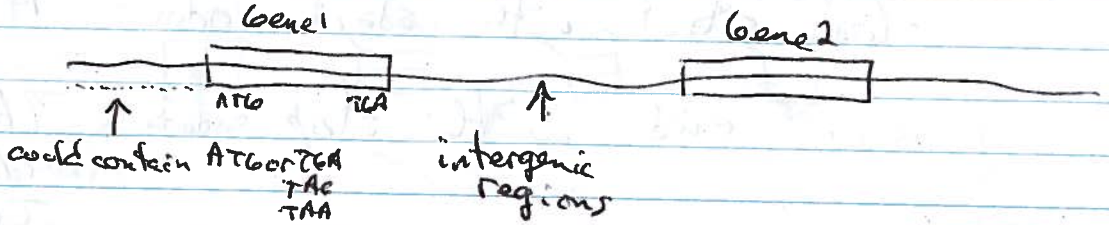
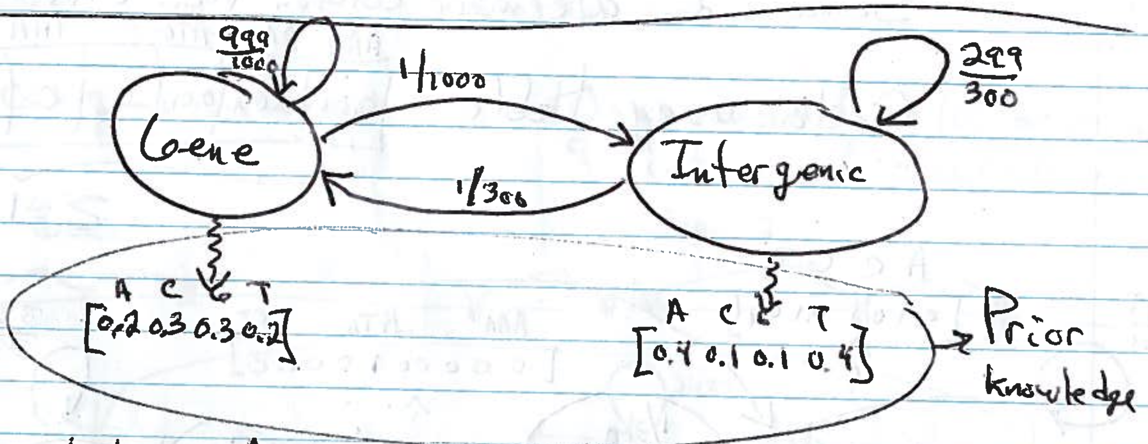
# HMMs for Gene Finding

Bacterial Genome:



Problem: Given: Genome sequence X

Find: Start/End position of each gene in X



Prior knowledge: Avg. gene length ~ 1000bp
Avg. intergenic length ~ 300bp

Genome X = ATAGATAA | CAGGCTCGTGGTC | ATAT

Viterbi Path = IIIIIIII | GGGGGGGGGGGGG | IIII

Integ | Gene | Integ

## Gene Properties:

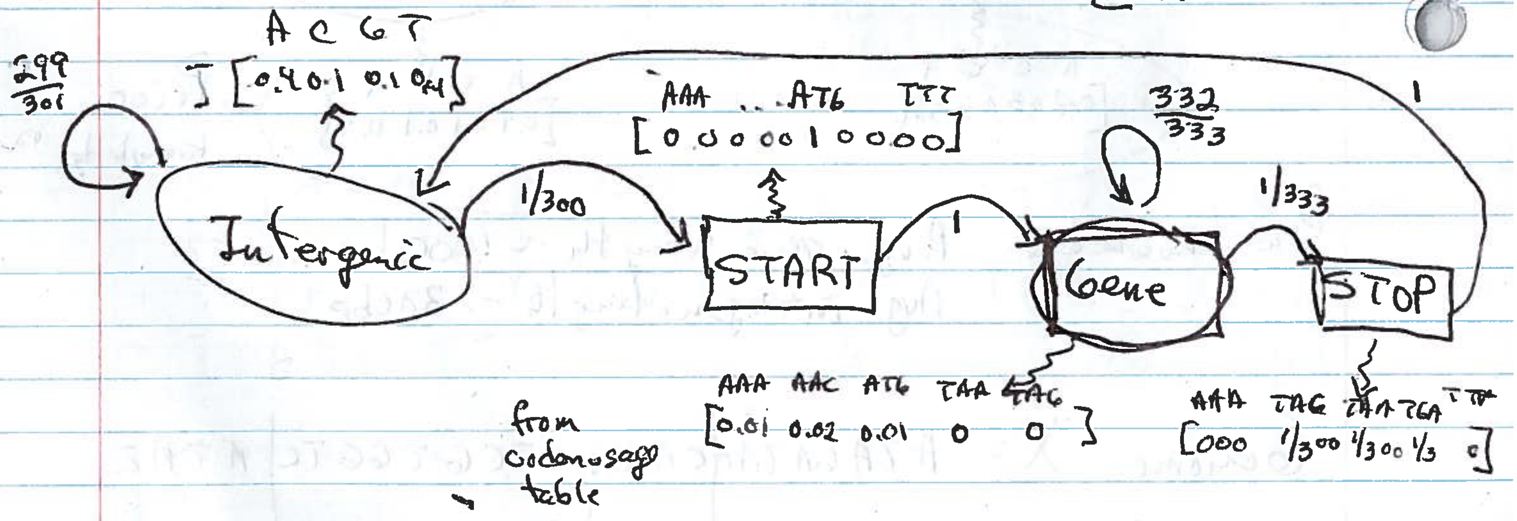
- Gene start with start codon: ATG
- end with stop codons TAG<br>TAA<br>TGA

