Graph Theory and its Application

Genome Analysis using Neo4j.

1) Adithya M, PES1UG20CS621,
Section: K
2) Rakesh Reddy, PES1UG20CS566,
Section: J

Introduction:

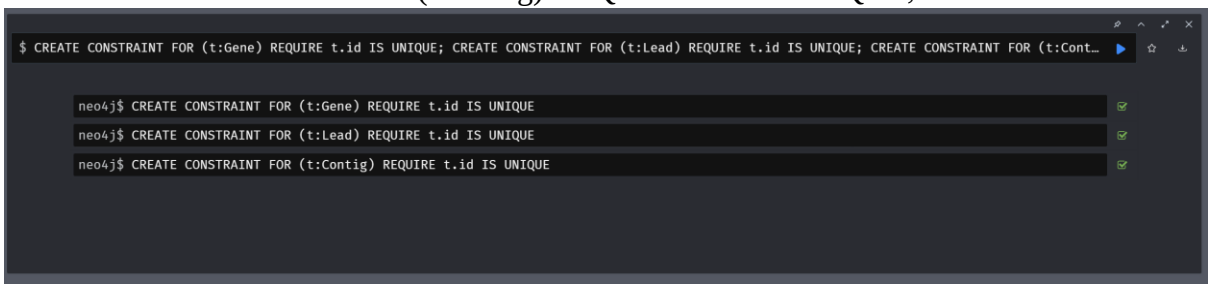
- A genome is the sum of all genetic materials inside an organism, be it a virus, a bacterium, or a human. With these genes, the cell can manufacture proteins to construct itself and fulfil various biochemical functions. Therefore, scientists can learn a lot about the organism by sequencing and analyzing its genes in its genome.

- For this analysis we use the complete chromosome of '*Halorhabdus tiamatea*' SARL4B by Werner et al. and the complete genome of '*Formosa agariphila'* KMM 3901 by Mann et al. in **JSON** format.

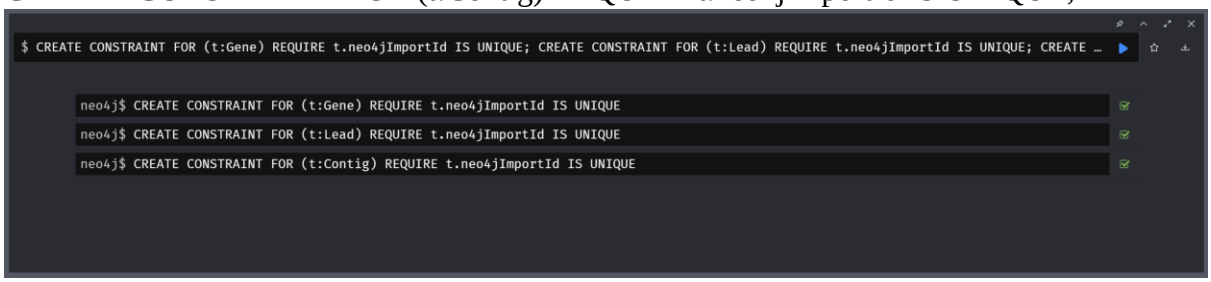
- Graph is a natural and intuitive way to represent the gene arrangement in genomes. Gene properties are stored as key-value pairs inside each node.

