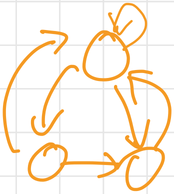
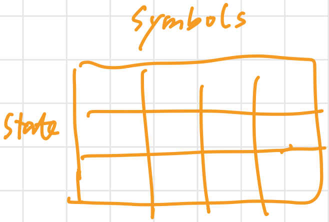
Review:

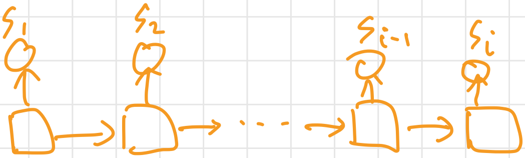
① HMM





transission T[p,p']

emission E[p, s].



$$D[i,p] = \max_{p'} D[i-1,p'] \times T[p',p] \times E[p,s_i].$$

② Gene

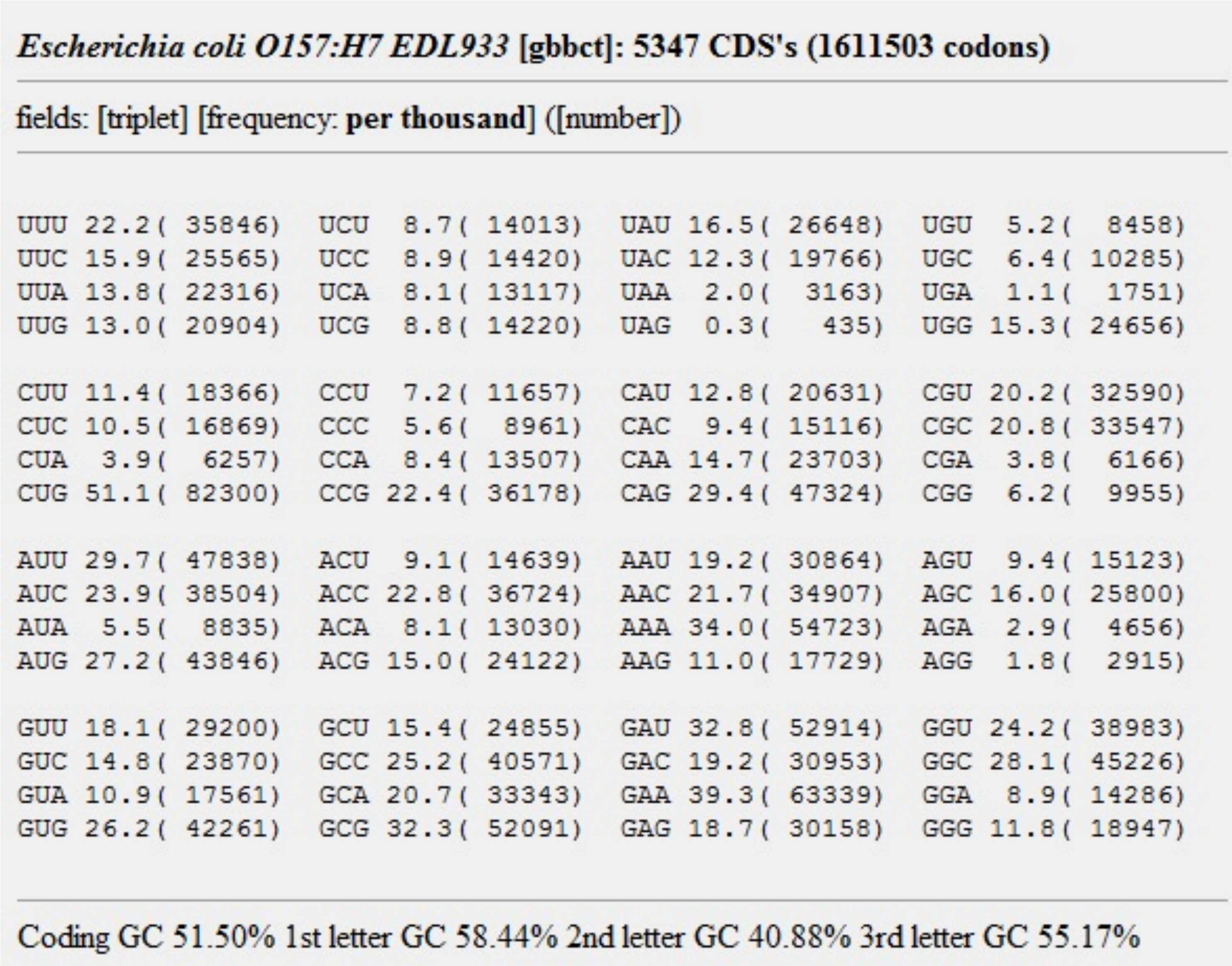
prokaryotes v.s eukaryotes

genetic code, start, stop, codons.

open reading frame.

codon bias.

http://www.kazusa.or.jp/codon/



Escherichia coli O157:H7 EDL933 **[gbbct]: 5347 CDS's (1611503 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

```
UUU 22.2( 35846)  UCU  8.7( 14013)  UAU 16.5( 26648)  UGU  5.2(  8458)
UUC 15.9( 25565)  UCC  8.9( 14420)  UAC 12.3( 19766)  UGC  6.4( 10285)
UUA 13.8( 22316)  UCA  8.1( 13117)  UAA  2.0(  3163)  UGA  1.1(  1751)
UUG 13.0( 20904)  UCG  8.8( 14220)  UAG  0.3(   435)  UGG 15.3( 24656)

CUU 11.4( 18366)  CCU  7.2( 11657)  CAU 12.8( 20631)  CGU 20.2( 32590)
CUC 10.5( 16869)  CCC  5.6(  8961)  CAC  9.4( 15116)  CGC 20.8( 33547)
CUA  3.9(  6257)  CCA  8.4( 13507)  CAA 14.7( 23703)  CGA  3.8(  6166)
