CoV2K: a Knowledge Base of SARS-CoV-2 Variant Impacts

Ruba Al Khalaf[0000−0002−5645−5886], Tommaso Alfonsi[0000−0002−4707−8425], Stefano Ceri[0000−0003−0671−2415], and Anna Bernasconi[0000−0001−8016−5750]

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Ponzio 34/5, 20133, Milan, Italy
{ruba.al,tommaso.alfonsi,stefano.ceri,anna.bernasconi}@polimi.it

Abstract. In spite of the current relevance of the topic, there is no universally recognized knowledge base about SARS-CoV-2 variants; viral sequences deposited at recognized repositories are still very few, and the process of tracking new variants is not coordinated. CoV2K is a manually curated knowledge base providing an organized collection of information about SARS-CoV-2 variants, extracted from the scientific literature; it features a taxonomy of variant impacts, organized according to three main categories (protein stability, epidemiology, and immunology) and including levels for these effects (higher, lower, null) resulting from a coherent interpretation of research articles.
CoV2K is integrated with ViruSurf, hosted at Politecnico di Milano; ViruSurf is globally the largest database of curated viral sequences and variants, integrated from deposition repositories such as COG-UK, GenBank, and GISAID. Thanks to such integration, variants documented in CoV2K can be analyzed and searched over large volumes of nucleotide and amino acid sequences, e.g., for co-occurrence and impact agreement; the paper sketches some of the data analysis tests that are currently under development.

Keywords: SARS-CoV-2 · Variant impact · COVID-19 · Knowledge base · Data integration · Statistical testing

1 Introduction

The global COVID-19 pandemic, caused by the SARS-CoV-2 viral infection, has impacted everyone's lives, with more than 100 million confirmed cases, including more than 2.2 million deaths worldwide, as reported by the World Health Organization on January 31st, 2021 (`https://covid19.who.int/`). Achieving genetic diversity is an essential aspect for the continuation of SARS-CoV-2 (similarly to other RNA viruses), because it brings viral survival, fitness, and pathogenesis [11]. Therefore, collecting information about genetic variation in SARS-CoV-2 becomes overly necessary, e.g., for studying its relationship with effects on the COVID-19 pandemic.

To accomplish this goal, we started to build CoV2K, a knowledge base fuelled by information extracted from published papers and preprints. Unlike in

other domains and contexts, automatic text mining methods are not applicable for building CoV2K, because there is not enough good quality material to instruct effective mining methods; therefore, we are building the knowledge base manually, following a systematic procedure.

In organizing the knowledge base, we designed categories that well-represent the impact of virus' variants on viral characteristics, arranged according to three main categories (protein stability, epidemiology, immunology) and including levels for these effects (higher, lower, null). The knowledge base is connected to a massive amount of publicly available data from heterogeneous sources (RefSeq, COG-UK, GenBank, NMDC, and GISAID); this connection is supported thanks to the ViruSurf database [5], a recently developed, integrated and curated resource, hosted by Politecnico di Milano. Furthermore, the knowledge base is connected to the VirusViz service (`http://genomic.deib.polimi.it/virusviz/`), which allows user-provided data to be analyzed and visualized.

The availability of CoV2K and its connection with ViruSurf and VirusViz opens opportunities for new discoveries, which can be achieved through statistical testing, e.g., about multiple variants co-occurrence and its spreading at given times within populations of particular geographical regions. Thus, it becomes possible to connect knowledge about variants published at a given time to the past and future diffusion of variants within publicly available sequences.

This paper is intended as a progress report and is organized as follows: Section 2 overviews how information is acquired and organized within the knowledge base; Section 3 describes its connection with the ViruSurf database; Section 4 elicits the expected data analysis activities exploiting statistical testing within parametric sub-populations; Section 5 finally concludes.

2 Knowledge base construction

2.1 Data acquisition

The data acquisition protocol is designed for building a comprehensive and well-organized knowledge base, considering both peer-reviewed articles and preprints, with peer-reviewed articles as the most valuable sources. We selected Google Scholar, PubMed, GISAID/COG-UK reports, MedRxiv, and BioRxiv as our data sources; we then selected search keywords, used individually or in pairs/triplets, of:

- virus terms (SARS-CoV-2, Coronavirus, 2019-nCoV, COVID-19, Spike, lineage, RBD, ...);
- known clinical impacts (transmission, fatality rate, monoclonal antibodies, phenotype, severe outcome, ...);
- known variants and/or lineages (variant of interest/VOC, D614G, N501Y, B.1.1.7, UK lineage, Brazilian variant, ...).