A) Importing dataset from JSON

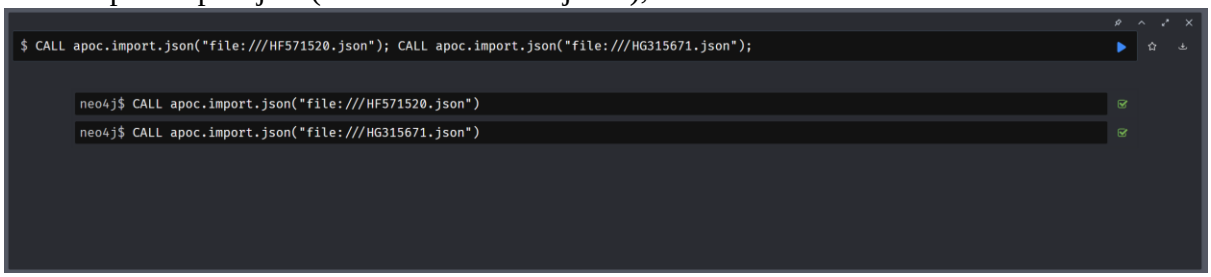
1) CREATE CONSTRAINT FOR (t:Gene) REQUIRE t.id IS UNIQUE;
   CREATE CONSTRAINT FOR (t:Lead) REQUIRE t.id IS UNIQUE;
   CREATE CONSTRAINT FOR (t:Contig) REQUIRE t.id IS UNIQUE;



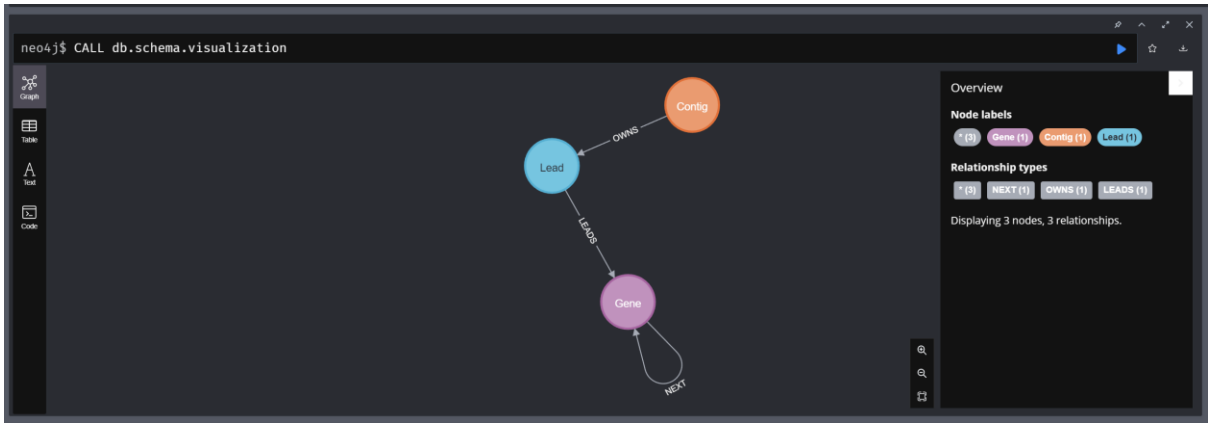
2) CREATE CONSTRAINT FOR (t:Gene) REQUIRE t.neo4jImportId IS UNIQUE;
   CREATE CONSTRAINT FOR (t:Lead) REQUIRE t.neo4jImportId IS UNIQUE;
   CREATE CONSTRAINT FOR (t:Contig) REQUIRE t.neo4jImportId IS UNIQUE;



3) CALL apoc.import.json("file:///HF571520.json");
   CALL apoc.import.json("file:///HG315671.json");



