A data-driven approach for the analysis of SARS-CoV-2 mutations

Samuele Del Bello, Stefano Destro, Alessio Frezza, Rita Numeroli, Carlos Santillán
Supervisor: Pietro Pinoli PhD

Politecnico di Milano
Applied Statistics Project
Mathematical and Computer Science Engineering

Abstract

We consider the problem of extracting meaningful mutation patterns from a SARS-CoV-2 mutation dataset by exploiting the large quantity of data available. In particular, we focus on identifying synergistic mutation pairs (or groups) via an association rule mining approach. This study also explores the main characteristics of the most significant lineages and their relationships. Finally, we show that the process of clade discovery can be aided or partially automated via hierarchical clustering techniques.

Introduction

The wide availability of SARS-CoV-2 tests has provided us with large amounts of data on the virus' amino acid mutations. We study a dataset from Nextstrain, a project that collects data on viral outbreaks. At first, we aimed to find synergistic mutation pairs and perform clustering to find well-defined groups within the lineages; as we explored the data, the connections between these two objectives became clear and merged into one.

Overview of the dataset

The dataset contained 4 million samples, with columns relative to the species of the subject, country, date, mutations w.r.t. the original Wuhan sequence, sequence length, lineage, and clade. Key terms:

- **Clade:** Group of organisms that share a common ancestor. Can be seen as a branch in a phylogenetic tree.
- **Lineage:** A single line of descent or linear chain in the phylogenetic tree.
- **Amino acid mutation:** A change in the proteins of the sample w.r.t. the original Wuhan sequence. If in protein S the 614th amino acid changed from D to G, the mutation would be labeled as S:D614G.

In total, we have 1397 lineages and 25 clades. Figure 1 shows the lineage distribution w.r.t. the number of samples

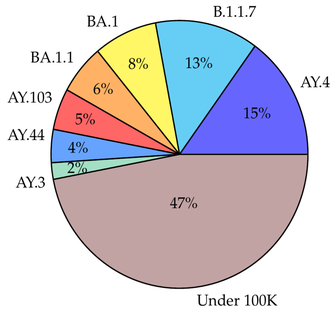

Figure 1: Lineage distribution w.r.t. number of samples

Main objectives

- Identification of mutation patterns within single lineages and association rules
- Clade discovery

Methods: Association Rule Mining

Consider each sample as a "transaction" that contains certain items (i.e., mutations). We provide some definitions:

- **Association rule:** An expression $X \to Y$, where X and Y are disjoint mutation subsets[1].
- **Support:** The support of a subset of mutations X, denoted by $sup(X)$, is the number of transactions (or samples) in a transaction dataset $\mathbf{D}$ that contain X.
- **Confidence:** The confidence of a rule is an estimate of the conditional probability that a transaction contains Y given that it contains X.

$$conf(X \to Y) = P(X|Y) = \frac{P(X \wedge Y)}{P(X)} = \frac{sup(X \cup Y)}{sup(X)}$$

Confidence can be a misleading metric in some cases, it is mostly used for filtering very unlikely rules, but not for assessing their quality.

- **Lift:** Ratio of the observed joint probability of X and Y to the expected joint probability if they were statistically independent.

$$lift(X \to Y) = \frac{P(X \cup Y)}{P(X) \cdot P(Y)} = \frac{conf(X \to Y) \cdot |\mathbf{D}|}{sup(X \cup Y)}$$

Pattern recognition within single lineages

We decided to study subsets of individual lineages. It is important to note that we performed a repeated sampling of 10K transactions, and the results are consistent throughout all samplings, except for negligible differences. We performed hierarchical clustering on all the features with Ward linkage and chose an appropriate number of clusters. Figure 2 shows the pairs plot (first four PCs) colored by clusters and the dendrogram for AY.103.

