# Introduction to Hidden Markov Models Practical 1

In this practical you will learn to build and apply a Hidden Markov Model for finding coding sequences in a bacterial genome.

1. **Download an annotated bacterial genome**

Go to NCBI Genome (<http://www.ncbi.nlm.nih.gov/genome>) and search for an annotated bacterial genome for a species of your choice. If you are looking for inspiration, you may pick from one of the following examples:

|  |  |
| --- | --- |
| *Bacillus anthracis* | Anthrax |
| *Campylobacter jejuni* | Food poisoning |
| *Clostridium botulinum* | Botulism |
| *Clostridium difficile* | Colitis |
| *Clostridium tetani* | Tetanus |
| *Escherichia coli* | Gastroenteritis, UTI |
| *Haemophilus influenzae* | Pneumonia (≠ influenza) |
| *Helicobacter pylori* | Gastric ulcers |
| *Legionella pneumophila* | Legionnaire's disease |
| *Listeria monocytogenes* | Food poisoning |
| *Mycobacterium leprae* | Leprosy |
| *Mycobacterium tuberculosis* | Tuberculosis |
| *Neisseria gonorrhoeae* | Gonorrhoea |
| *Neisseria meningitidis* | Meningitis |
| *Salmonella enterica* serovar Typhi | Typhoid |
| *Staphylococcus aureus* | Soft tissue infections |
| *Streptococcus pneumoniae* | Pneumonia |
| *Streptococcus pyogenes* | Scarlet fever |
| *Treponema pallidum* | Syphilis |
| *Vibrio cholerae* | Cholera |
| *Yersinia pestis* | Plague |

* On the species overview page that the search takes you to, click on "Genome Assembly and Annotation Report".
* Choose a high quality annotated reference genome that has a single scaffold or none listed. Under the "Chromosomes", "RefSeq" column, click the link beginning "NC\_".
* Expand the "Customize View" panel on the right hand side, select the "Customize" button and select "Show sequence". Click "Update View" to make the changes. The page will now show a description of the genome in GenBank format with genes and coding sequences (CDS) annotated and the full genome sequence at the bottom.
* Click "Send" in the top right hand corner, under "Choose Destination" select "File", under "Format" select "**GenBank (full)**" and click "Create File". When the dialog box appears, click "Save File". It will save a file to your default downloads folder called "sequence.gb".
* Click "Send" again, under "Choose Destination" select "File", under "Format" select "**FASTA**" and click "Create File". When the dialog box appears, click "Save File". It will save a file to your default downloads folder called "sequence.fasta".

1. **Data exploration**

Open the sequence (in FASTA format) and annotation (in GenBank format) in R using the functions provided in practical1.R. Code is provided to translate the six possible reading frames. In what follows, you are encouraged to find the solutions yourself but may use the provided code if stuck or running short on time.

1. For the first 15kb, say, of the genome, plot the positions of stop codons in all six reading frames, and compare the positions of annotated genes.
2. What is the frequency of stop codons in the six reading frames?
3. What proportion of the overall genome encodes proteins?
4. **Markov model of stop codon clustering**
5. Focusing on one of the six reading frames in what follows, tabulate the number of transitions from amino acids to stop codons and back.
6. Using this table, estimate the parameters of a Markov model with two states (amino acid and stop codon).
7. What is the mean length of an open reading frame, in kilobases?
8. Write out the transition matrix and stationary distribution of the Markov chain.
9. Simulate a sequence of, e.g. 15 kb, and plot it to compare clustering of stop codons in your simulated data to an equivalent length of sequence in the downloaded genome. How good is the resemblance?
10. **Hidden Markov model for genome annotation**
11. Using the GenBank annotation, tabulate the number of transitions from coding sequences (CDS) to non-coding sequences (NCDS) in one of the reading frames of your downloaded genome. Treat a triplet, not a nucleotide, as an individual site.
12. Use the table to estimate the transition probability matrix of a Hidden Markov Model with two hidden (also called latent) states (CDS and NCDS).
13. What is the mean length of a CDS in your genome, in kilobases?
14. What is the mean length of NCDS in one reading frame of your genome, in kilobases?
15. Now tabulate the number of emissions of an observed state (amino acid versus stop codon) within CDS and NCDS.
16. Use the table to estimate the emission probability matrix of the HMM.
17. What is the frequency of stop codons in CDS and NCDS respectively?
18. Simulate a sequence using the HMM and estimated parameters, and plot the distribution of stop codons to compare it with the downloaded genome. What qualitative differences are there to the Markov model and why?
19. **Detecting coding sequences in a bacterial genome**
20. Implement the Viterbi algorithm.
    * Test it on your simulated data by tabulating (i) the proportion of triplets inferred to be in CDS and NCDS. Compare this to the stationary distribution and (ii) the proportion of simulated CDS and NCDS correctly inferred by the Viterbi algorithm.
21. Apply your algorithm to all six reading frames of the downloaded genome.
22. What proportion of CDS and NCDS in each reading frame does the algorithm correctly infer?
23. What proportion of CDS and NCDS in the genome as a whole does the algorithm correctly infer?
24. How accurate is the estimation of translation start and end sites? Are there any patterns? Could you improve the model?
25. Compare your results to bacterial species analysed by others in the practical. What are the similarities and differences?