CUG 51.1( 82300)  CCG 22.4( 36178)  CAG 29.4( 47324)  CGG  6.2(  9955)

AUU 29.7( 47838)  ACU  9.1( 14639)  AAU 19.2( 30864)  AGU  9.4( 15123)
AUC 23.9( 38504)  ACC 22.8( 36724)  AAC 21.7( 34907)  AGC 16.0( 25800)
AUA  5.5(  8835)  ACA  8.1( 13030)  AAA 34.0( 54723)  AGA  2.9(  4656)
AUG 27.2( 43846)  ACG 15.0( 24122)  AAG 11.0( 17729)  AGG  1.8(  2915)

GUU 18.1( 29200)  GCU 15.4( 24855)  GAU 32.8( 52914)  GGU 24.2( 38983)
GUC 14.8( 23870)  GCC 25.2( 40571)  GAC 19.2( 30953)  GGC 28.1( 45226)
GUA 10.9( 17561)  GCA 20.7( 33343)  GAA 39.3( 63339)  GGA  8.9( 14286)
GUG 26.2( 42261)  GCG 32.3( 52091)  GAG 18.7( 30158)  GGG 11.8( 18947)
```

Coding GC 51.50% 1st letter GC 58.44% 2nd letter GC 40.88% 3rd letter GC 55.17%

$$\frac{Pr(seq \mid gene)}{Pr(seq \mid random)}$$

A Better Gene Finder

- We can use the log likelihood ratio score to evaluate each ORF. Each codon XYZ contributes score
- $\log \frac{P(XYZ)}{P(X)P(Y)P(Z)}$
- An ORF is predicted as a gene if the sum of codon score is above a threshold.
- This is better. But it does not catch the correlation between adjacent codons.

