10: Biological Applications for HMMs

Machine Learning and Real-world Data (MLRD)

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

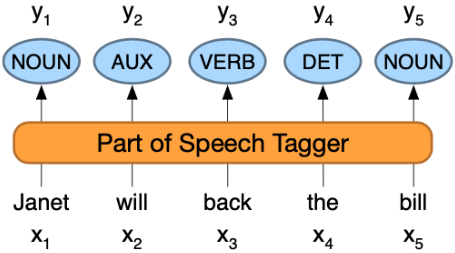
Department of Computer Science and Technology
University of Cambridge

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.
- Let's look at some sequence learning in the real world.

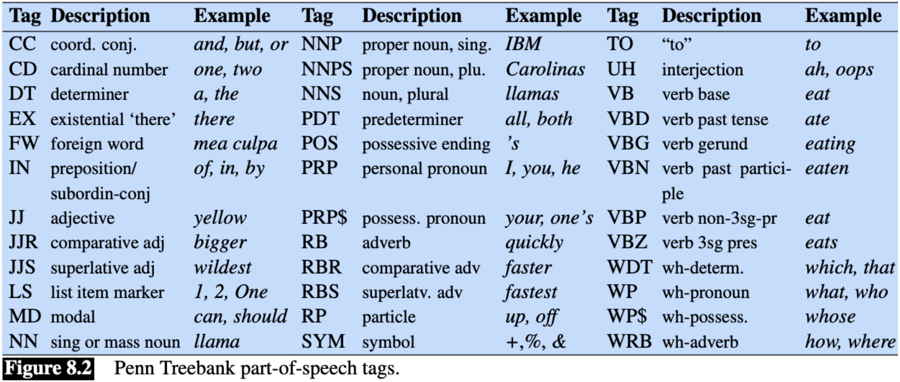
HMMs for parts of speech tagging

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



Copyright ©2020 Daniel Jurafsky & James H. Martin. From: Speech and Language Processing.

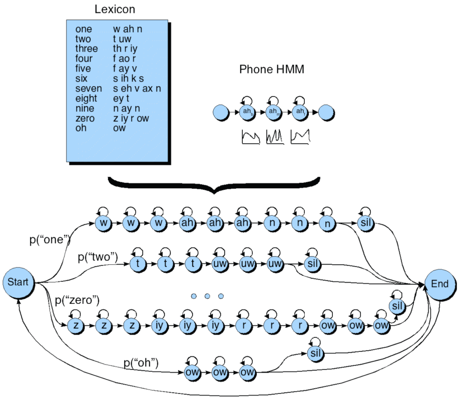
There are many hidden states in POS tagging



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

Figure 8.2 Penn Treebank part-of-speech tags.

HMMs in Automatic Speech Recognition (ASR)



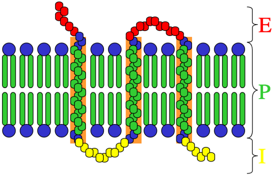
Lexicon

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

- 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



By Mouagip (talk)This W3C-unspecified vector image was created with Adobe Illustrator. - Transmembrane_receptor.png, CC-BY-SA 3.0 https://commons.wikimedia.org/w/index.php?curid=11317884

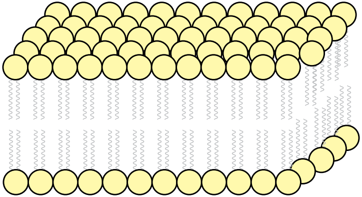
Transmembrane Protein analysis

```
#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
- 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)



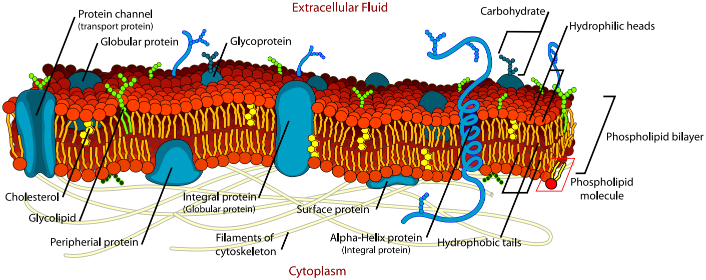
Jerome Walker - Own work, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=915557

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)

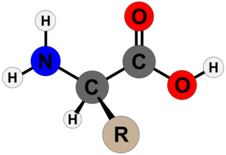


By LadyofHats Mariana Ruiz - Own work. https://commons.wikimedia.org/w/index.php?curid=6027169

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 α-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, hydrophile, hydrophobe
- If alpha-carbon is adjacent to nitrogen atom, amino acid is called alpha amino acid

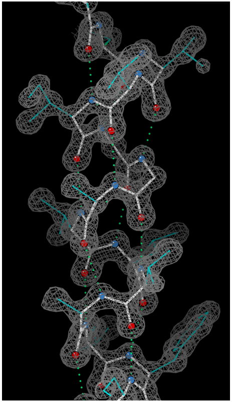
By Techguy78 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=89198282

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

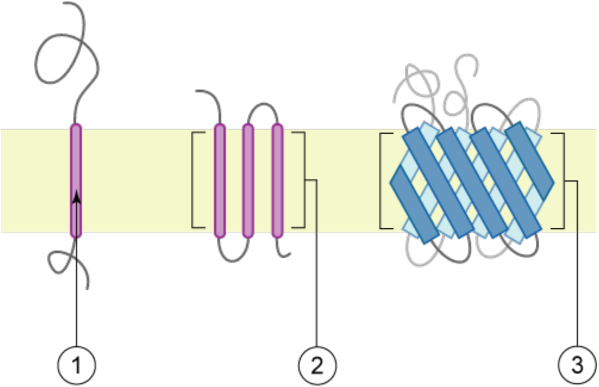
Alpha helix

- 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 α-helix in ultra-high-resolution electron density contours

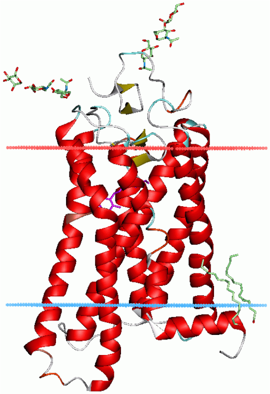
Transmembrane protein: schematic diagram



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

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
- 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

```
#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
- 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