- Gene is made of codons (triplets of nuc)
- Some codons are more common than others

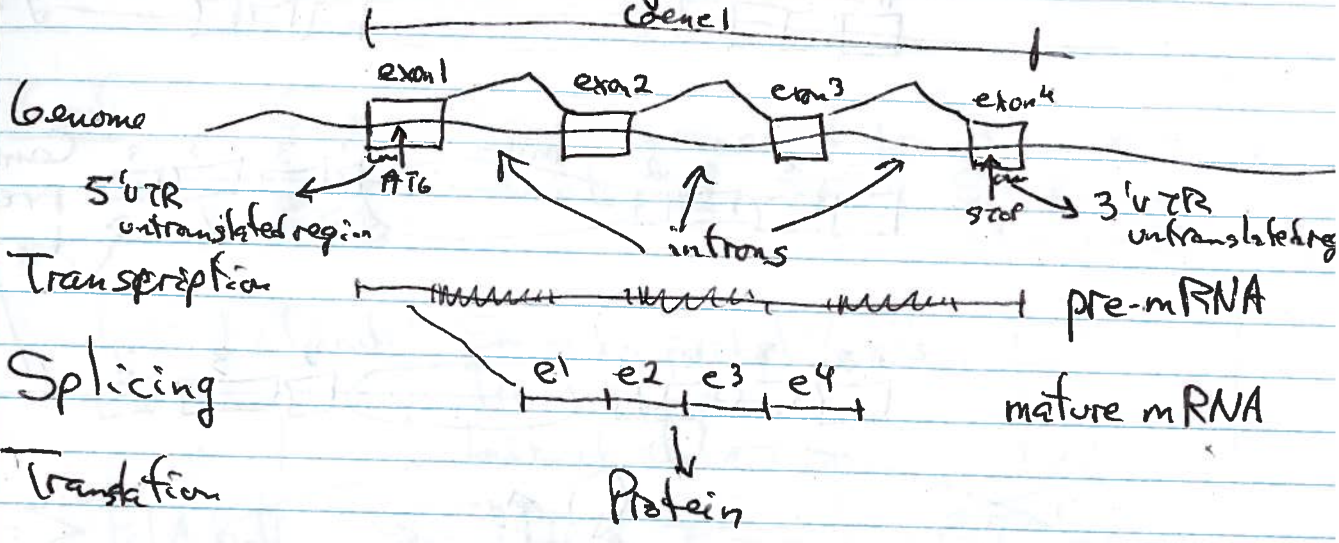
Codon usage table

| AAA | AAC | ATG | ... | TAA | TAG | TGA | |
|---|---|---|---|---|---|---|---|
| 0.01 | 0.02 | 0.01 | | 0 | 0 | 0 | |

