Subsequence Pattern Models

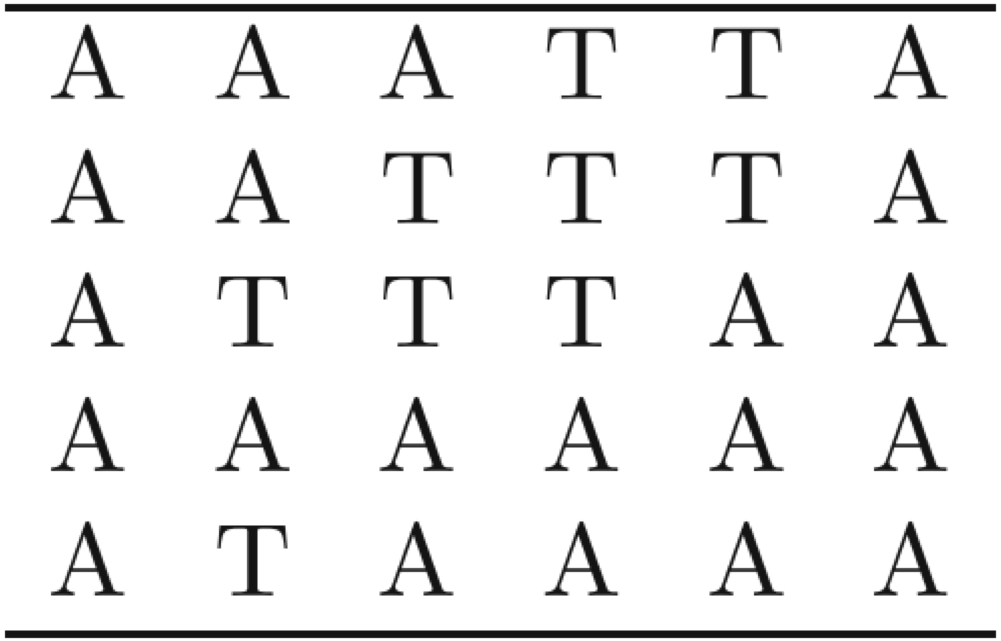
(Section 5)

Prepared By :

T.A / Mina E. Karam

Position Dependent Markov Model

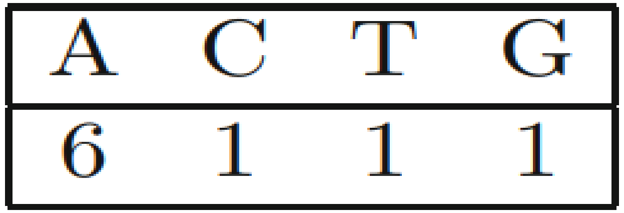
- Example 1: Using position dependent markov model calculate the probability of matching the subsequence ATTCAT on the following MSA:



A	A	A	T	T	A
A	A	T	T	T	A
A	T	T	T	A	A
A	A	A	A	A	A
A	T	A	A	A	A

Position Dependent Markov Model

1. Claculate the frequencies for position one. (Using Laplace Rule)



A	C	T	G
6	1	1	1

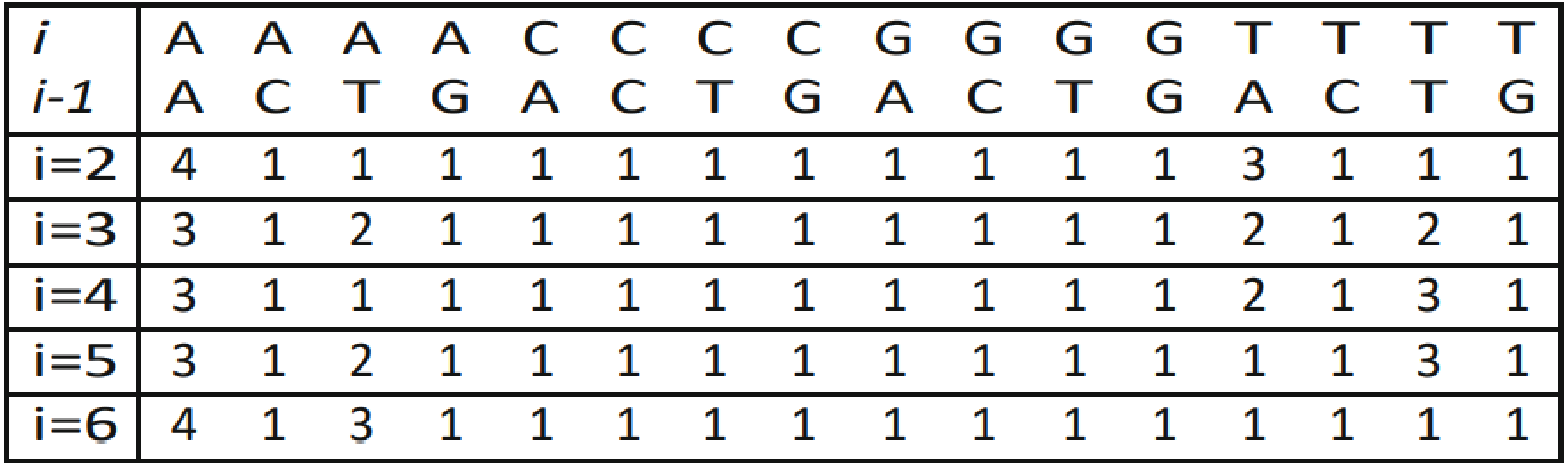
- So the probabilities of position one are ($P_1(A) = 6/9$, $P_1(C) = 1/9$, $P_1(G) = 1/9$, $P_1(T) = 1/9$)