4) CALL db.schema.visualization



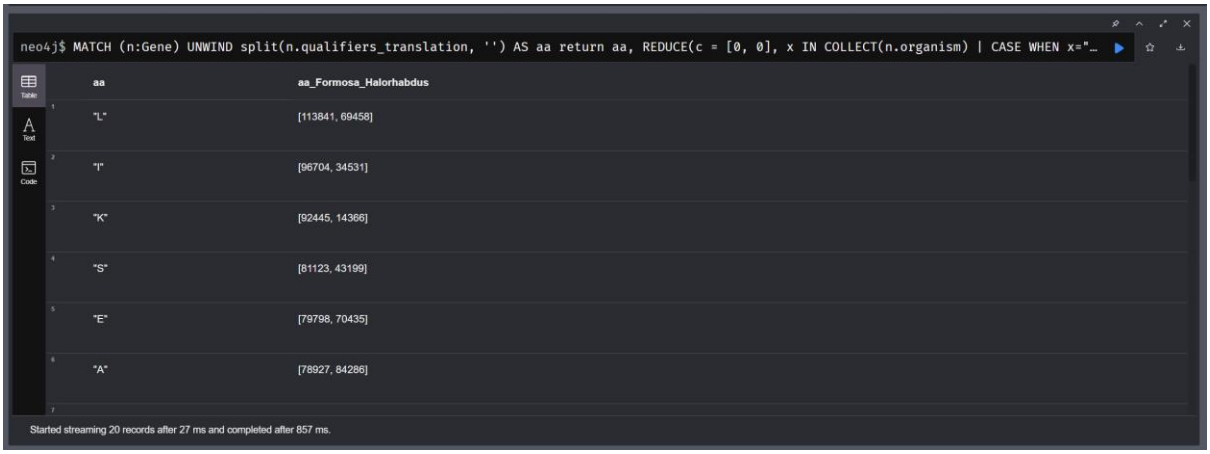
B) Neo4j Commands:

5) Count the amount of GH in the chromosome of *Halorhabdus tiamatea*:

```
MATCH (n:Gene)
WHERE n.organism='Halorhabdus tiamatea SARL4B'
  AND n.name  =~ '.+GH\d+.*'
RETURN COUNT(n.name)
```

6) Display the raw amino acid counts between the two genomes:

```
MATCH (n:Gene)
UNWIND split(n.qualifiers_translation, '') AS aa
return aa,
REDUCE(c = [0, 0], x IN COLLECT(n.organism) | CASE WHEN
x="Formosa agariphila KMM 3901" THEN [c[0]+1, c[1]] ELSE [c[0],
c[1]+1] END) AS aa_Formosa_Halorhabdus
ORDER BY aa_Formosa_Halorhabdus DESC
```



7) Return gene clusters that are bookended by a SusD-like protein and a TonB-dependent receptor as long as the fragment is shorter than 10,000 base pairs in *Formosa agariphila*:

```
MATCH p=(g1:Gene) -[:NEXT*] ->(g2:Gene)
WHERE g1.organism='Formosa agariphila KMM 3901'
  AND g1.name='SusD-like protein'
  AND g2.organism='Formosa agariphila KMM 3901'
```

```
 AND g2.name='TonB-dependent receptor'
 AND apoc.coll.max([g1.location_end, g2.location_end]) -
apoc.coll.min([g1.location_start, g2.location_start]) < 10000
RETURN p;
```

8) Return all the GH16-centric gene clusters with five neighbors on each side in *F. agariphila*:

```
MATCH p=(f0:Gene) -[:NEXT*5] -> (f1:Gene) -[:NEXT*5] ->(f2:Gene)
WHERE f1.name =~ '.+GH16[^a-zA-Z\d\s:]*' AND f0.organism='Formosa agariphila
KMM 3901'
RETURN p;
```

C) Analysis:

1) Amount of GH in chromosome of *Halorhabdus tiamatea*: 42 which is in agreement with the number reported by Werner et al..

2) The two branched-chain amino acids leucine (L) and isoleucine (I) are more frequent in *Formosa* than in *Halorhabdus*. The two genomes have something in common: the two sulfuric amino acids: methionine (M) and cysteine (C) appear at the bottom of the list for both.

3) There are 18 clusters with a TonB-dependent receptor and SusD-like protein. This gives more insight into the polysaccharide degradation ability of the bacterium.

4) There are 5 such 'Polysaccharide Utilization Loci' or PUL.

   a. Three of them have exactly eleven genes.
   b. The fourth has twelve.
   c. And a super PUL with 35 genes is revealed.

D) Conclusion:

1) This project demonstrates how Neo4j could be used to browse various genomes and perform relatively tough queries and visualize the same.

2) This could be further expanded to store and manipulate metagenomes with a similar design and thus accelerate new discoveries in genomics and metagenomics.

3) This design also allows to export to EMBL format for further study of the genome.