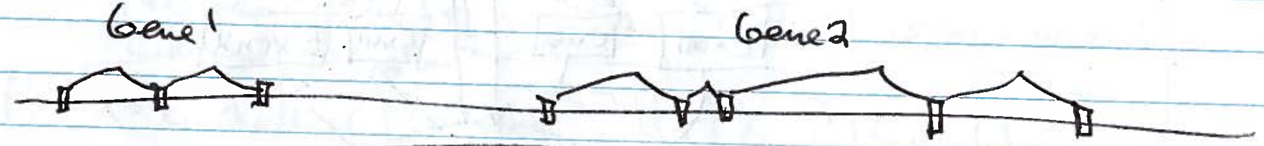
$\Sigma = 1$

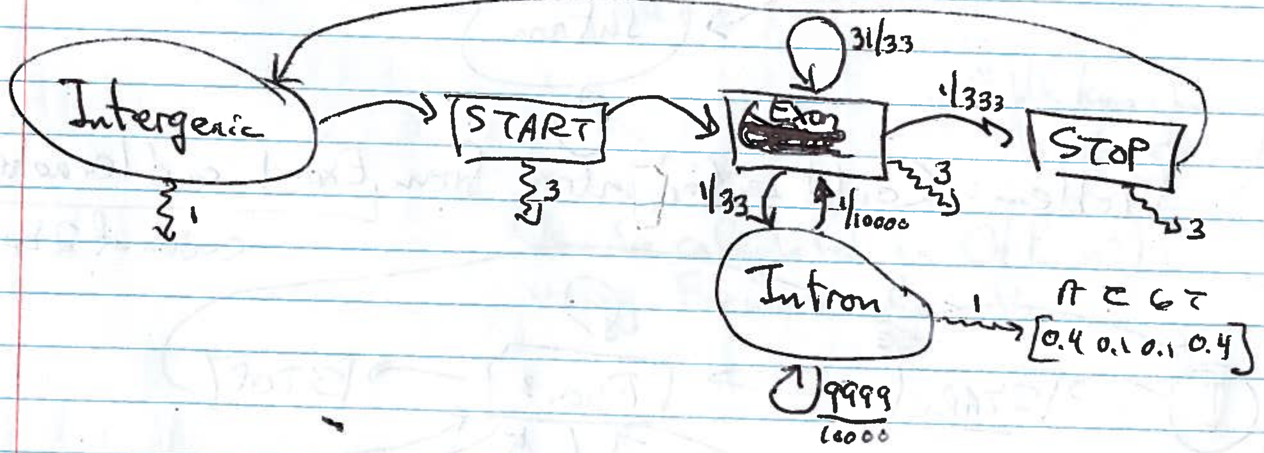


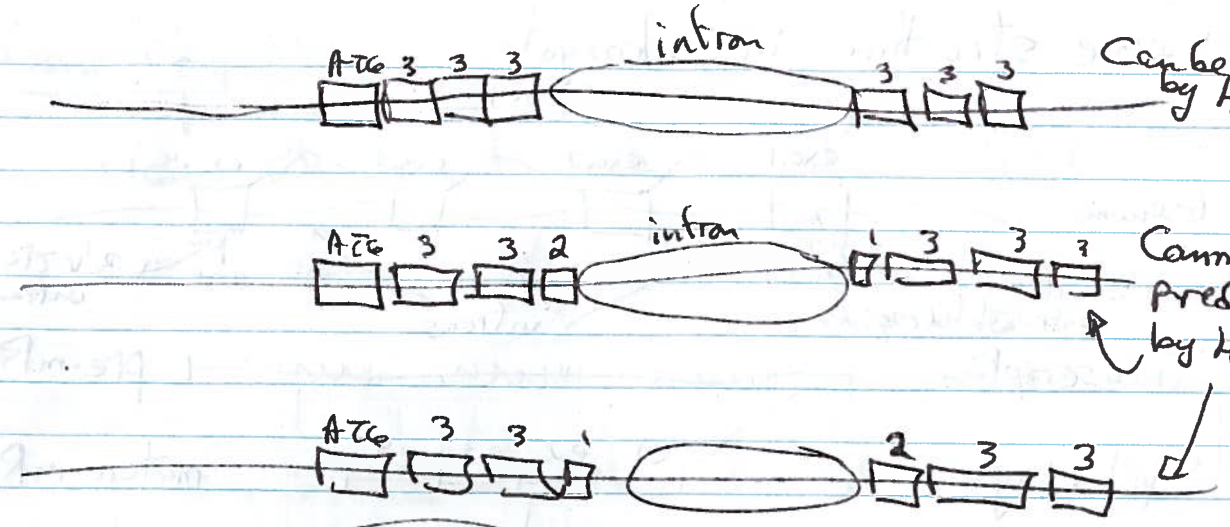
# Gene structure in eukaryote



Typical gene in human : 10 exons of ~100 bp on averag
9 introns of ~10000 bp on aver