HMM

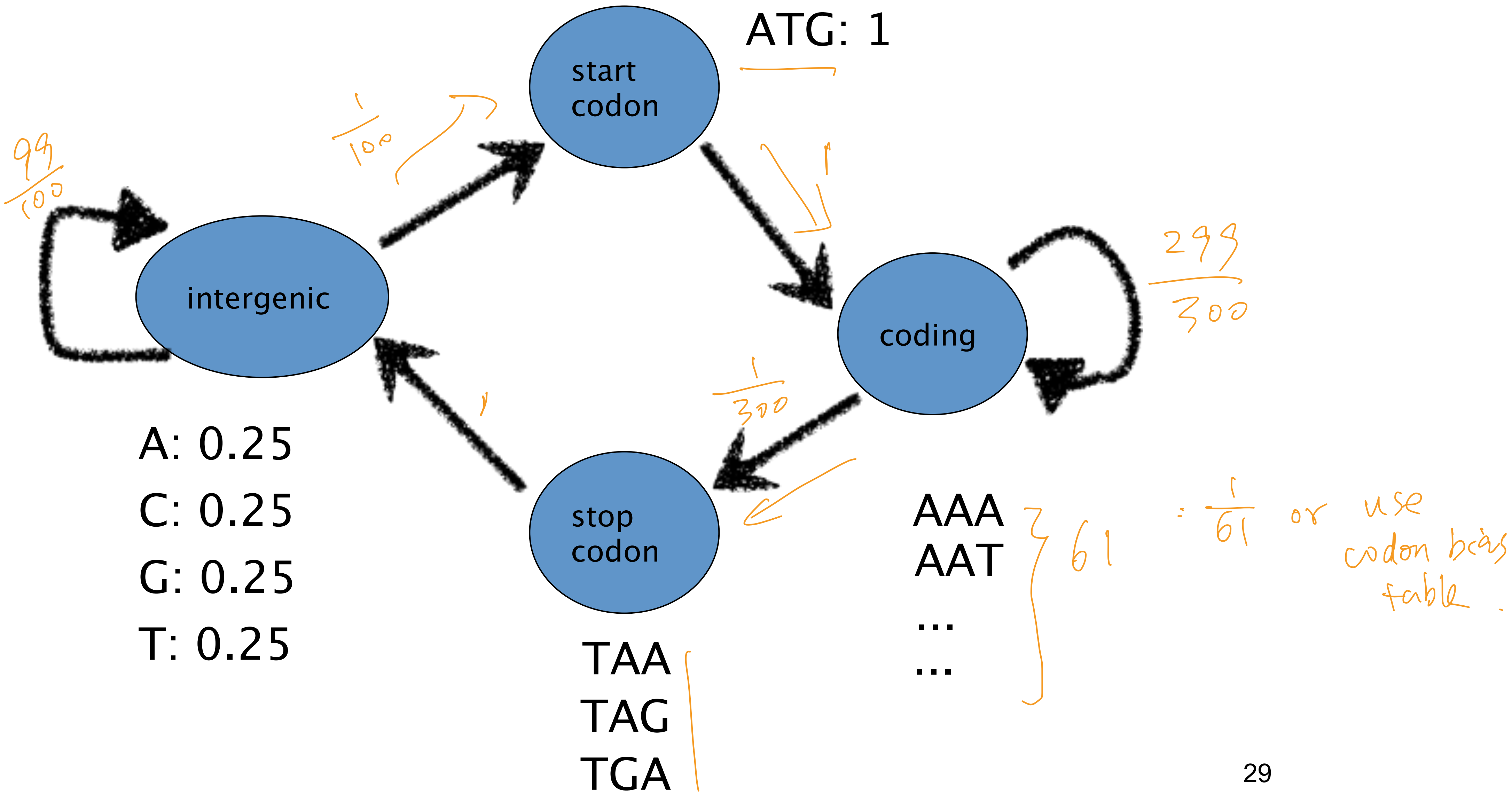
- We have used HMM in the classroom example to catch correlations between adjacent events.
- This can be used to model gene prediction.
- For example:
  - Symbols: Nucleotide bases.
  - States: start codon, stop codon, coding, non-coding (intergenic).