Position Dependent Markov Model

2. Claculate the frequencies for positions from 2 to N. (Using Laplace Rule)

- We will create a table that includes the frequencies of conditional dimers involved during each position in the MSA from the second position (the second column) to the last position (the last column).
- We will then go through each column separately, starting from the second column to the last column and each time:
  - Select the column representing the position (i) and the column preceding it (i - 1).
  - Count the frequency of the included conditional dimers across the target columns i , i – 1.
  - Then write the frequency in the cell resulting from the intersection of the row expressed in the target position i with the column expressed in the conditional dimer (i | i-1) .

Position Dependent Markov Model



i	A	A	A	A	C	C	C	C	G	G	G	G	T	T	T	T
i-1	A	C	T	G	A	C	T	G	A	C	T	G	A	C	T	G
i=2	4	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1
i=3	3	1	2	1	1	1	1	1	1	1	1	1	2	1	2	1
i=4	3	1	1	1	1	1	1	1	1	1	1	1	2	1	3	1
i=5	3	1	2	1	1	1	1	1	1	1	1	1	1	1	3	1
i=6	4	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1

- The frequency values listed in the above table may be converted to probability values by multiplying it with (1/21).
- Note that the number 21 indicates the sum of the frequencies in each row.

Position Dependent Markov Model

3. Calculate the probability of matching the subsequence ATTCAT using the following rule:

$$P(x|M) = P_1(x_1) \prod_{i=2,\ldots,n} P_i(x_i|x_{i-1})$$

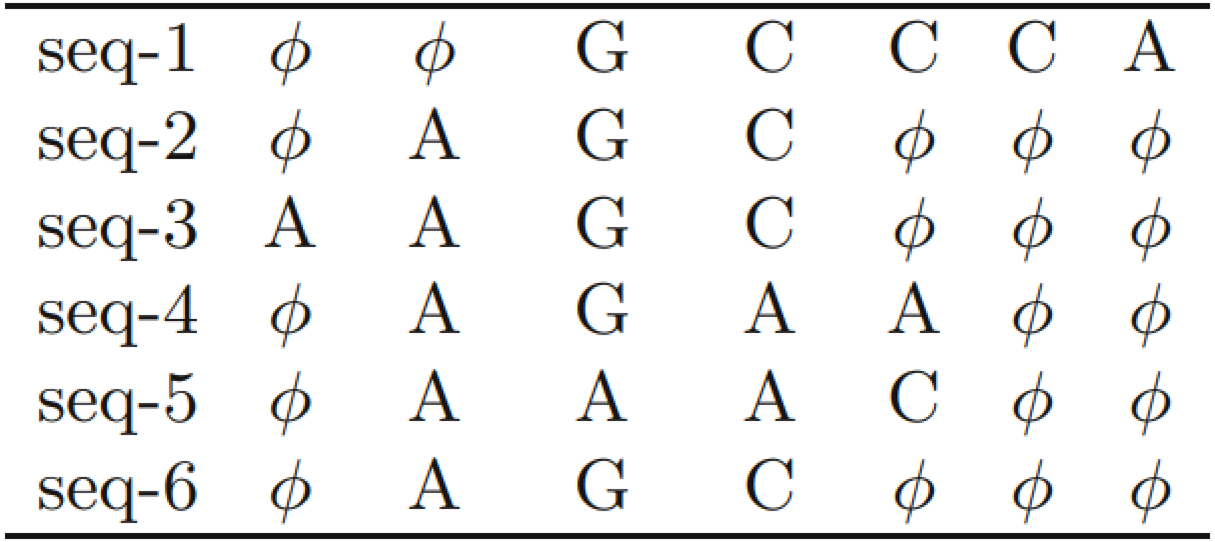
- P(ATTCAT | M) = $P_1(A) * P_2(T | A) * P_3(T | T) * P_4(C | T) * P_5(A | C) * P_6(T | A)$

  6 / 9 * 3 / 21 * 2 / 21 * 1 / 21 * 1 / 21 * 1 / 21 = 9.79×10^{-7}

Hidden Markov Model (HMM)

- HMM-profile is parameterized using λ = {π , A, B}
  - π ------ The initial state distribution.
  - A ------ Matrix of the transition probabilities.
  - B ------ Matrix of the emission probabilities.

