Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220696

Analyzing SARS-CoV-2 Sequence Patterns by Semantic Trajectories

Wissame LADDADA[a,1], Cecilia ZANNI-MERK[a] and Lina F. SOUALMIA[a]
[a]*Normandie Univ, UNIROUEN, INSAR, LITIS UR 4108, F-76000 Rouen, France*

Abstract. Since the beginning of the pandemic due to the SARS-CoV-2 emergence, several variants has been observed all over the world. One of the last known, Omicron, caused a large spread of the virus in few days, and several countries reached a record number of contaminations. Indeed, the mutation in the Spike region of the virus played an important role in altering its behavior. Therefore, it is important to understand the virus evolution by extracting and analyzing the virus structure of each variant. In this work we show how patterns sequence could be analyzed and extracted by means of semantic trajectories modeling. To do so, we designed a graph-based model in which the genome organization is handled using nodes and edges to represent respectively the nucleotides and sequence connection (point of interest and routes for trajectories). The modeling choices and pattern extraction from the graph allowed to retrieve a region where a mutation occurred in Omicron (NCBI version:OM011974.1).

Keywords. SARS-CoV-2, Pattern analysis, Graph, Neo4j, Cypher

1. Introduction

The SARS-CoV-2 is a positive single-stranded RNA genome (*30 kilobases*) that gathers several proteins. The genome structure can be illustrated as a sequence of nucleotide (...AUG...), amino acids (...MFV...), or proteins (...S, ORF3a, E...), where each amino acid is identified by a combination of three nucleotides. Genomes are dynamic entities that change over time as a result of the cumulative effects of small-scale sequence alterations caused by mutation [2]. A mutation is a change in the nucleotide sequence of a short region of a genome by the substitution, insertion, or deletion of one or a few nucleotides [2]. These genome mutations may occur during the replication process. This was designed and simulated in a previous work through an ontological-based approach [3], and then by combining the ontology with discrete event system specification (DEVS) modeling [1]. However, apart from the replication process, an interesting way to understand genomes evolution is to extract and analyze the sequences by means of patterns. By considering the link chain of these sequences, we propose to use a graph-based model in order to design the virus organization, and as an effective approach to analyze and detect mutations. Some work has proven the effectiveness of graph databases when the problem to address is to analyze interconnected biomedical data [6]. For example, the approach presented in [5] designs RNA sequencing data using the Neo4j graph database.

[1] Wissame Laddada, Laboratoire LITIS, Université de Rouen Normandie, UFR Sciences et Techniques, Avenue de l'université, 76800 Saint-Étienne-du-Rouvray, E-mail: wissame.laddada@univ-rouen.fr.

We present in this study a graph-based model to design the original genome structure of SARS-CoV-2 at the micro-level, by considering the nucleotides sequence. To test our approach, the graph database is enriched with a region of the genome (the full genome sequence is expected in a future work) to compare it with the first genome of SARS-CoV-2. To do so, we use the abstraction modeling of semantic trajectories: the genome either follows or not the same path (same point of interest of a trajectory) that designs the original sequence. In this first study, the deletion and the insertion mutations were not taken into account, we addressed only the substitution mutation.

2. Data Graph Modeling

When a mutation occurs in a genome, a short region changes during the replication process due to the substitution, deletion, or insertion of amino acids sequence in a macro-level and of nucleotides sequence in a micro-level. When the replication machinery terminates without errors, the generated genome sequence follows the same pattern as the original virus genome, i.e., if the original genome sequence is a trajectory where the points of interest are nucleotides and their connection are routes, then the newly replicated genome should pass through the same point of interest (i.e., the same nucleotide at the same rank) as the original genome. If the replication takes a different path, we consider this alternative as a mutation.

In our case, we take into account the micro-level analysis by designing the nucleotides sequence as a graph.

2.1. Graph Formalization

Let $G<(N),(E)>$ be a graph, where N and E represent respectively a set of nodes and edges. Each node N_i can be typed by several labels ($V_1...V_n$) to classify it. Also, for better description, each node is defined by properties P_j, that are common to all the labels $V_1...V_n$. Hence, the global definition of a node is N_i= ($V_1|V_n$ {$P_1...P_k$}). Likewise, each edge can be associated with a set of properties Q_j. However, an edge E_i must have only one label R to type the relation. The set of edges is therefore defined by E_i=(R{$Q_1...Q_m$}). Considering our domain of interest, the node represents the nucleotides N= (*Nucleotide*{*Nucleobase, Rank*}) where *Nucleotide* is the label and the set of properties *{Nucleobase, Rank}* represents respectively the name of the nucleotide (*U, A, G, C*) and the rank of each nucleotide. The edges represent the chain links between nucleotides. We define hence the set of edges E= (*HAS_NEXT*{*Virus_name*}), where *HAS_NEXT* represents edge's type, and the *Virus_name* property represents the genome that follows the path specified by the relation.

2.2. Data Integration in Neo4j

We have chosen the Neo4j graph database to integrate our graph. It is one of the world's leading open-source NoSQL (Not only SQL) Graph Database Management Systems (GDBMS) [4] (https://db-engines.com/en/ranking). The data semantics can be enhanced by using properties on edges which is an option handled by Neo4j and not in other representation systems (such as RDF (Resource Description Framework) triples). Cypher is the search engine. Figure 1, represents the obtained graph database structure.