symbols

Input: A C A T G C T G

Hidden States:

Prokaryote gene finding HMM



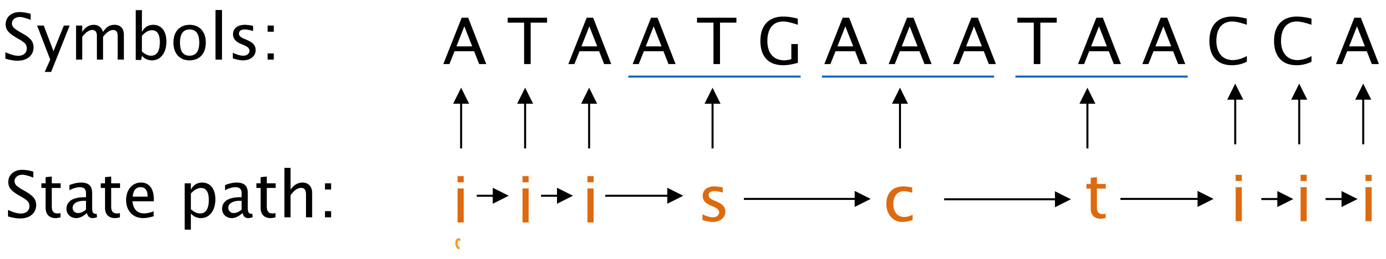
A: 0.25
C: 0.25
G: 0.25
T: 0.25

TAA
TAG
TGA

AAA
AAT
...
...

61 : 1/61 or use codon bias table.

Gene Prediction as HMM



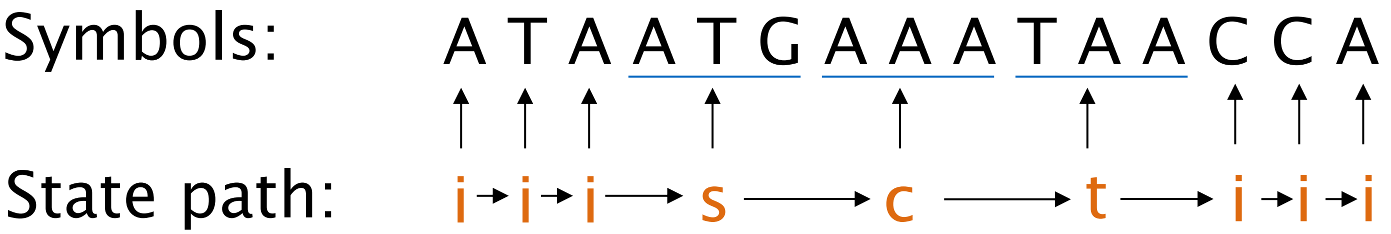
Symbols: A T A ATG AAA TAA C C A

State path: i → i → i → s → c → t → i → i → i

start codon
coding
intergenic
stop codon

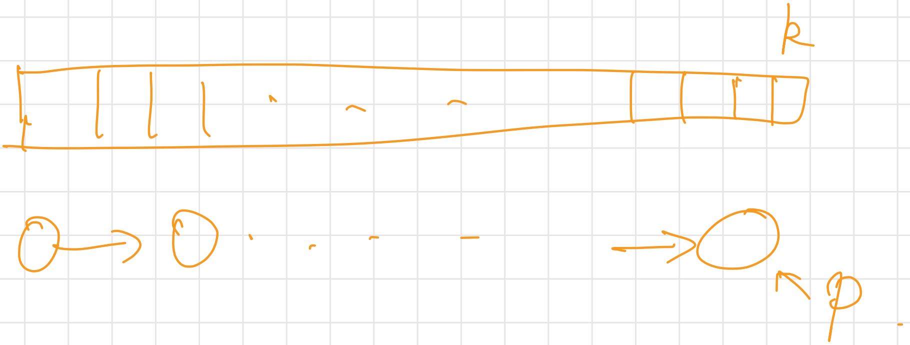
- Annotated the sequence with most probable path of states. This provides a reasonable answer to gene prediction.
- A difference here: emission is not fixed length. But this does not forbid us from solving it with dynamic programming.