- Example 1 : Construct HMM-profile for the following MSA.

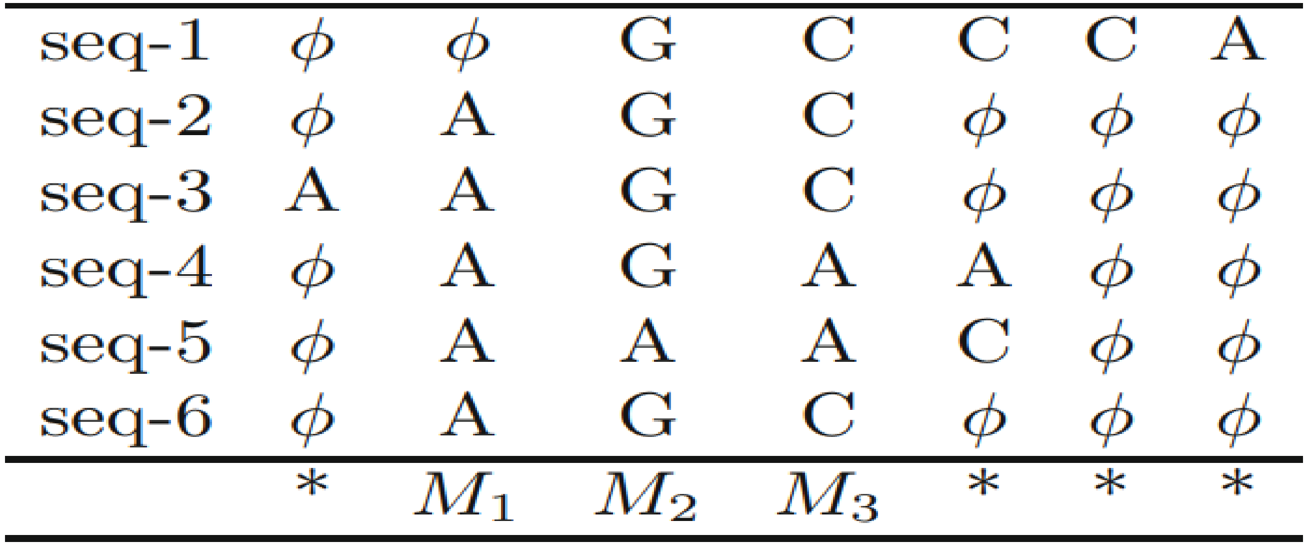


seq-1	ϕ	ϕ	G	C	C	C	A
seq-2	ϕ	A	G	C	ϕ	ϕ	ϕ
seq-3	A	A	G	C	ϕ	ϕ	ϕ
seq-4	ϕ	A	G	A	A	ϕ	ϕ
seq-5	ϕ	A	A	A	C	ϕ	ϕ
seq-6	ϕ	A	G	C	ϕ	ϕ	ϕ

Hidden Markov Model (HMM)

1. Create the consensus sequence (Model Topology):

- Go through the columns of the MSA sequentially and through each column determine whether the dominant element is a letter (A, C, G or T) or a gap.
- If there is a letter that is the highest frequency of 50% or more of the column's frequencies, put the letter M_i in front of the column in consensus sequence.(i starts with 1 and increases with 1 with each column state M).
- If the gap is the highest frequency during the column, put * in front of the column in consensus sequence.



seq-1	ϕ	ϕ	G	C	C	C	A
seq-2	ϕ	A	G	C	ϕ	ϕ	ϕ
seq-3	A	A	G	C	ϕ	ϕ	ϕ
seq-4	ϕ	A	G	A	A	ϕ	ϕ
seq-5	ϕ	A	A	A	C	ϕ	ϕ
seq-6	ϕ	A	G	C	ϕ	ϕ	ϕ
	*	M_1	M_2	M_3	*	*	*

Hidden Markov Model (HMM)

2. Create the initial state distribution (π):

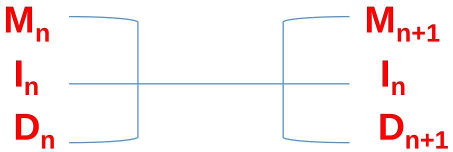
- Go through each sequence in the MSA in order, so that you will represent the sequence using a code statement so that the code statement is generated using 3 units expressing the operations involved in the sequence (M → Matching , I→ Insertion , D→ Deletion).

