



# Gene Finding Project (Cont.)

---

Charles Yan



# Gene Finding

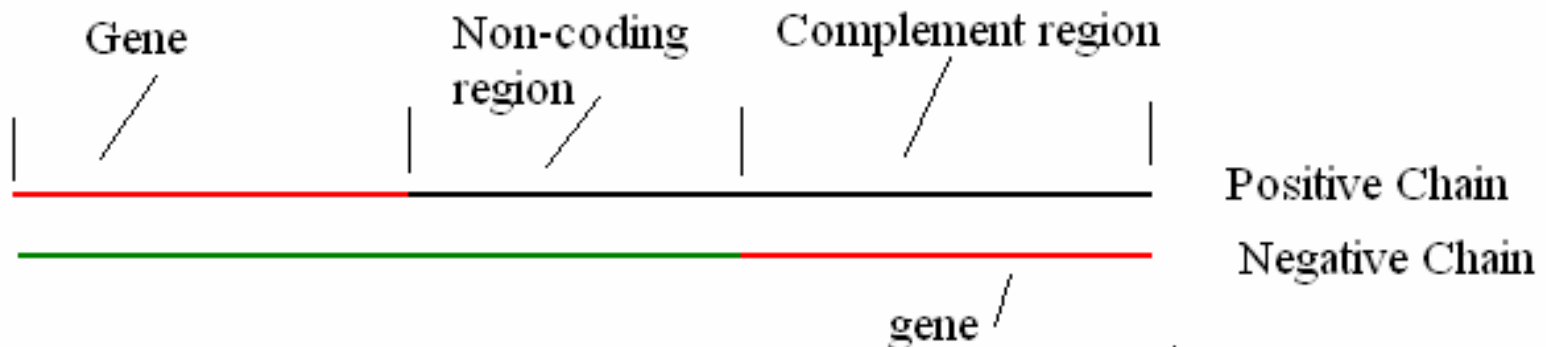
---

- **Summary of the project**
  - Download the genome of E. Coli K12
  - Gene-finding using  $k^{\text{th}}$ -order Markov chains, where  $k = 1, 2, 3$
  - Gene-finding using inhomogeneous Markov chains

# Non-Coding Regions

The rest of the genome that are not labeled as gene or complement does not encode genetic information. These regions are *non-coding regions*.

The following figure shows that the positive chain is divided into three types of region: gene, non-coding region and complement region.





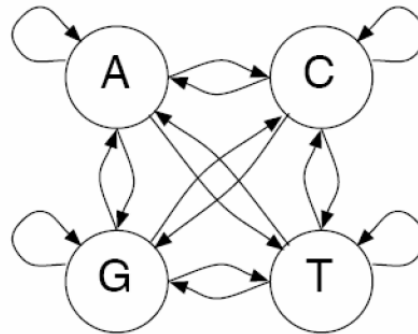
# 1<sup>st</sup>-Order Markov Chain

---

- Since there are three types of regions on the sequence we have, we will develop three models corresponding to them: *gene model*, *non-coding model* and *complement model*.

# 1<sup>st</sup>-Order Markov Chain

- For these models, we use the same structure as we shown in the example of identifying CpG island.



States: A,C,G,T

Emissions: corresponding letter

Transitions:  $a_{st} = P(x_i = t \mid x_{i-1} = s)$

The structure of the 1<sup>st</sup>-order Markov chain model.

# 1<sup>st</sup>-Order Markov Chain

■ Then, each model is reduced to a transition probability table. Here is an example for the gene model (1st-order Markov chain). **We will need to estimate the probabilities for each model.**

This is the model for gene

End state

Gene	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	<u>0.274</u>	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

Starting state

The probability that the C state transit into G state. In another word, it the probability that you will see a C followed by a G on the genome sequence.



