Course: BINF-6211-LAB-6
Name: Drashti Mehta, SID: 801262877

Part 1 (10pts).
If you look at the database schema for miRBase, you might notice several issues with it. For example, the attributes author and journal in the table literature_references are problematic. Please, think why and complete the following sentences:
The attribute author in the table literature_references is problematic because…
The attribute journal in the table literature_references is problematic because…
Other issues with the miRBase database schema include…

A1.
With literature_references table:

- **Author** column: There are so many names in author column and it can get difficult and confusing
- **Journal** column: We have journal name, page information and publication date in a single column which is like a chaos. Hence it can get easier if we separate it into different columns.

Another table: **mirna_species:**

- **Taxonomy** column: If we want to find a particular family or genus, it gets difficult. The whole taxonomical hierarchy is explained in a single column making it messy. We can separate it into different columns.

Part 2 (10pts).
In this lab, we'll fix the issue with the journal attribute. To fix the issue, you might need to add additional columns to the table literature_references. Think about what columns you need to add. When naming the table columns, try to be consistent with the naming conventions for other columns in miRBase.
Report SQL statements for adding columns to literature_references.

A2.
The main issue with the journal column was that when we performed a group by query for the literature_references table, it did not give proper grouped values. So, for that we make 2 columns, "journal_name" and "publication_year".

Query:
Alter table add column journal_name
Alter table add column publication_year

Part 3 (10pts).
Now, it's time to populate our new columns. You might need to use this SQL functions:
INSTR(string, substring) returns the position of substring in the given string.
Example: "SELECT INSTR('SQLite Tutorial', 'Tutorial') AS Position" returns 8, which is the position of the substring 'Tutorial' in the string 'SQLite Tutorial'.
SUBSTR(string, start, length) returns the substring of the string, of given length, starting at position start.
Example: "SELECT substr('SQLite substr', 1, 6);" return the substring 'SQLite'.

A3.

Course: BINF-6211-LAB-6
Name: Drashti Mehta, SID: 801262877

Populating publication year:
update literature_references set publication_year = substr(journal, -6, 4)

Publicating journal name:
update literature_references set journal_name = SUBSTR(journal,1,INSTR(journal,'.')-1)

Part 4 (5pts).
In the previous lab, you had to find the number of RNAs per journal. Try to do it again with the new structure of table literature_references. Did you get a different or the same result this time? Please, explain why.

A4:
Previous_lab (lab-5):

SELECT literature_references.journal, COUNT(DISTINCT mirna.sequence) FROM mirna JOIN mirna_literature_references JOIN literature_references on mirna.auto_mirna=mirna_literature_references.auto_mirna AND mirna_literature_references.auto_lit=literature_references.auto_lit GROUP BY literature_references.journal;
Result: **677** rows

This lab (lab - 6):

SELECT literature_references.journal_name, COUNT(DISTINCT mirna.sequence) FROM mirna JOIN mirna_literature_references JOIN literature_references on mirna.auto_mirna=mirna_literature_references.auto_mirna AND mirna_literature_references.auto_lit=literature_references.auto_lit GROUP BY literature_references.journal_name;
Result: **171** rows

The issues that we discussed in the above question about grouping the data, that got solved and thus we have lesser rows as the journal names column gives proper grouping for journals.