- Each code statement begins with M_0 and ends with M_{n+1} so that n is the last state mentioned in the consensus sequence. So all code statements in this example start with M_0 and end with M_4.

- We create the code statement by passing each element in the sequence and comparing it with the corresponding element in the consensus sequence so that if the element in the sequence is:
  - A gap and the corresponding element in the consensus sequence is a gap, **Do not add anything to the code statement .**
  - A gap and the corresponding element in the consensus sequence is M, **Add to the code statement D.**
  - A letter and the corresponding element in the consensus sequence is M, **Add to the code statement M.**
  - A letter and the corresponding element in the consensus sequence is a gap, **Add to the code statement I**.

Hidden Markov Model (HMM)

- Each element in the code statement has an index to express its state, so we will show the rule used to determine the index written under each element during the code statement:



M_n

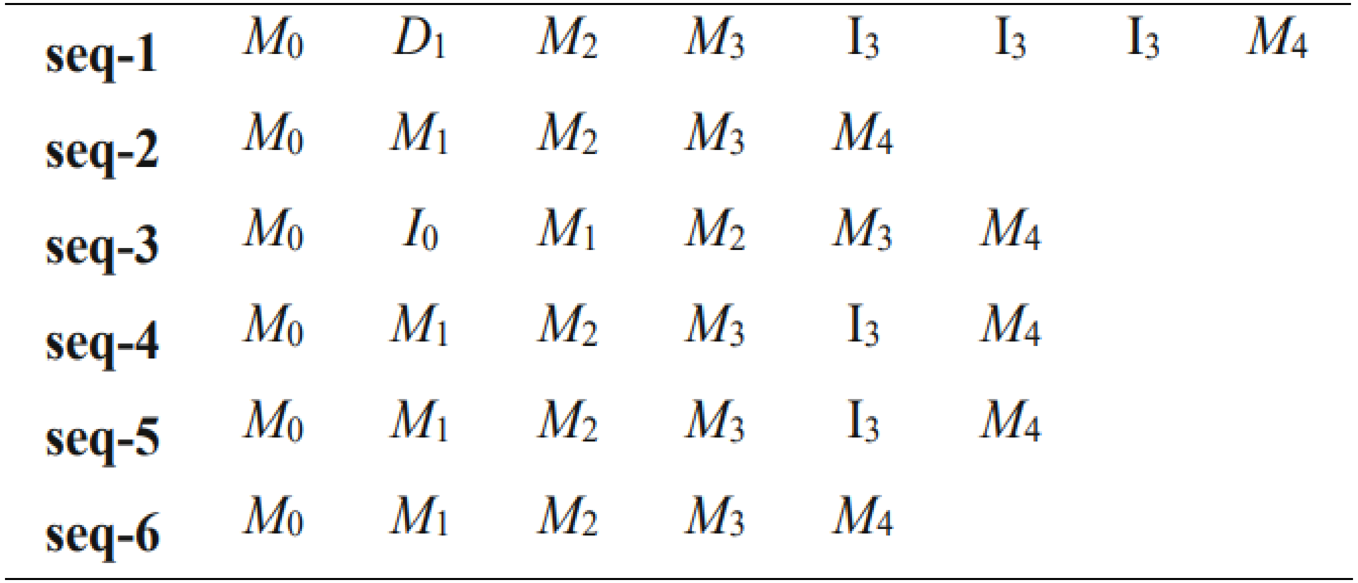
I_n

D_n

M_{n+1}

I_n

D_{n+1}



seq-1	M_0	D_1	M_2	M_3	I_3	I_3	I_3	M_4
seq-2	M_0	M_1	M_2	M_3	M_4			
seq-3	M_0	I_0	M_1	M_2	M_3	M_4		
seq-4	M_0	M_1	M_2	M_3	I_3	M_4		
seq-5	M_0	M_1	M_2	M_3	I_3	M_4		
seq-6	M_0	M_1	M_2	M_3	M_4			