# 1<sup>st</sup>-Order Markov Chain

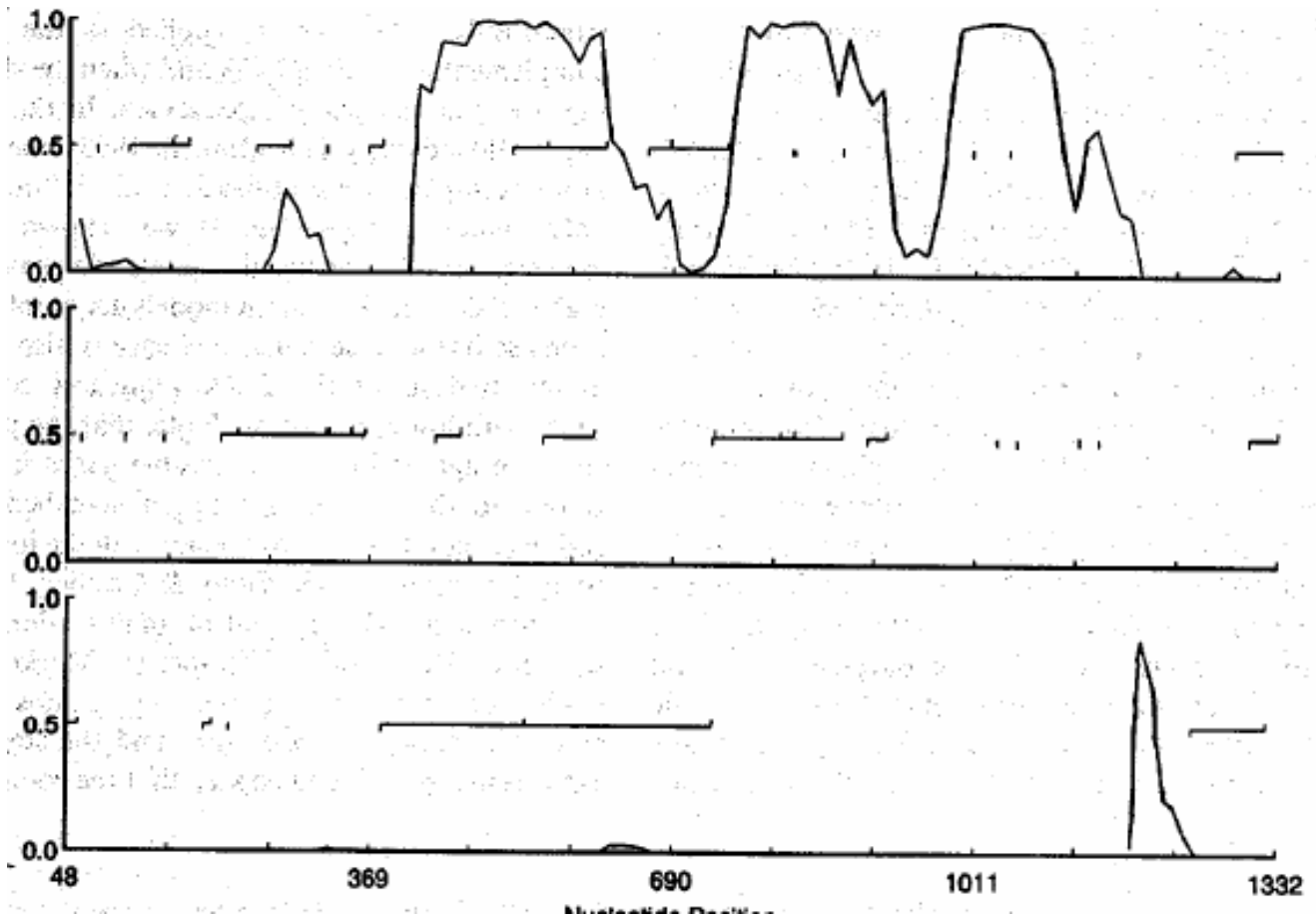
---

Now we have three models, i.e. three probability tables. Then, how are we going to make predictions using them? For example, we have a sequence  $X=x_1x_2x_3\dots x_n$ , where  $x_i \in \{A, T, C, G\}$ . Then, **the probability that this sequence belongs to gene is given by**

$$\begin{aligned}P_{gene}(X) &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2x_1) \cdot \dots \cdot P_{gene}(x_n | x_{n-1}x_{n-2}\dots x_1) \\ &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2) \cdot \dots \cdot P_{gene}(x_n | x_{n-1}) \\ &= P_{gene}(x_1) \cdot a_{gene}(x_1, x_2) \cdot a_{gene}(x_2, x_3) \cdot \dots \cdot a_{gene}(x_{n-1}, x_n) \\ &= P_{gene}(x_1) \cdot \prod_{i=2}^n a_{gene}(x_{i-1}, x_i)\end{aligned}$$

Where  $a_{gene}(x_{i-1}, x_i)$  is the transition probability from  $x_{i-1}$  to  $x_i$  in the gene model, and  $P_{gene}(x_1)$  is the prior probability that  $x_1$  appears in genes, that is, the fraction of the letters in genes that are  $x_1$ .

# 1<sup>st</sup>-Order Markov Chain

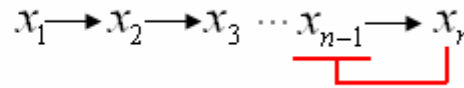
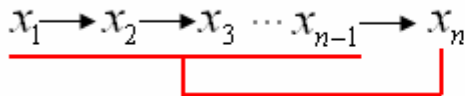


# K<sup>th</sup>-Order Markov Chain

$$\begin{aligned}P_{gene}(X) &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2 x_1) \cdot \dots \cdot P_{gene}(x_n | x_{n-1} x_{n-2} \dots x_1) \\ &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2) \cdot \dots \cdot P_{gene}(x_n | x_{n-1}) \\ &= P_{gene}(x_1) \cdot a_{gene}(x_1, x_2) \cdot a_{gene}(x_2, x_3) \cdot \dots \cdot a_{gene}(x_{n-1}, x_n) \\ &= P_{gene}(x_1) \cdot \prod_{i=2}^n a_{gene}(x_{i-1}, x_i)\end{aligned}$$

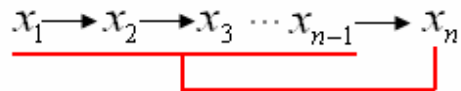
The assumption by 1<sup>st</sup>-order Markov chain is that  $x_n$  is independent of  $x_{n-2} x_{n-3} \dots x_1$  given  $x_{n-1}$ , i.e.,

$$P_{gene}(x_n | x_{n-1} x_{n-2} \dots x_1) = P_{gene}(x_n | x_{n-1})$$



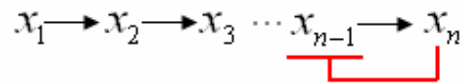


# $K^{\text{th}}$ -Order Markov Chain



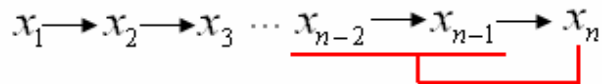
$$P_{\text{gene}}(x_n | x_{n-1}x_{n-2}\cdots x_1)$$

1<sup>st</sup>-order



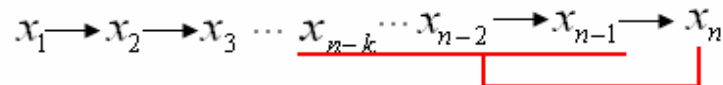
$$P_{\text{gene}}(x_n | x_{n-1}x_{n-2}\cdots x_1) = P_{\text{gene}}(x_n | x_{n-1})$$

2<sup>nd</sup>-order



$$P_{\text{gene}}(x_n | x_{n-1}x_{n-2}\cdots x_1) = P_{\text{gene}}(x_n | x_{n-1}x_{n-2})$$

$k^{\text{th}}$ -order



$$P_{\text{gene}}(x_n | x_{n-1}x_{n-2}\cdots x_1) = P_{\text{gene}}(x_n | x_{n-1}x_{n-2})$$



# $K^{\text{th}}$ -Order Markov Chain

---

- When  $K=2$  is used, the changes in the method include:
  - (1) The size of the transition probability table for each model will become  $16 \times 4$ .

