

# Problem 3475: DNA Pattern Recognition

## Problem Information

Difficulty: Medium

Acceptance Rate: 0.00%

Paid Only: No

## Problem Description

Table:

Samples

+-----+-----+ | Column Name | Type | +-----+-----+ | sample\_id | int | |  
dna\_sequence | varchar | | species | varchar | +-----+-----+ sample\_id is the unique  
key for this table. Each row contains a DNA sequence represented as a string of characters  
(A, T, G, C) and the species it was collected from.

Biologists are studying basic patterns in DNA sequences. Write a solution to identify

sample\_id

with the following patterns:

Sequences that

start

with

ATG

(a common

start codon

)

Sequences that

end

with either

TAA

,

TAG

, or

TGA

(

stop codons

)

Sequences containing the motif

ATAT

(a simple repeated pattern)

Sequences that have

at least

3

consecutive

G

(like

GGG

or

GGGG

)

Return

the result table ordered by

sample\_id in

ascending

order

.

The result format is in the following example.

Example:

Input:

Samples table:

```
+-----+-----+-----+ | sample_id | dna_sequence | species |
+-----+-----+-----+ | 1 | ATGCTAGCTAGCTAA | Human | | 2 |
GGGTCAATCATC | Human | | 3 | ATATATCGTAGCTA | Human | | 4 | ATGGGGTCATCATAA
| Mouse | | 5 | TCAGTCAGTCAG | Mouse | | 6 | ATATCGCGCTAG | Zebrafish | | 7 |
CGTATGCGTCGTA | Zebrafish | +-----+-----+-----+
```

Output:

sample_id	dna_sequence	species	has_start	has_stop	has_atat	has_ggg
1	ATGCTAGCTAGCTAA	Human	1	1	0	0
2	GGGTCAATCATC	Human	0	0	0	1
3	ATATATCGTAGCTA	Human	0	0	1	0
4	ATGGGGTCATCATAA	Mouse	1	1	0	1
5	TCAGTCAGTCAG	Mouse	0	0	0	0
6	ATATCGCGCTAG	Zebrafish	0	1	1	0
7	CGTATGCGTCGTA	Zebrafish	0	0	0	0

Explanation:

Sample 1 (ATGCTAGCTAGCTAA):

Starts with ATG (has\_start = 1)

Ends with TAA (has\_stop = 1)

Does not contain ATAT (has\_atat = 0)

Does not contain at least 3 consecutive 'G's (has\_ggg = 0)

Sample 2 (GGGTCAATCATC):

Does not start with ATG (has\_start = 0)

Does not end with TAA, TAG, or TGA (has\_stop = 0)

Does not contain ATAT (has\_atat = 0)

Contains GGG (has\_ggg = 1)

Sample 3 (ATATATCGTAGCTA):

Does not start with ATG (has\_start = 0)

Does not end with TAA, TAG, or TGA (has\_stop = 0)

Contains ATAT (has\_atat = 1)

Does not contain at least 3 consecutive 'G's (has\_ggg = 0)

Sample 4 (ATGGGGTCATCATAA):

Starts with ATG (has\_start = 1)

Ends with TAA (has\_stop = 1)

Does not contain ATAT (has\_atat = 0)

Contains GGGG (has\_ggg = 1)

Sample 5 (TCAGTCAGTCAG):

Does not match any patterns (all fields = 0)

Sample 6 (ATATCGCGCTAG):

Does not start with ATG (has\_start = 0)

Ends with TAG (has\_stop = 1)

Starts with ATAT (has\_atat = 1)

Does not contain at least 3 consecutive 'G's (has\_ggg = 0)

Sample 7 (CGTATGCGTCGTA):

Does not start with ATG (has\_start = 0)

Does not end with TAA, "TAG", or "TGA" (has\_stop = 0)

Does not contain ATAT (has\_atat = 0)

Does not contain at least 3 consecutive 'G's (has\_ggg = 0)

Note:

The result is ordered by sample\_id in ascending order

For each pattern, 1 indicates the pattern is present and 0 indicates it is not present

## Code Snippets

### MySQL:

```
# Write your MySQL query statement below
```

### MS SQL Server:

```
/* Write your T-SQL query statement below */
```

### PostgreSQL:

```
-- Write your PostgreSQL query statement below
```

### Oracle:

```
/* Write your PL/SQL query statement below */
```

### Pandas:

```
import pandas as pd

def analyze_dna_patterns(samples: pd.DataFrame) -> pd.DataFrame:
```

## Solutions

### MySQL Solution:

```
# Write your MySQL query statement below
```

### MS SQL Server Solution:

```
/* Write your T-SQL query statement below */
```

### PostgreSQL Solution:

```
-- Write your PostgreSQL query statement below
```

**Oracle Solution:**

```
/* Write your PL/SQL query statement below */
```

**Pandas Solution:**

```
import pandas as pd

def analyze_dna_patterns(samples: pd.DataFrame) -> pd.DataFrame:
```