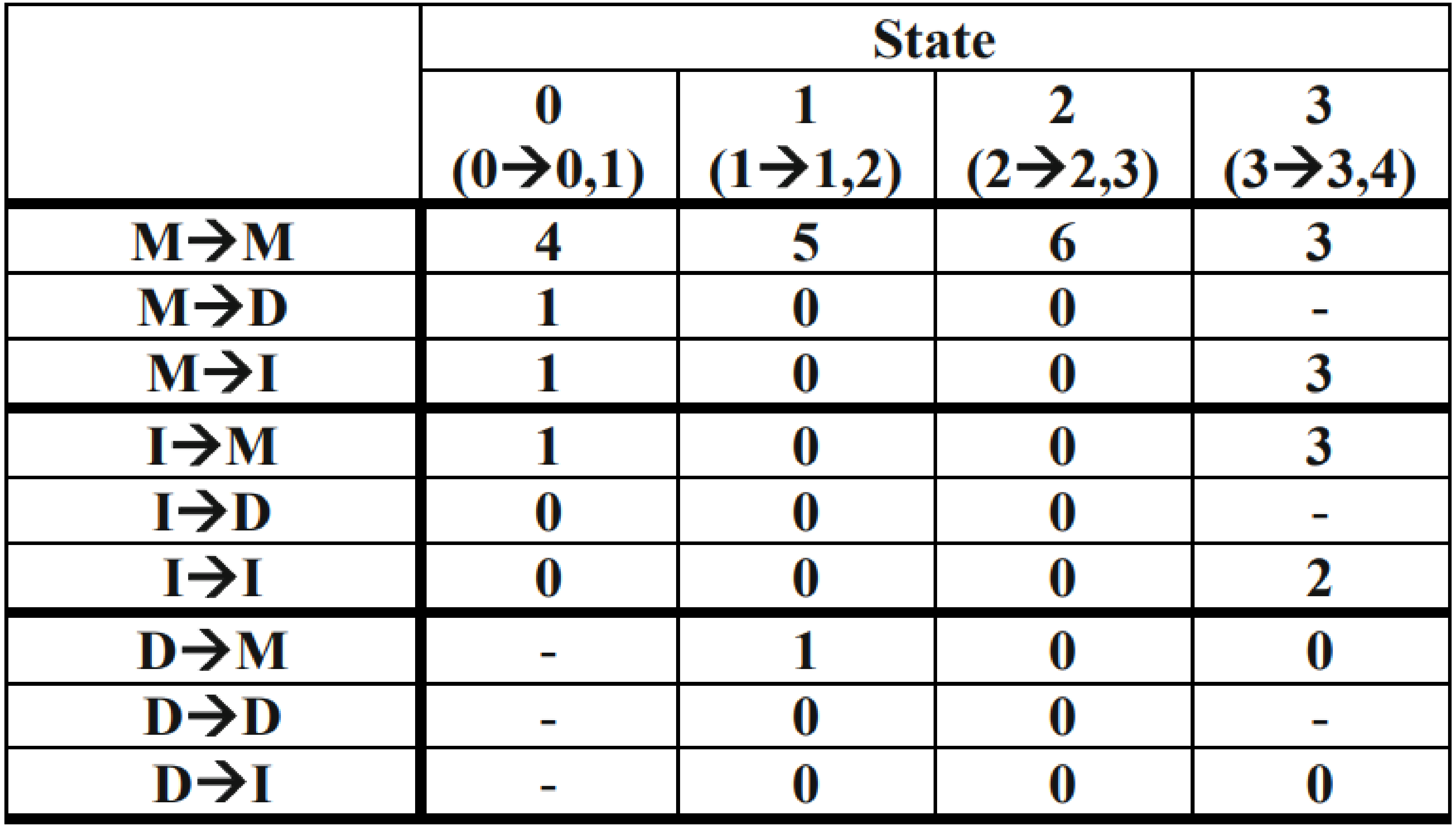
Hidden Markov Model (HMM)

3. Create a table that includes the transition frequencies:

- The columns of the table represent the basic states included during the consensus sequence 1, 2 and 3 in addition to the zero state so that each of them comes after a state and this fulfills the concept of transition, so we do not use the state included in the end point 4 because nothing comes after it.
- Table rows express the transition from each letter to another using permutations and combinations.

- **Notes that will help you to count frequencies faster:**
  - There are cells in the table that represent transitions that are impossible to occur:
    - In **column 0** all the cells that represent the transition from **D_0 → letter** because in the zero state there is no deletion process, So we will put in these cells **--** .
    - In the last column the transitions from **letter → D_i** so that the **i** is an index representing the state for the end point, So we will put in these cells **--** .
  - You must identify the letters of the **I , D** that are not present in the code table in order to put in the cells expressing the transitions that include these non-existent letters **0**.
  - After all of the above, there will be some cells left in the table. Learn about the transition represented in each cell. Then go to the code table in step 2 and count the frequency of the presence of this transition, then write it in the corresponding cell in the current table .