	A	C	T	G
AA				
AT				
AC				
AG				
TA				
TT				
TC				
TG				
...				



# K<sup>th</sup>-Order Markov Chain

---

(2) The formula used in calculating the probabilities change to

$$\begin{aligned} P_{gene}(X) &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2 x_1) \cdot \dots \cdot P_{gene}(x_n | x_{n-1} x_{n-2} \dots x_1) \\ &= P_{gene}(x_1) \cdot P_{gene}(x_2 | x_1) \cdot P_{gene}(x_3 | x_2 x_1) \cdot \dots \cdot P_{gene}(x_n | x_{n-1} x_{n-2}) \\ &\dots \\ &= P_{gene}(x_1 x_2) \cdot \prod_{i=3}^n a_{gene}(x_{i-2} x_{i-1}, x_i) \end{aligned}$$

When other value of K is used, similar changes apply.



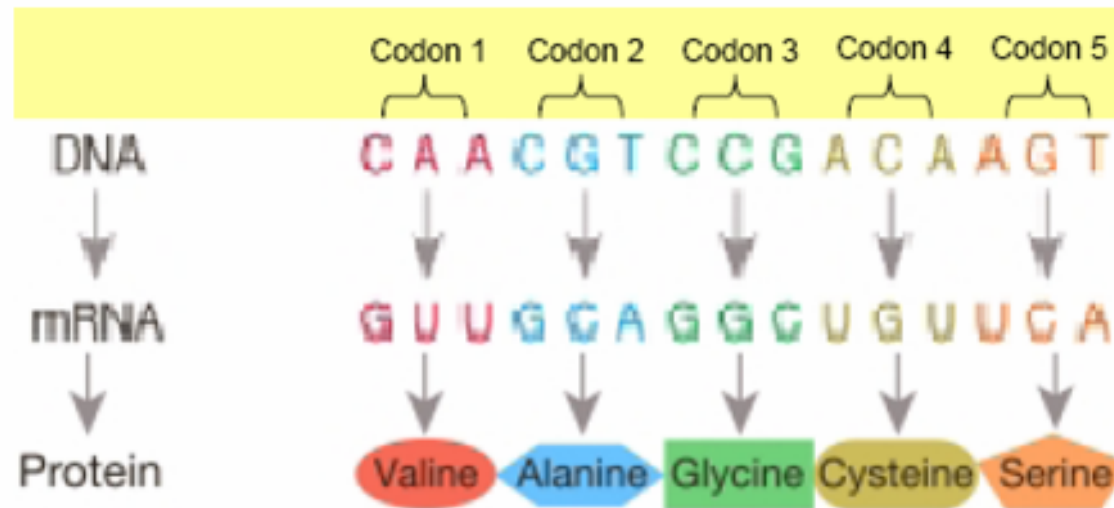
# $K^{\text{th}}$ -Order Markov Chain

---

- $K=1,2 \dots$
- One model for each type of sequence: *gene*, *complement*, and *non-coding*.
- Every position in the same type of sequence is considered the same.
- But there are some differences between different positions.
- We need new methods to address these differences.

# Inhomogeneous Markov Chains

- When DNA is translated into proteins, three bases (the letters for DNA, which are A, T, G, and C) make up a codon and encode one amino acid residue (the letter for Proteins).





# Inhomogeneous Markov Chains

---

- Each codon has three positions. In the previous models (referred as *homogeneous models*), we do not distinguish between the three positions. In this section, we will build different models (referred as *inhomogeneous models*) for different positions.

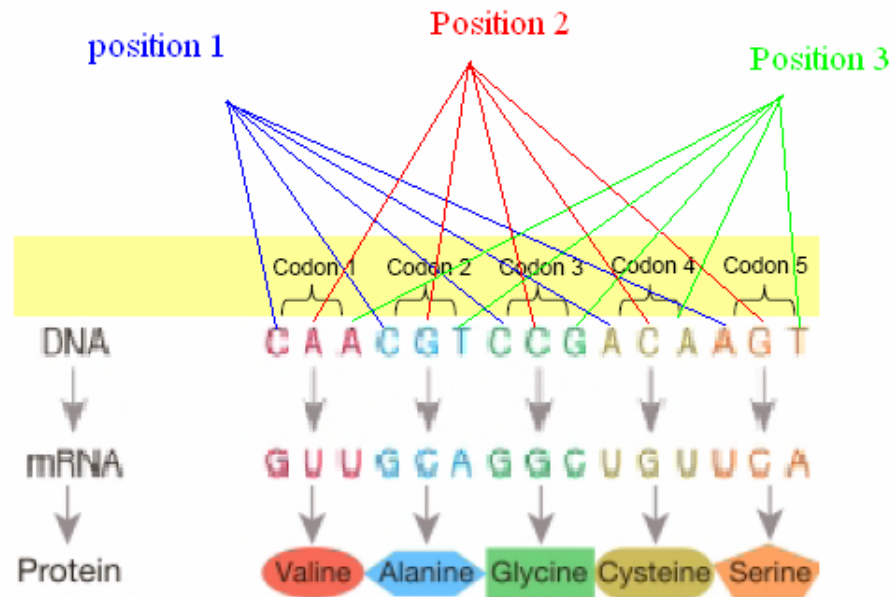
# Gene Models

- The gene model will be split into three models, each for one codon position:

$a_{gene\_code1}$

$a_{gene\_code2}$

$a_{gene\_code3}$





# Gene Models

---

$\alpha_{gene\_code1}$

Gene Position 1	A	C	G	T
A				
C				
G				
T				

$\alpha_{gene\_code2}$

Gene Position 2	A	C	G	T
A				
C				
G				
T				

$\alpha_{gene\_code3}$

Gene Position 3	A	C	G	T
A				
C				
G				
T				

# Gene Models

- In the prediction stage, for a given sequence  $X = x_1 x_2 x_3 \dots x_n$ , we will have to calculate three probabilities.

(1) The probability that  $X$  is a gene sequence starting with position 1.

$$\underbrace{x_1 x_2 x_3}_{\text{Codon}} \underbrace{x_4 x_5 x_6}_{\text{Codon}} \underbrace{x_7 \dots x_n}_{\text{Codon}}$$

$$P_{\text{gene\_code1}}(X) = P_{\text{gene\_code1}}(x_1) \cdot a_{\text{gene\_code1}}(x_1, x_2) \cdot a_{\text{gene\_code2}}(x_2, x_3) \cdot a_{\text{gene\_code3}}(x_3, x_4) \cdot a_{\text{gene\_code1}}(x_4, x_5) \cdot a_{\text{gene\_code2}}(x_5, x_6) \cdot a_{\text{gene\_code3}}(x_6, x_7) \dots$$

(2) The probability that  $X$  is a gene sequence starting with position 2.

$$\underbrace{x_1 x_2 x_3}_{\text{Codon}} \underbrace{x_4 x_5 x_6}_{\text{Codon}} \underbrace{x_7 \dots x_n}_{\text{Codon}}$$

$$P_{\text{gene\_code2}}(X) = P_{\text{gene\_code2}}(x_1) \cdot a_{\text{gene\_code2}}(x_1, x_2) \cdot a_{\text{gene\_code3}}(x_2, x_3) \cdot a_{\text{gene\_code1}}(x_3, x_4) \cdot a_{\text{gene\_code3}}(x_4, x_5) \cdot a_{\text{gene\_code1}}(x_5, x_6) \cdot a_{\text{gene\_code2}}(x_6, x_7) \dots$$

(3) The probability that  $X$  is a gene sequence starting with position 3.

$$\underbrace{x_1 x_2 x_3}_{\text{Codon}} \underbrace{x_4 x_5 x_6}_{\text{Codon}} \underbrace{x_7 \dots x_n}_{\text{Codon}}$$

$$P_{\text{gene\_code3}}(X) = P_{\text{gene\_code3}}(x_1) \cdot a_{\text{gene\_code3}}(x_1, x_2) \cdot a_{\text{gene\_code3}}(x_2, x_3) \cdot a_{\text{gene\_code1}}(x_3, x_4) \cdot a_{\text{gene\_code2}}(x_4, x_5) \cdot a_{\text{gene\_code3}}(x_5, x_6) \cdot a_{\text{gene\_code1}}(x_6, x_7) \dots$$



# Complement Region Models

---

- We treat complement region the same way as we do genes. Three models will be built for three positions.
- Three probabilities will be calculated when prediction is to be made for a sequence  $X = x_1x_2x_3 \dots x_n$ ,

(1)  $P_{\text{complement\_code1}}(X)$ : The probability that X is a complement region starting with position 1.

(2)  $P_{\text{complement\_code2}}(X)$ : The probability that X is a complement region starting with position 2.

(3)  $P_{\text{complement\_code3}}(X)$ : The probability that X is a complement region starting with position 3.



# Non-Coding Region Model

---

- Since the non-coding region does not contain codons, every position will be considered the same. There is no change to the non-coding region model. will be calculated as described in the homogeneous models.



# Inhomogeneous Markov Chains

---

Then for each input sequence, we need to calculate and compare 7 probabilities:

$$P_{non-coding}(X)$$

$$P_{complement\_code3}(X)$$

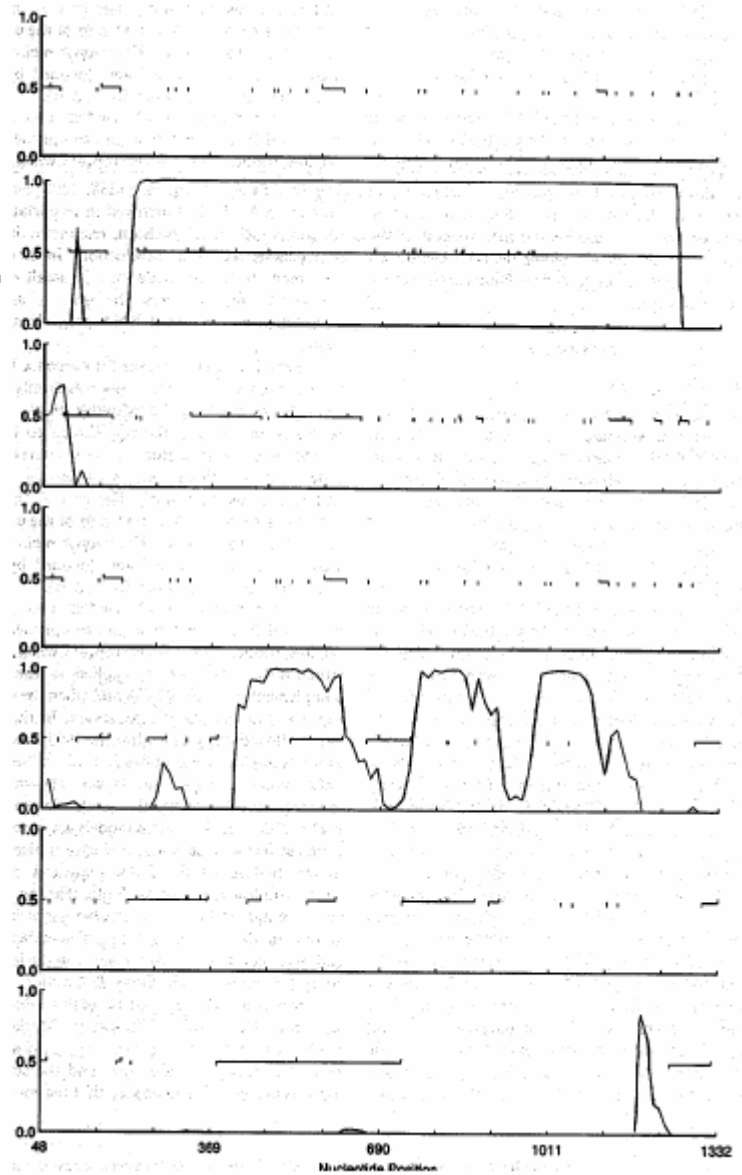
$$P_{complement\_code2}(X)$$

$$P_{complement\_code1}(X)$$

$$P_{gene\_code1}(X)$$

$$P_{gene\_code2}(X)$$

$$P_{gene\_code3}(X)$$





# Markov Chains

---

100

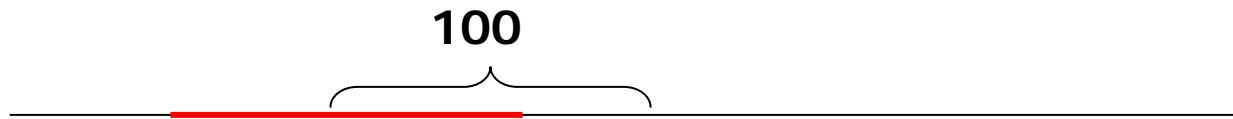


- The test sequence is divided into fragments using a sliding window of 100 letters.
- Predictions are made for each window.
- Each prediction for a window.
- We need new methods that can make prediction for each letter.



# Markov Chains

---



- What if the window is on the boundary of a gene?
- The methods can not predict boundaries precisely.
- We need methods that can do so!







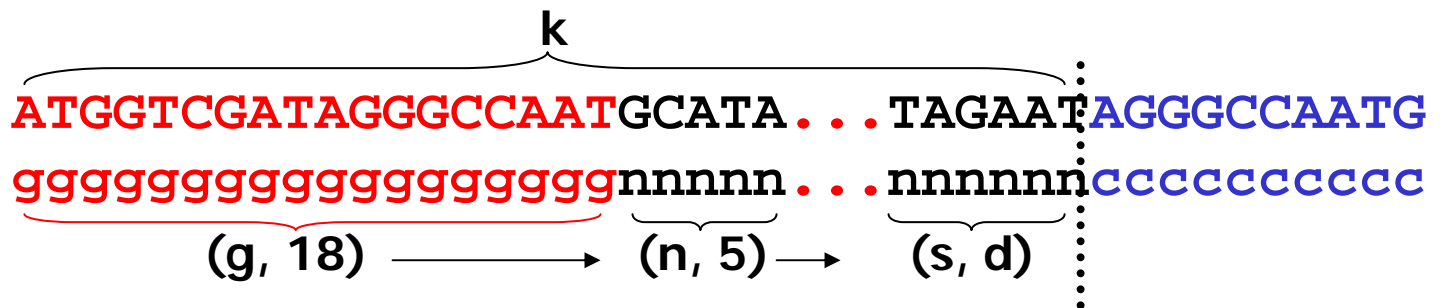
# Hidden Markov Model (HMM) with Duration

---

- We can use dynamic programming to find the optimal path for a given sequence.
- Let  $X = x_1 x_2 x_3 x_4 x_5 \dots x_n$  be the sequence about which predictions are to be made.

# Hidden Markov Model (HMM) with Duration

- Let  $Z_k(s,d)$  be the maximum probability that subsequence  $x_1x_2x_3\dots x_k$  is generated by a path ending with  $(s,d)$ .
- Let  $S_k$  be the maximum probability of generating subsequence  $x_1x_2x_3\dots x_k$  using any path.
  - Then 
$$S_k = \arg \max_{(s,d)} Z_k(s,d)$$
- Let  $P_k$  be the last note of the optimal path that generates subsequence  $x_1x_2x_3\dots x_k$ .  $P_k \rightarrow s$  refers to its state and  $P_k \rightarrow d$  refers to the duration of its state. Note that  $S_k = Z_k(P_k \rightarrow s, P_k \rightarrow d)$