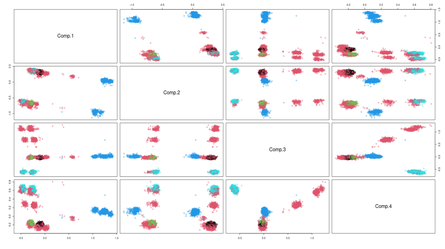

Figure 2: Pairs plot colored by cluster. Clustering has captured what we can assess by eye, although the data is high-dimensional

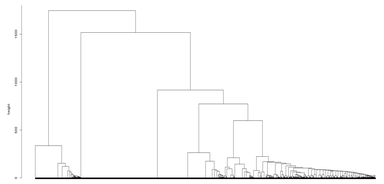

Figure 3: Dendrogram for AY.103 with Ward linkage. Choosing five balanced clusters we suspect there might be subpatterns within the lineage

Figure 4 shows a comparison of the mutation frequency barplots between AY.103 and one of its clusters.

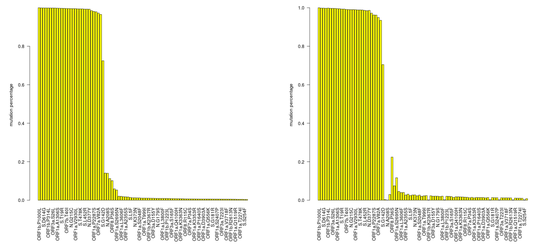

Figure 4: Original AY.103 mutation frequency barplot (left) and cluster 1 barplot (right). The mutation spikes in cluster 1 allow us to focus our attention on possibly meaningful rules

The spikes in mutations {*ORF81a:S2695N*} and {*ORF8:P36S*} suggest the rule {*ORF81a:S2695N*}→{*ORF8:P36S*}. After rule mining, we find such a rule with high confidence (0.989) and high lift (10.335). So we have exploited the information obtained via clustering and redirected our attention to potentially meaningful rules. The biological interpretation is that the mutations may be synergistic or beneficial to the virus or, more importantly, predominant in lineages generated or related to AY.103. Pairs plot colored by cluster. The three clusters found have an accurate correspondence with the clades

Clade discovery

Can we apply the previous pipeline to clade discovery? We consider five lineages (belonging to three clades) and randomly extract 1600 samples from each. Table 1 shows the lineages and their corresponding clades.

Lineages	Clade
AY.103, AY.4	21J (Delta)
B.1.1.7	20I (Alpha)
BA.1, BA.1.1	21K (Omicron)

Table 1: Lineages - clade table

We perform clustering, and the dendrogram suggests three balanced clusters. Figure 5 shows the pairs plot (first four PCs, over 92% of variability).

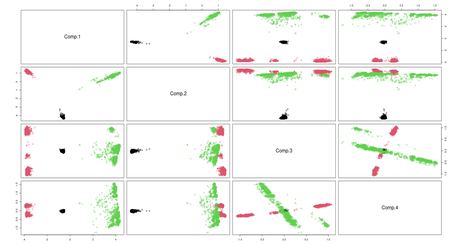

Figure 5: Pairs plot colored by cluster. The three clusters found have an almost exact correspondence with the clades

Our clustering has an almost perfect correspondence with the real clades (the coloring by clade shows an almost identical plot). Therefore, it is reasonable to assume that with more sophisticated techniques, we could apply this pipeline to lineages, which are more fine-grained than clades.

Conclusion

We have shown that standard techniques such as clustering and association rules can successfully exploit the large amount of data available. For example, smaller datasets, the case with any other virus, would have yielded poor results because of the lack of large-scale random sampling. We found non-trivial mutation patterns because of the reduction of the rule search space via the information provided by clustering. However, these results also have a biological interpretation; mutations tend to appear in pairs, possibly because of synergy between them (i.e., beneficial effect on the virus). Finally, our analysis suggests a data-driven lineage discovery method since clade discovery was successful.

Forthcoming Research

Future research focuses on predicting novel sublineages by studying the mutation pairs (or groups) found via association rules. Another unexplored aspect of the data is their functional nature; each sample has the date of sampling, and the analysis of the growth rates of single mutations within lineages may aid in the study of the evolution of new sublineages, such would be the case of the latest Omicron variant, lineage BA.5.

References

[1] Agrawal, Rakesh, Imieliński, and Swami. "Mining association rules between sets of items in large databases." In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp. 207-216. 1993.