Hidden Markov Model (HMM)



	State			
	0 (0→0,1)	1 (1→1,2)	2 (2→2,3)	3 (3→3,4)
M→M	4	5	6	3
M→D	1	0	0	-
M→I	1	0	0	3
I→M	1	0	0	3
I→D	0	0	0	-
I→I	0	0	0	2
D→M	-	1	0	0
D→D	-	0	0	-
D→I	-	0	0	0

Hidden Markov Model (HMM)

4. Create a table that includes the transition probabilities:

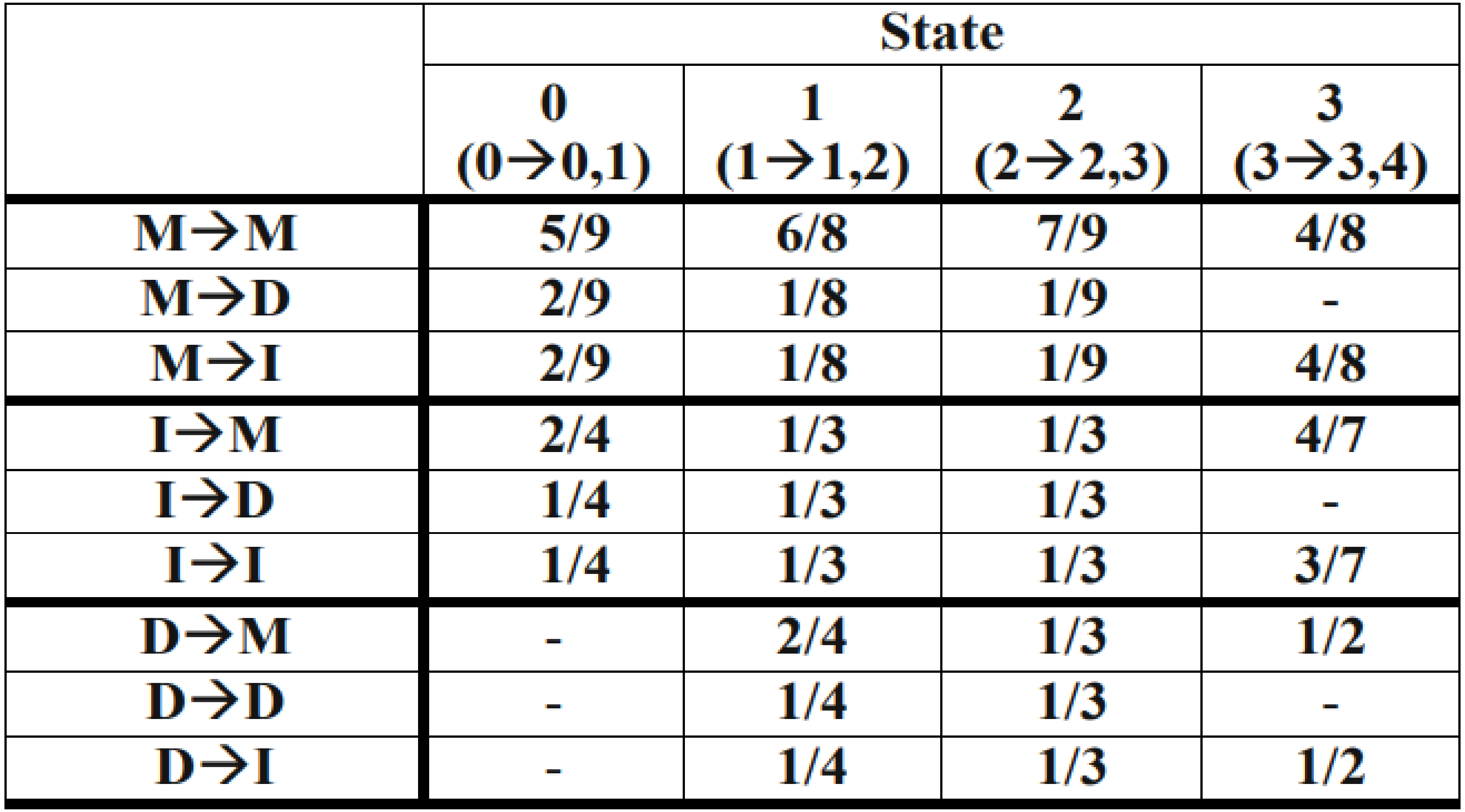
- From the previous table, it is clear that each column is divided into 3 parts, and each part is divided into 3 cells. So that each part expresses a process during a specific state, such as the first 3 cells in column 0 expressing M_0 transitions for all letters. ($M_0 \rightarrow M_1$, $M_0 \rightarrow I_0$, $M_0 \rightarrow D_1$) .

- We will create the same previous table and pass over each part in the transition frequency table (part means 3 cells expressing the same process during a specific state) and by applying **Laplace's rule** we will increase the frequency of each cell by 1, then divide by the sum of the cells included during the part. Then we write the resulting probabilities in the corresponding part of the new table.

- Transition frequency table cells that represent impossible transitions also do not include probabilities.

- So, for example, in the last column of the previous table, you will notice that each part consists of 3 cells, but the second cell is impossible to occur, so we do not include it in calculating the probabilities.