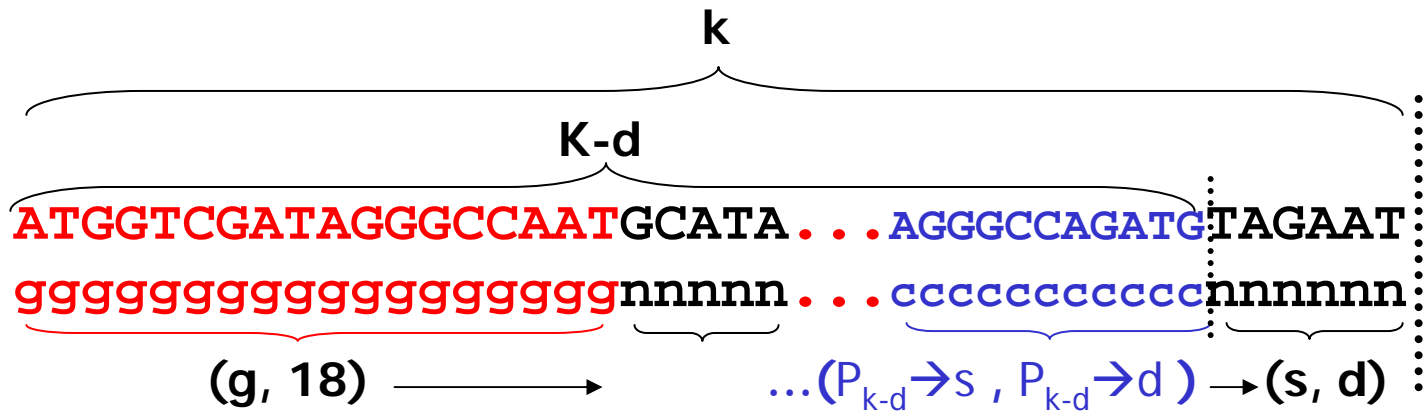


# Hidden Markov Model (HMM) with Duration

- The recursion to calculate  $Z_k(s,d)$  is:

- if  $s \neq P_{k-d} \rightarrow s$

$$Z_k(s,d) = S_{k-d} * Q(P_{k-d} \rightarrow s, s) * D(s,d) * E_s(x_{k-d}x_{k-d+1}\dots x_k)$$



$Q(P_{k-d} \rightarrow s, s)$ : Transition probability from  $P_{k-d} \rightarrow s$  to  $s$

$D(s,d)$ : Probability that state  $s$  has a duration of  $d$

$E_s(x_{k-d}x_{k-d-1}\dots x_k)$ : Probability that state  $s$  generates  $x_{k-d}x_{k-d-1}\dots x_k$

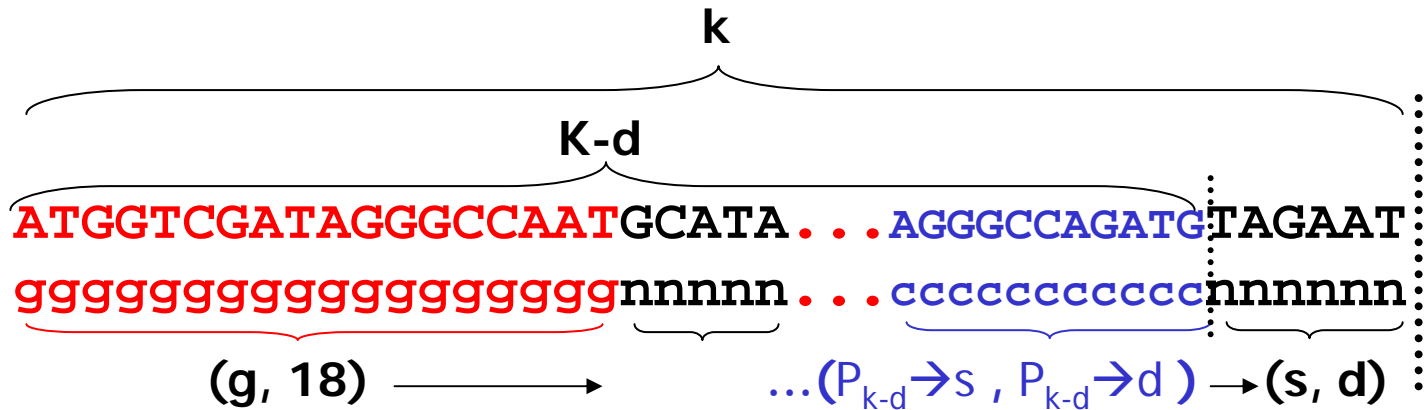
# Hidden Markov Model (HMM) with Duration

Since  $S_k = Z_k(P_k \rightarrow s, P_k \rightarrow d)$ , then

$$S_{k-d} = Z_{k-d}(P_{k-d} \rightarrow s, P_{k-d} \rightarrow d)$$

Thus

$$\begin{aligned} Z_k(s, d) &= S_{k-d} * Q(P_{k-d} \rightarrow s, s) * D(s, d) * E_s(x_{k-d} x_{k-d+1} \dots x_k) \\ &= Z_{k-d}(P_{k-d} \rightarrow s, P_{k-d} \rightarrow d) * Q(P_{k-d} \rightarrow s, s) * D(s, d) * E_s(x_{k-d} x_{k-d+1} \dots x_k) \end{aligned}$$





# Hidden Markov Model (HMM) with Duration

- Now we have the recursion function

$$Z_k(s, d) = \text{Max} \left\{ \begin{array}{l} Z_{k-d}(P_{k-d} \rightarrow s, P_{k-d} \rightarrow d) * Q(P_{k-d} \rightarrow s, s) * D(s, d) * E_s(x_{k-d} x_{k-d+1} \dots x_k) , \\ Z_{k-d-P_{i-1} \rightarrow d}(P_{k-d-P_{i-1} \rightarrow d} \rightarrow s, P_{k-d-P_{i-1} \rightarrow d} \rightarrow d) * Q(P_{k-d-P_{i-1} \rightarrow d} \rightarrow s, s) * D(s, d) * E_s(x_{k-d-P_{i-1} \rightarrow d} x_{k-d-P_{i-1} \rightarrow d+1} \dots x_k) \end{array} \right\}$$