Dynamic Programming



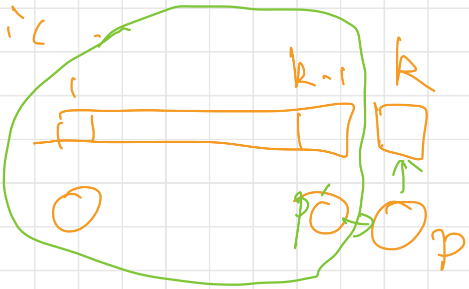
start codon
coding
intergenic
stop codon

- Define D[k,p] be the max probability achieved by first k symbols for a path with the last state being p.



$$D[k,p]$$

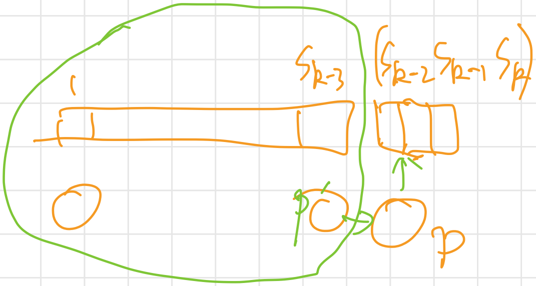
Case 1: $p =$ intergenic



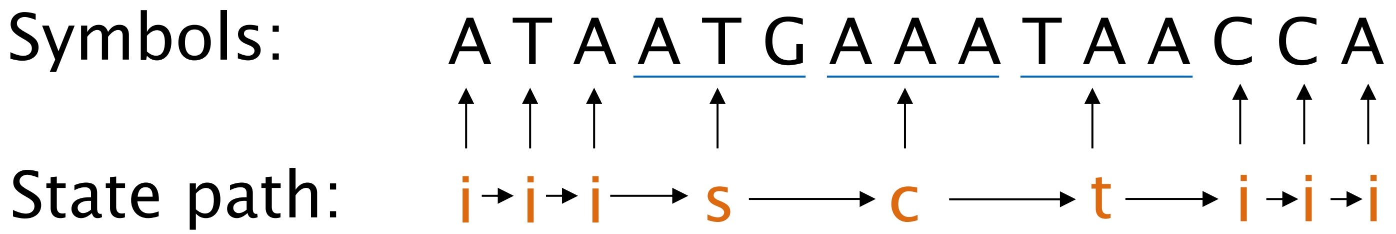
$$D[k,p] = \max_{p'} D[k-1,p'] \times T[p',p] \times E[p,s_k]$$

Case 2: $p \neq$ intergenic

$$D[k,p] = \max_{p'} D[k-3,p'] \times T[p',p] \times E[p, s_{k-2}s_{k-1}s_k]$$



Recurrence Relation



start codon
coding
intergenic
stop codon

For p=intergenic

$$D[k,p] = \max_{p'} D[k-1,p'] \Pr(p|p') \Pr(S_k|p)$$

For p=start, coding, or stop

$$D[k,p] = \max_{p'} D[k-3,p'] \Pr(p|p') \Pr(S_{k-2}S_{k-1}S_k|p)$$

Dynamic Programming

- Once the recurrence relation is obtained. It is straightforward to work out a dynamic programming algorithm.

Easy enough to implement?

- This is very easy to implement.
- If desired, one can also use a higher-order HMM.
- Parameter training must be done carefully.

