

10: Biological Applications for HMMs

Machine Learning and Real-world Data (MLRD)

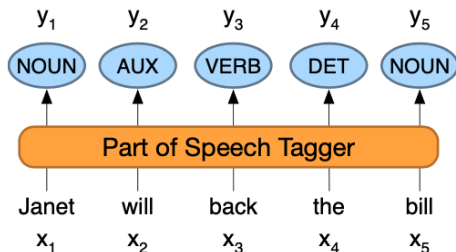
Simone Teufel
(based on slides created by Ann Copestake
and Simone Teufel)

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.

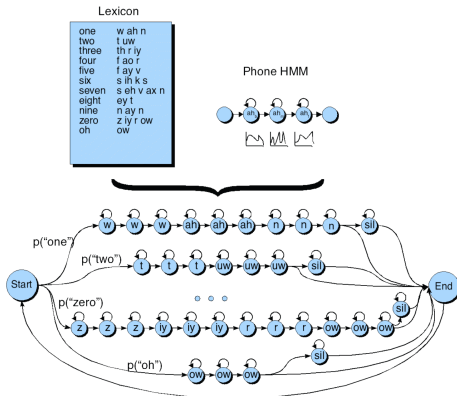
There are many hidden states in POS tagging

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

Figure 8.2 Penn Treebank part-of-speech tags.

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

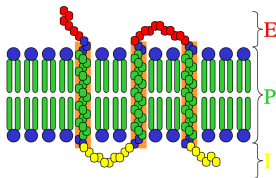
HMMs in Automatic Speech Recognition (ASR)



- 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

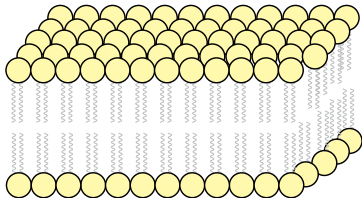
- Goal: Find which sections of proteins are in cell membranes
- States: zones relating to cells
- Observations: amino acids



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>

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)

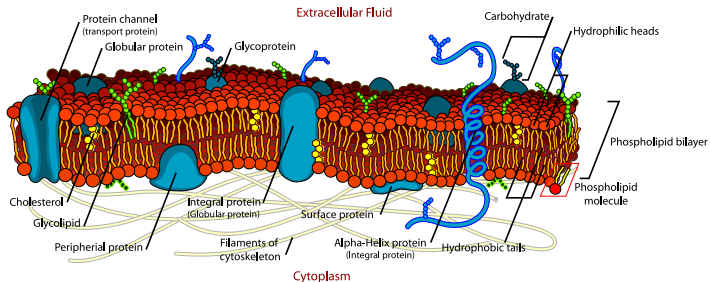


Proteins

- 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
- 20 amino acids are coded for directly by DNA
- amino acid sequences fold into very complex 3-D protein structure

Cell membranes and proteins

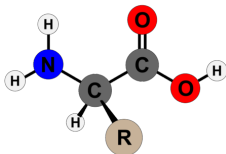
- cell membranes have to let things in and out of the cell (e.g., water, glucose, sodium ions, calcium ions)
- proteins which are part of the cell membrane allow this (membrane proteins do other things too)



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 -helices in the membrane

Types of amino acids



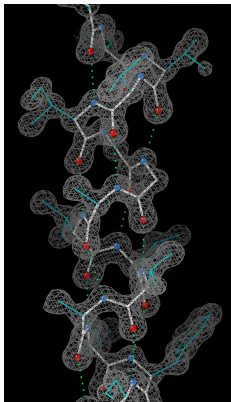
- all amino acids have one amine (NH_2) and one carboxyl (COOH) group
- they also have a sidechain that differs from amino acid to amino acid
- properties of sidechain: weak acid, strong base, hydrophile, hydrophobe
- If alpha-carbon is adjacent to nitrogen atom, amino acid is called alpha amino acid

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

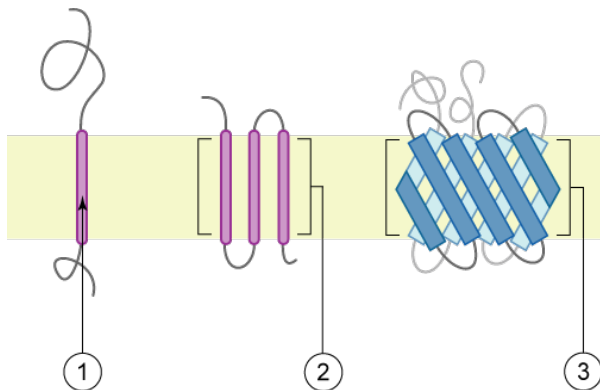


An α -helix in ultra-high-resolution electron density contours

Dcrjsr - Own work, CC BY 3.0,

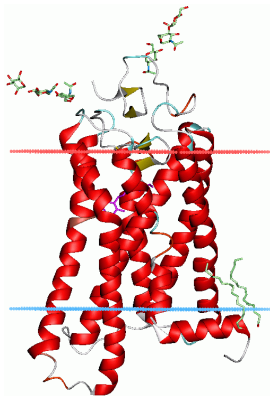
<https://commons.wikimedia.org/w/index.php?curid=1000000>

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)

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

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. Report Precision and Recall.

For Task 10 (next week), you will need to download gephi (graph visualization): please do this in advance of the scheduled session if possible.