- We will discuss  $Q(P_{k-d} \rightarrow s, s)$ ,  $D(s, d)$ , and  $E_s(x_{k-d} x_{k-d-1} \dots x_k)$  later. Here, let's assume that they are known.
- We can calculate  $Z_k(s, d)$  for all  $k, s, d$  where  $k \leq n$ , and  $d \leq k$

# Hidden Markov Model (HMM) with Duration

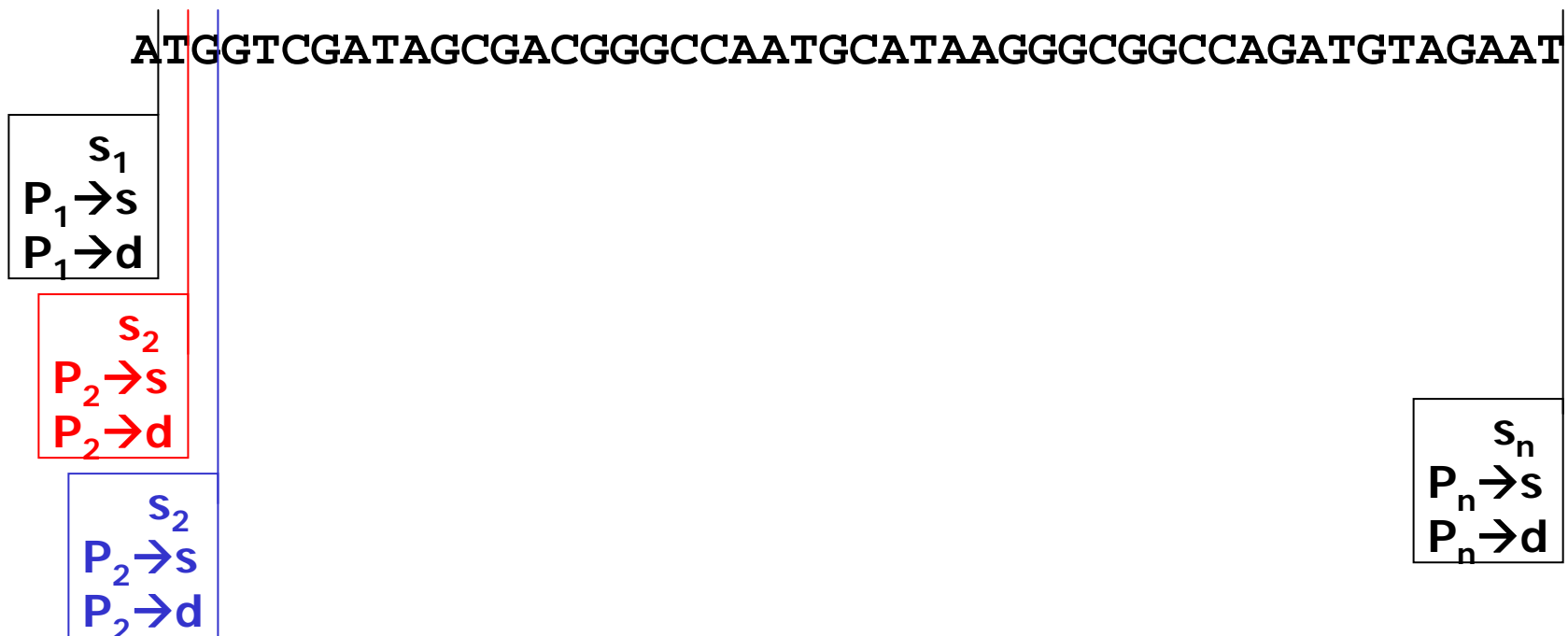
- $K=0,$ 
  - $Z_0(g,0)=1, S_0=1$
- $K=1,$ 
  - $Z_1(g,1)=1, Z_1(c,1)=1, Z_1(n,1)=1$
  - $Z_1(g,0)=1, Z_1(c,0)=1, Z_1(n,0)=1$
  - $S_0=1$
- $2 \leq K \leq n$

$$Z_k(s, d) = \text{Max} \left\{ \begin{array}{l} Z_{k-d}(P_{k-d} \rightarrow s, P_{k-d} \rightarrow d) * Q(P_{k-d} \rightarrow s, s) * D(s, d) * E_s(x_{k-d} x_{k-d+1} \dots x_k) , \\ Z_{k-d-P_{k-d} \rightarrow d}(P_{k-d-P_{k-d} \rightarrow d} \rightarrow s, P_{k-d-P_{k-d} \rightarrow d} \rightarrow d) * Q(P_{k-d-P_{k-d} \rightarrow d} \rightarrow s, s) * D(s, d) * E_s(x_{k-d-P_{k-d} \rightarrow d} x_{k-d-P_{k-d} \rightarrow d+1} \dots x_k) \end{array} \right\}$$

$$S_k = \arg \max_{(s,d)} Z_k(s, d) \longrightarrow \begin{array}{l} (P_k \rightarrow s, \\ P_k \rightarrow d) \end{array}$$

# Hidden Markov Model (HMM) with Duration

- $S_k, P_k \rightarrow s, P_k \rightarrow d$  (for all  $1 \leq k \leq n$ ) are the three tables we need to keep during this dynamic programming.