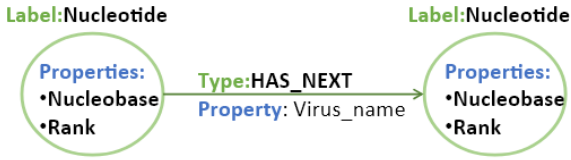

Figure 1. Graph model for genome patterns represented in the graph database Neo4j.

3. Retrieving Omicron Genome Mutation

Figure 3, represents the data graph enrichment with the mutation region S371L in the Spike protein S. It illustrates the Omicron (M011974.1) genome mutation extraction pattern (substitution mutation) by comparing it to the SARS-CoV-2 (NC_045512.2), and Delta (MZ724414.1). The extraction of such a substitution pattern (without considering the insertion and deletion mutation) is handled by adding to each node a *weight* property which represents the degree of incoming relations to the node. There are three genomes to be compared in this case study. If the weight value of a node is less than three, then this node represents a substitution mutation. The amino acid is a combination of three nucleotides. Hence, the mutation represents a pattern of three nodes, and the rank of the third node is a multiple of three. Considering the data enrichment with the protein S of the three genomes, the Cypher query and results (Figure 2) allow extracting some of the region mutations and the corresponding nucleotides sequence. For instance, the mutation K417N means that the Lysine (i.e., k) amino acid has been substituted to the Asparagine (i.e., N) amino acid. In our results, the Asparagine amino acid substitution is represented by the pattern "a", "a", "u".

```
MATCH p=(n1)-[r1]→(n2)-[r2]→(n3)
WHERE (n3.Weight<3 OR n2.Weight<3 OR n1.Weight<3)
AND (n3.Rang)%3=0 AND r1.Virus_Name='OM011974.1'
AND r2.Virus_Name='OM011974.1'
RETURN DISTINCT  n3.Rang/3 AS MUTATIONS,
n1.Nucleobase, n2.Nucleobase, n3.Nucleobase
ORDER BY MUTATIONS ASC
```

"MUTATIONS"	"n1.Nucleobase"	"n2.Nucleobase"	"n3.Nucleobase"
371	"c"	"u"	"c"
373	"c"	"c"	"a"
375	"u"	"u"	"c"
417	"a"	"a"	"u"

Figure 2. Patterns extraction of Omicron mutation regions and their corresponding nucleotides in the S protein allowed by a Cypher query on the data graph.

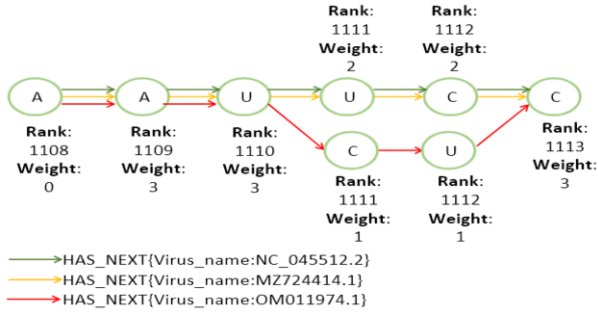

Figure 3. Graph data enrichment with the S371L region mutation in the spike.

4. Discussion and Conclusions

The graph-oriented approach presented in this study allowed to extract and analyze sequence patterns of a genome. These first results showed that this approach is effective to extract genome mutations caused by nucleotides substitution.

As a future work, a deeper analysis will be conducted at the macro-level (amino acids sequences) for a better abstraction of the mutation. We also planned to generalize our approach by adding other nodes classification: e.g., a node labeled as an amino acid will be connected to its ending nucleotide to mark its nucleobase sequence. Our graph can also be enriched with a huge volume of sequences from the GISAID (Global Initiative on Sharing Avian Influenza Data) database (https://www.gisaid.org) since the Neo4j allows the management of millions of relations and nodes. Furthermore, Neo4j has a bunch of useful data science algorithms (Centrality, Community Detection, Similarity, etc.) that will be applied to our graph model to extract more semantics for sequence patterns. Moreover, since the genome organization was described through an ontology(https://bioportal.bioontology.org/ontologies/ONTOREPLICOV) [2] an hybrid approach, that combines semantic web technologies with an expressive graph database, would be of a great interest for a better semantic expressiveness.

References

[1] Ayadi A, Frydman C, Laddada W, Soualmia LF, Zanni-Merk C, L'Hote I, Grellet E, Imbert I. Combining devs and semantic technologies for modeling the sars-cov-2 replication machinery, In Annual Modeling and Simulation Conference (ANNSIM); 2021p. 1–12.

[2] Brown TA, Genomes. 2nd ed. Oxford : New York : Wiley-Liss, 2002. 572 p.

[3] Laddada W, Soualmia LF, Zanni-Merk C, Ayadi A, Frydman C, L'Hote I, Imbert I., Ontoreplicov: an ontology-based approach for modeling the sars-cov-2 replication process. In the 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES);2021; Poland: Szczecin, Procedia Computer Science c2021. p.487–496.

[4] Pokorný J. Graph Databases: Their Power and Limitations. In the IFIP International conference on Computer Information Systems and Industrial Management (CISIM);2015; Lecture Notes in Computer Sience, Springer, Cham. c2015. p.58-69.

[5] Simpson CM, Gnad F. Applying graph database technology for analyzing perturbed co-expression networks in cancer. Database (Oxford). 2020.

[6] Timón-Reina S, Rincón M, Martínez-Tomás R. An overview of graph databases and their applications in the biomedical domain. Database (Oxford). 2021.