Gene Prediction

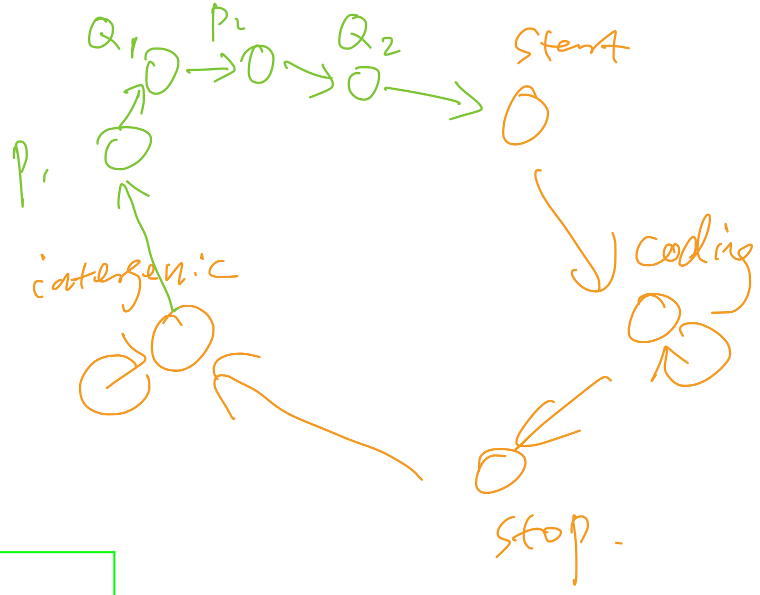
- Besides the codon bias that can be captured by HMM, there are other signals in a gene structure that can be employed by a gene prediction program.
  - E.g. the promoter of a gene is a region of DNA sequence located near the start codon.

```
                     -10   -5
 <-- upstream                                                    downstream -->
5'-XXXXPPPPPPXXXXXXXXXPPPPPPXXXXGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGXXXX-3‘
       -35            -10       Gene to be translated
```

Promoters

```
<-- upstream                                                downstream -->
5'-XXXXPPPPPPXXXXXXXXXPPPPPPXXXXGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGXXXX-3'
      -35 P1 Q1    -10 P2 Q2 Gene to be translated
```

-10:	T	A	T	A	A	T
	77%	76%	60%	61%	56%	82%
-35:	T	T	G	A	C	A
	69%	79%	61%	56%	54%	54%



- These rules are only approximately correct.
- The presence of promoters allow a very high transcription rate.
- Exercise: How to assign a score to the promoter.

PSWM

Positional Specific Weight Matrix

$$\begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline A & 0.1 & 0.76 & 0.1 & 0.61 & 0.56 & 0.1 \\ C & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.04 \\ G & 0.03 & 0.1 & 0.2 & 0.1 & 0.2 & 0.04 \\ T & 0.77 & 0.04 & 0.6 & 0.19 & 0.14 & 0.82 \end{array}$$

$p_i(a)$ $\qquad$ $q(a) = \frac{1}{4}.$

$$\text{score}(s_1 \cdots s_6) = \log \frac{\Pr(s_1 \cdots s_6 \mid \text{promoter})}{\Pr(s_1 \cdots s_6 \mid \text{random})}$$

$$= \log \frac{\prod_{i=1}^{6} p_i(s_i)}{\prod_{i=1}^{6} q(s_i)}$$

$$= \log \prod_{i=1}^{6} \frac{p_i(s_i)}{q(s_i)}$$

$$= \sum_{i=1}^{6} \log \frac{p_i(s_i)}{q(s_i)}$$

Summary

- HMM is a general model to predict some hidden states by examining emitted symbols.
- HMM can be used in gene prediction to harvest the codon bias and adjacent codon correlation.
- Gene prediction can use more information about the gene structure than codon bias.
- We only talked about prokaryote gene prediction. Eukaryote gene prediction is harder because of introns.