# Hidden Markov Model (HMM) with Duration

---

- At the end of the dynamic programming we will have  $S_k, P_k \rightarrow s, P_k \rightarrow d$  for all  $1 \leq k \leq n$ . Then we can make predictions for the whole sequence using a back-track approach.

**ATGGTCGATAGCGACGGGCCAATGCATAAGGGCGGCCAGATGTAGAAT**

$P_n \rightarrow s = n$
$P_n \rightarrow d = 5$



# Hidden Markov Model (HMM) with Duration

---

- At the end of the dynamic programming we will have  $S_k, P_k \rightarrow s, P_k \rightarrow d$  for all  $1 \leq k \leq n$ . Then we can make predictions for the whole sequence using a back-track approach.

ATGGTCGATAGCGACGGGCCAATGCATAAGGGCGGCCAGATGT**AGAAT**  
**nnnnn**

$P_n \rightarrow s = n$
$P_n \rightarrow d = 5$

# Hidden Markov Model (HMM) with Duration

- At the end of the dynamic programming we will have  $S_k, P_k \rightarrow s, P_k \rightarrow d$  for all  $1 \leq k \leq n$ . Then we can make predictions for the whole sequence using a back-track approach.

ATGGTCGATAGCGACGGGCCAATGCATAAGGGCGGCCAGATGT **AGAAT**  
**nnnnn**

$P_k \rightarrow s = g$   
 $P_k \rightarrow d = 30$

$P_n \rightarrow s = n$   
 $P_n \rightarrow d = 5$













## Hidden Markov Model (HMM) with Duration

---

- Let's go back to the statement "We will discuss  $Q(P_{k-d} \rightarrow s, s)$ ,  $D(s, d)$ , and  $E_s(x_{k-d} x_{k-d-1} \dots x_k)$  later. Here, let's assume that they are known."
- Now, we need to know how to estimate these functions.



# Hidden Markov Model (HMM) with Duration

---

- Let's go back to the statement "We will discuss  $Q(P_{k-d} \rightarrow s, s)$ ,  $D(s, d)$ , and  $E_s(x_{k-d} x_{k-d-1} \dots x_k)$  later. Here, let's assume that they are known."
- Now, we need to know how to estimate these functions.

$Q(P_{k-d} \rightarrow s, s)$ : Transition probability from  $P_{k-d} \rightarrow s$  to  $s$

$D(s, d)$ : Probability that state  $s$  has a duration of  $d$

$E_s(x_{k-d} x_{k-d-1} \dots x_k)$ : Probability that state  $s$  generates  $x_{k-d} x_{k-d-1} \dots x_k$

# Hidden Markov Model (HMM) with Duration

- $Q(P_{k-d} \rightarrow s, s)$ : Transition probability from  $P_{k-d} \rightarrow s$  to  $s$   
There are three states: gene (g), complement (c), and non-coding (n)

Q( )	g	c	n
g		$Q(g,c)$	
c			
n			

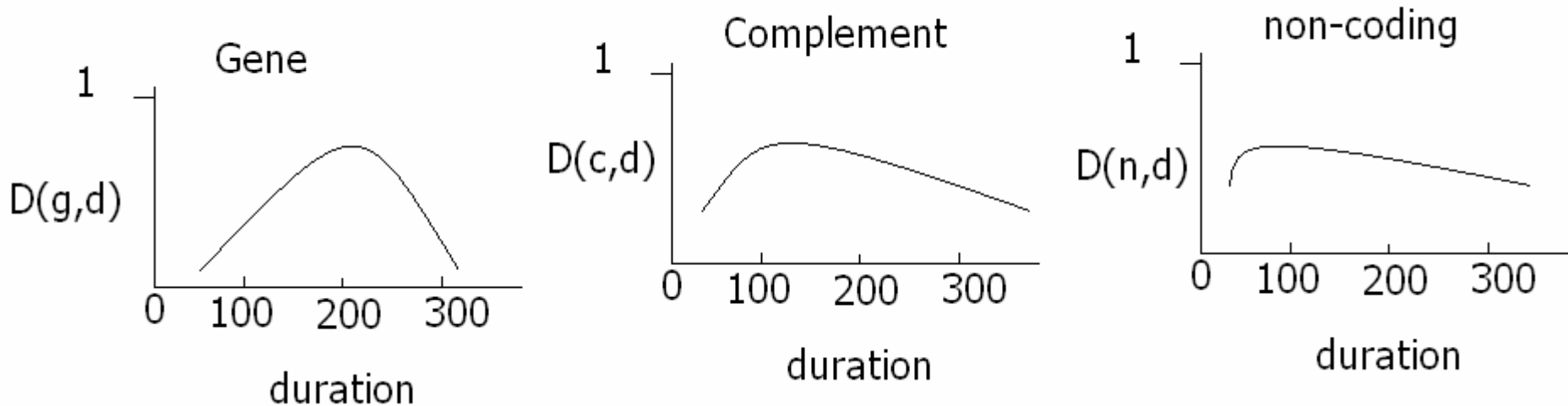
$$= N_{gc} / N_c$$

$N_c$ : Number of genes

$N_{gc}$ : Number times of a gene is followed by a complement

# Hidden Markov Model (HMM) with Duration

- $D(s,d)$ : Probability that state  $s$  has a duration of  $d$
- We just need find out the length distribution for gene, complement and non-coding.





# Hidden Markov Model (HMM) with Duration

---

- $E_s(x_{k-d}x_{k-d-1}\dots x_k)$ : Probability that state  $s$  generates  $x_{k-d}x_{k-d-1}\dots x_k$
- This is the probability that a short sequence is generated by gene, complement or non-coding state.
- This can be calculated using the homogenous or non-homogenous Markov chains we introduced in the beginning of the class.