Hidden Markov Model (HMM)



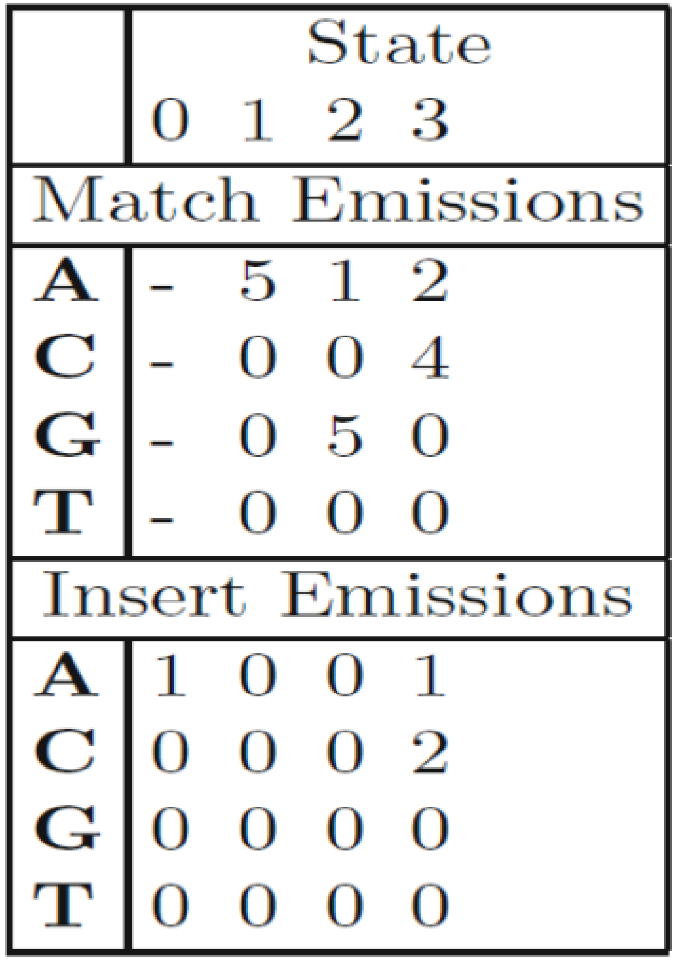
	State			
	0 (0→0,1)	1 (1→1,2)	2 (2→2,3)	3 (3→3,4)
M→M	5/9	6/8	7/9	4/8
M→D	2/9	1/8	1/9	-
M→I	2/9	1/8	1/9	4/8
I→M	2/4	1/3	1/3	4/7
I→D	1/4	1/3	1/3	-
I→I	1/4	1/3	1/3	3/7
D→M	-	2/4	1/3	1/2
D→D	-	1/4	1/3	-
D→I	-	1/4	1/3	1/2

Hidden Markov Model (HMM)

5. Create a table that includes the emission frequencies:

- The columns of the table represent the same cases mentioned in the table of transition frequencies.(0,1,2 and 3).
- There are two types of emissions such as match emissions and insert emissions, and they will be represented throughout the table, so we divide the table into two parts to express the two types.
- Each type of emissions will be represented in terms of its nucleotides (A,C,G and T) during each state in the table.
- To calculate the frequencies of match emissions go to step 1 and count the frequencies of each letter through the basic states of the match **(M_1 , M_2 , M_3)** and then write the resulting frequencies in the match emissions part (note that the match in the zero state **M_0** is not a process but just a starting point for the code statement, So we will put the symbol **--** in front of the zero state of the match emissions).
- To calculate the frequencies of insert emissions, it must be remembered that the insertion process is achieved when we compare a letter in the sequence to a gap in the consensus sequence, so we will go back to step 1 and look at the first column * and the fourth column * so that all the letters in the first column express **I_0** while all the letters in the fourth column expresses **I_3** because it comes after M_3, so during the two columns, count the frequencies of the letters represented by **I_0** and **I_3** and put their frequencies in the insert emissions part in front of the zero state and the third state (note that **I_1, I_2** are not mentioned, so we put in front of their 0).

