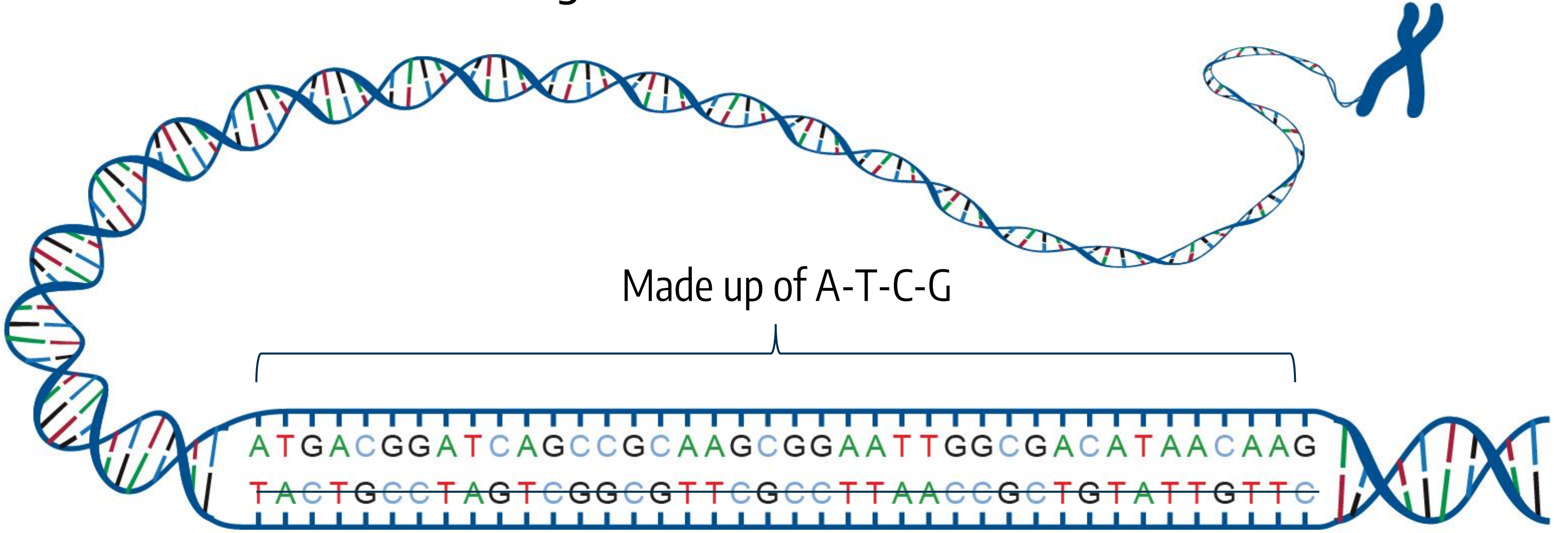
Hidden Markov Model in Gene Detection

By Wei Jun Tan

Biological Background

What is DNA?

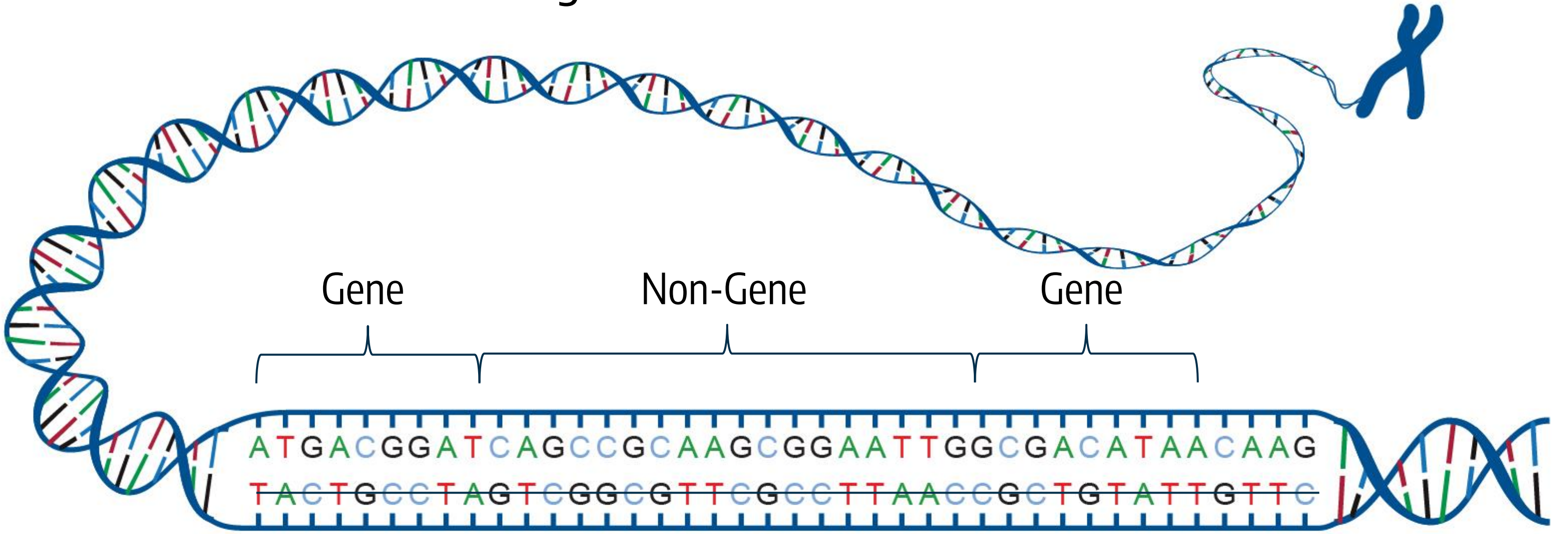
A chemical compound that contains genes, which encodes protein sequence that defines the characteristics of organism


ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAACAAG
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATTGTTC

Made up of A-T-C-G

What is DNA?

A chemical compound that contains genes, which encodes protein sequence that defines the characteristics of organism


Gene
Non-Gene
Gene
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAACAAG
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATTGTTC

Why find gene?

Improve our understanding on gene can help:

- Identify mutations that cause diseases (e.g. cancer)
- Cure chromosomal and genetical diseases (e.g. Down Syndrome)
- ...

Introduction to HMM

Hidden Markov Model (HMM)

- Observation: A, T, C, G
- States (Hidden): Gene, Non-Gene

- Transition Probability

	Gene	Non-Gene
Gene		
Non-Gene		

$$P(state_{i+1} = Gene | state_i = Non - Gene)$$

- Emission Probability

	A	T	C	G
Gene				
Non-Gene				

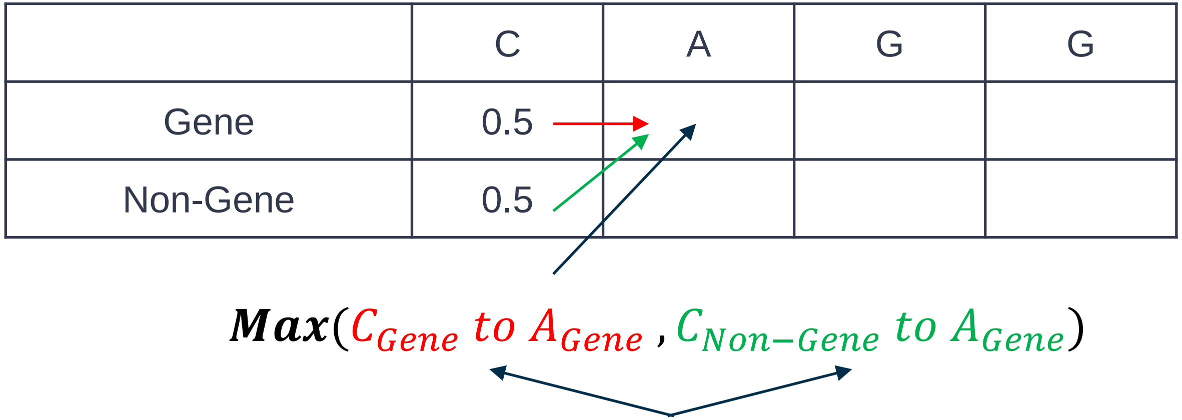
$$P(observe_i = T | state_i = Gene)$$

Viterbi

Q: Given a sequence of DNA (i.e. ATCG) and a HMM (transition and emission matrix), how do we find the gene?

A: Use Viterbi to compute the most probable path of hidden states!