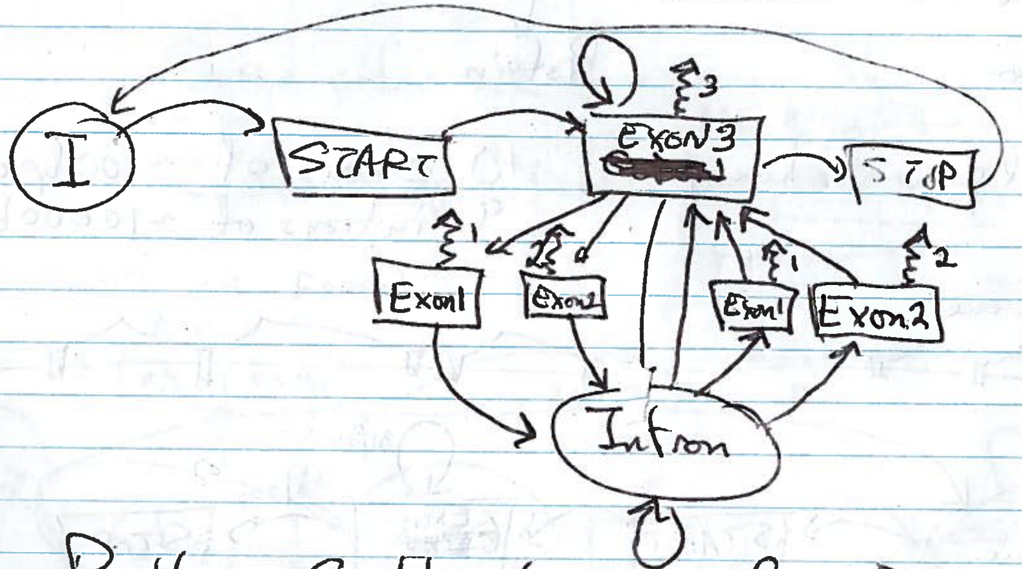




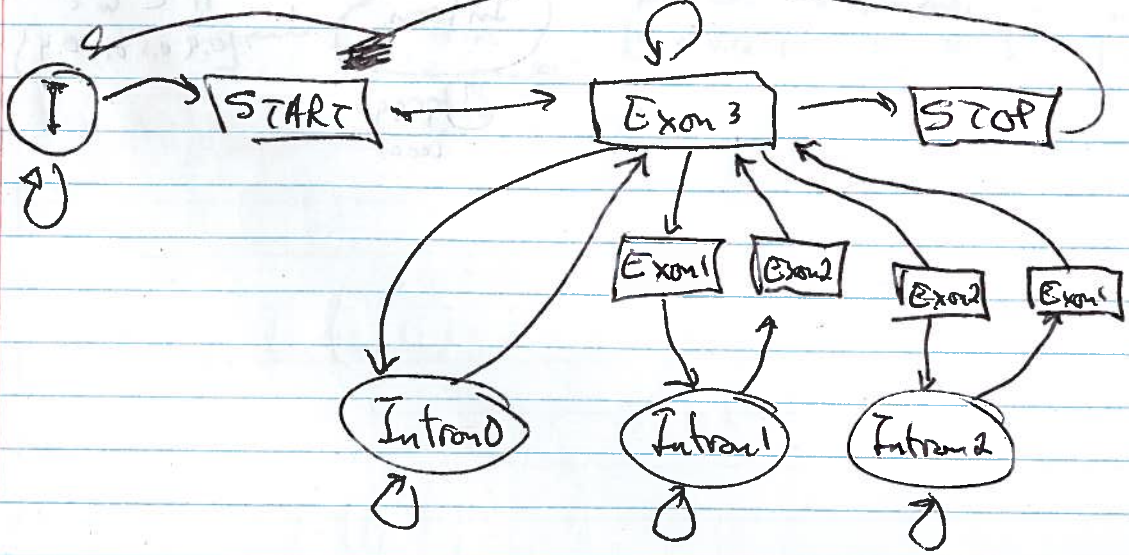


Can be predicted by HMM

Cannot be predicted by HMM



Problem: Could enter intron from, $\underbrace{\text{Exon1 and leave from Exon1}}_{\text{codon of 2bp}}$


I
START
Exon3
STOP
Exon1
Exon2
Exon2
Exon1
Intron0
Intron1
Intron2