Hidden Markov Model (HMM)

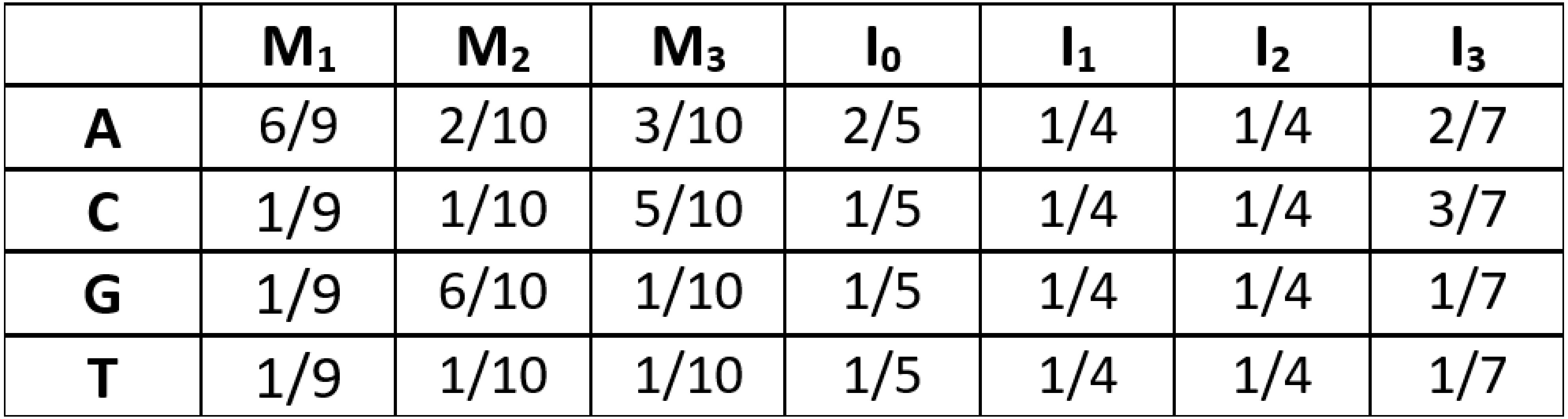


	State			
	0	1	2	3
Match Emissions				
A	-	5	1	2
C	-	0	0	4
G	-	0	5	0
T	-	0	0	0
Insert Emissions				
A	1	0	0	1
C	0	0	0	2
G	0	0	0	0
T	0	0	0	0

Hidden Markov Model (HMM)

6. Create a table that includes the emission probabilities:
    - In the previous table, we see that there are two types of emissions, and each type includes states expressed in terms of the four nucleotides (A, C, G, T). Every 4 cells have the same state during the same type of emissions is considered an independent part.
    - The probabilities for each part will be calculated independently according to **Laplace's rule** as mentioned earlier in calculating the transition probabilities.

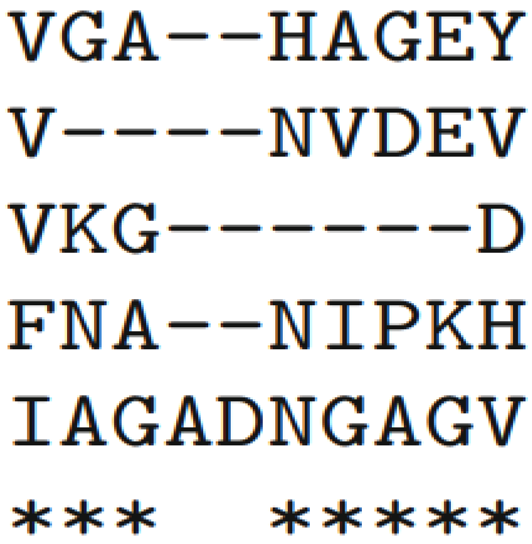
Hidden Markov Model (HMM)



	M_1	M_2	M_3	I_0	I_1	I_2	I_3
A	6/9	2/10	3/10	2/5	1/4	1/4	2/7
C	1/9	1/10	5/10	1/5	1/4	1/4	3/7
G	1/9	6/10	1/10	1/5	1/4	1/4	1/7
T	1/9	1/10	1/10	1/5	1/4	1/4	1/7

Hidden Markov Model (HMM)

- Example 2 : Consider the following protein sequence alignment, where * denotes the match state:



```
VGA--HAGEY
V----NVDEV
VKG------D
FNA--NIPKH
IAGADNGAGV
***  *****
```

a) Find the emission probabilities of match state 1.

b) Find the transition probabilities of match state 1.

Hidden Markov Model (HMM)

a) Solution:

- Emission frequencies are **freq(V) = 3 , freq(F) = 1, freq(I) = 1** and the frequencies of the other amino acids = 0 (17 amino acids not found).

- **So e_{M1}(V)= 4/25 , e_{M1}(F)= 2/25 , e_{M1}(I)= 2/25 and the emission probability for all other amino acids are 1/25.**

b) Solution:

- Transition frequencies are **freq($M_1 \rightarrow M_2$) = 4** , **freq($M_1 \rightarrow D_2$) = 1, freq($M_1 \rightarrow I_1$) = 0.**

- **So Transition probabilities are p($M_1 \rightarrow M_2$) = 5/8 , p($M_1 \rightarrow D_2$) = 2/8 , p($M_1 \rightarrow I_1$) = 1/8.**