</td>
<td>what, who</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td>can, should</td>
<td>RP</td>
<td>particle</td>
<td>up, off</td>
<td>WP$</td>
<td>wh-possess.</td>
<td>whose</td>
</tr>
<tr>
<td>NN</td>
<td>sing or mass noun</td>
<td>llama</td>
<td>SYM</td>
<td>symbol</td>
<td>+,%, &amp;</td>
<td>WRB</td>
<td>wh-adverb</td>
<td>how, where</td>
</tr>
</tbody>
</table>

**Figure 8.2** Penn Treebank part-of-speech tags.

Copyright ©2020 Daniel Jurafsky & James H. Martin. From: Speech and Language Processing.
HMMs in Automatic Speech Recognition (ASR)

- **Goal**: determine from signal which words were said
- **States**: words
- **Observations**: phones (classified by acoustic classifier from acoustic inputs in signal)
A biological application: Protein analysis

- **Goal:** Find which sections of proteins are in cell membranes
- **States:** zones relating to cells
- **Observations:** amino acids
Transmembrane Protein analysis

#MNQGKIWTVVNPAYIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA

- 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)

Source: Jerome Walker - Own work, CC BY 2.5, https://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
- 20 amino acids are coded for directly by DNA
- amino acid sequences fold into very complex 3-D protein structure
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 channels formed by the protein allow ions and molecules through, in a controlled way.
- 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.
Types of amino acids

- All amino acids have one amine (NH$_2$) and one carboxyl (COOH) group.
- They also have a sidechain that differs from amino acid to amino acid.
- Properties of sidechain: weak acid, strong base, hydrophilic, hydrophobic.
- If alpha-carbon is adjacent to nitrogen atom, amino acid is called alpha amino acid.

By Techguy78 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=89198282
Peptides

- two or more amino acids can combine to form a peptide (short chains of between 2 and 50 amino acids)
- in peptides, amino acids are connected by a peptide backbone, and what remains of each amino acid is called a residue (the side chain)
- alpha-peptides and beta-peptides have different secondary protein structure
Alpha helix

- Alpha helix is most extreme, most predictable, most prevalent of secondary protein structures.
- Every backbone N-H group hydrogen bonds to the backbone C=O groups of the amino-acid located 3 or 4 residues earlier.
- Inner section is formed by tightly-coiled main chain.
- Side chains extend outwards in helical array.
- In crystallographic electron density image left: O atoms red; N atoms blue; hydrogen bonds as green dotted lines.

An $\alpha$-helix in ultra-high-resolution electron density contours

Dcrjsr - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9131613
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
   (bitopic=single-span, polytopic=multi-span)
Transmembrane protein: Bovine rhodopsin

- one of the visual pigments
- found in the rods of the retina (vertebrates)
- extremely sensitive to light (photobleaching)
- accurate structure via x-ray crystallography: difficult and time-consuming, membrane location undetermined
A biological application

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
- Explore semi-supervised learning via self-training, i.e. using a trained model to annotate unlabelled data which in turn will be used for training
- Use 10-fold cross validation
- Evaluate reporting Precision and Recall