Example

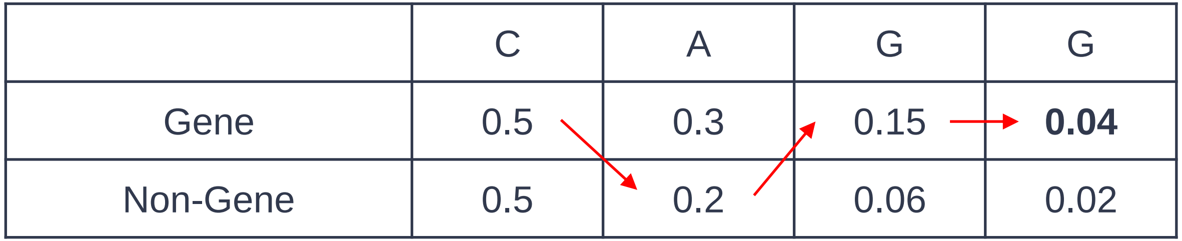


	C	A	G	G
Gene	0.5			
Non-Gene	0.5			

$$\boldsymbol{Max}(C_{Gene}\ to\ A_{Gene}\ , C_{Non-Gene}\ to\ A_{Gene})$$

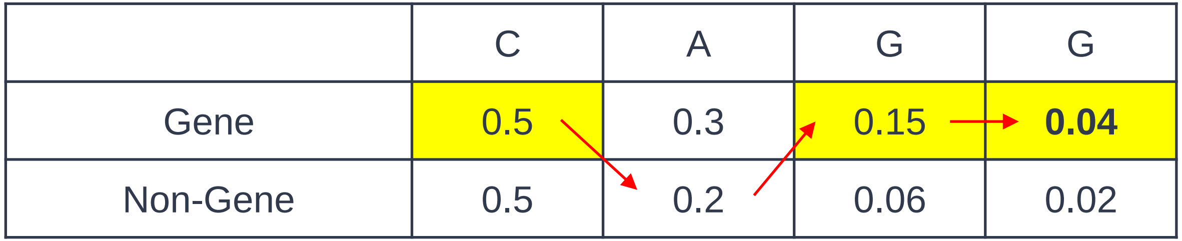
Can be computed using transition and emission matrix!

Example



	C	A	G	G
Gene	0.5	0.3	0.15	**0.04**
Non-Gene	0.5	0.2	0.06	0.02

Example



	C	A	G	G
Gene	0.5	0.3	0.15	**0.04**
Non-Gene	0.5	0.2	0.06	0.02

Two Genes Found!

Use HMM to predict gene

Given an annotated genome

Training

- Use human chromosome 21 (3M+ long) to train our model
- Count the number of transitions and emissions
- Normalize the count to get the transition and emission matrix

	Gene	Non-Gene
Gene	0.999982	0.000037
Non-Gene	0.000018	0.999963

	A	T	C	G
Gene	0.2587	0.2747	0.2354	0.2312
Non-Gene	0.2871	0.2967	0.2112	0.2050

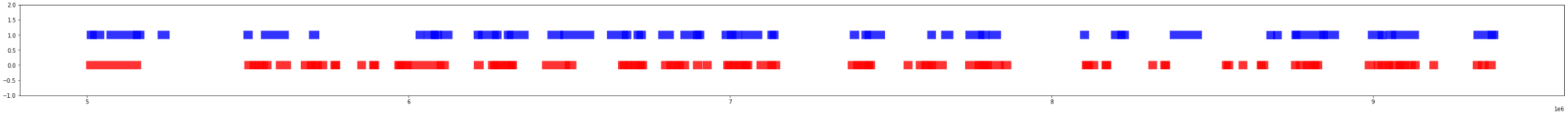
Testing

- Use human chromosome 20 (64M+ long) to test our model
- Use Viterbi to find the predicted path and compare with the actual states

Visualization

Actual Gene V.S. Predicted Gene

Training Set



* Testing Set is not shown as it is too large for readable visualization

Result

Testing Set

- Accuracy: 0.6145
- Precision: 0.6018
- Recall: 0.5479
- Baseline Accuracy: 0.4367

Training Set

- Accuracy: 0.5575
- Precision: 0.4777
- Recall: 0.4015
- Baseline Accuracy: 0.6646

Surprisingly, we perform better in testing set than training set!

Result

Testing Set

- Accuracy: 0.6145
- Precision: 0.6018
- Recall: 0.5479
- Baseline Accuracy: 0.4367

Training Set

- Accuracy: 0.5575
- Precision: 0.4777
- Recall: 0.4015
- Baseline Accuracy: 0.6646

Precision: Among all your predictions, how many of them are correct?
Recall: Among all actual genes, how many of them do you predict?

Result

Testing Set

- Accuracy: 0.6145
- Precision: 0.6018
- Recall: 0.5479
- Baseline Accuracy: 0.4367

Training Set

- Accuracy: 0.5575
- Precision: 0.4777
- Recall: 0.4015
- Baseline Accuracy: 0.6646

Baseline Accuracy: The accuracy of a model that always predict Non-gene (i.e. the frequency of genes)

Improvement

Problem: Underfitting due to strong assumption in the model

- Encode more specific features (dinucleotide, codon, start codon, end codon)
- Encode more specific hidden states (gene types)

Summary

- Use a HMM to detect gene
- Needs improvement
- Please do SHARE your questions!