Then, we filtered the outcome according to the relevance to the research's topic. We daily perform data searches, noting that the COG-UK report is updated on a bi-weekly basis[1].

CoV2K structure is composed of three sections. The first section represents the variant characterization, including the protein encoded by the gene (called product), the type of variant (i.e., substitution, insertion or deletion), the amino acid variation (composed of a reference sequence, its position on the reference genome, and an alternative sequence), as well as the identifier of the reference genome used in the study. The second section, which is better defined next, describes the variant's impact (i.e., how the virus behaviour is influenced by the presence of that variant). The third section links the variant to the source of information, defined by the manuscript's author, the DOI, and type of publication. We next define the second section in details.

2.2 Taxonomy of effects

According to epidemiological studies and definitions, we organize variant effects into three categories, as follows:

Protein Stability. In this category, we organize all the variants that could lead to a change in the produced protein's stability (see R203K and G204R in Table 1, as examples). These are reported in structure-related studies focusing on the "stability" of viral proteins. Several genetic variations are non-synonymous, thus altering the amino acid composition of viral proteins, which will produce a protein with different degrees of stability [4].

Epidemiology. This category is important to understand SARS-CoV-2 evolutionary epidemiology, viral kinetics and dynamics related studies. It includes:

- *Viral transmission*, the virus capability to pass from a host to another host [9]. See P323L, D614G, and N501Y in Table 1.
- *Infectivity*, the capability of a transmitted virus to actually establish infection [7]. See V367F and D614G in Table 1.
- *Disease severity*, an assessment of systematic symptoms caused by the virus [14]. See D614 in Table 1.
- *Fatality rate*, the proportion of persons who die after the viral infection over the number of confirmed infected people [7]. See Q57H and P323L in Table 1.

Immunology. This category is concerned with immune response and virus-host interactions related studies, including any immune system process that happened in response to a virus-host interaction [9]. This category is important in the vaccine and therapeutic development studies, and it includes three subcategories:

- *Sensitivity to convalescent sera*: as in other infections, the convalescent serum from recovered individuals might be used for prevention and treatment of

[1] Last accessed report (January 31st, 2021) before submission deadline: `https://www.cogconsortium.uk/wp-content/uploads/2021/01/Report-2_COG-UK_SARS-CoV-2-Mutations.pdf`.

Table 1. Example CoV2K variants with the related effect and level, captured on the specified literature publication of different type (publ. = published; prep = preprint).

Variant signature					Impact		Publication		
Product	Type	Orig.	Position	Alt.	Effect	Level	Author	DOI	Type
N	SUB	R	203	K	protein stability	L	Parvez et al.	https://doi.org/10.1016/j.compbiolchem.2020.107413	publ
N	SUB	G	204	R	protein stability	L	Parvez et al.	https://doi.org/10.1016/j.compbiolchem.2020.107413	publ
NS3	SUB	Q	57	H	fatality rate	L	Oulas et al.	https://doi.org/10.1101/2020.08.17.253484	publ
NSP12	SUB	P	323	L	viral transmission	H	Wang et al.	https://doi.org/10.21203/rs.3.rs-49671/v1	prep
NSP12	SUB	P	323	L	fatality rate	H	Toyoshima et al.	https://doi.org/10.1038/s10038-020-0808-9	publ
Spike	SUB	V	367	F	infectivity	H	Junxian et al.	https://doi.org/10.1101/2020.03.15.991844	prep
Spike	SUB	D	614	G	infectivity	H	Korber et al.	https://doi.org/10.1016/j.cell.2020.06.043	publ
Spike	SUB	D	614	G	viral transmission	H	Zhang et al.	https://doi.org/10.1038/s41467-020-19808-4	publ
Spike	SUB	D	614	G	viral transmission	H	Volz et al.	https://doi.org/10.1016/j.cell.2020.11.020	publ
Spike	SUB	D	614	G	disease severity	N	Volz et al.	https://doi.org/10.1016/j.cell.2020.11.020	publ
Spike	SUB	N	501	Y	viral transmission	H	Teruel et al.	https://doi.org/10.1101/2020.12.16.423118	prep
Spike	SUB	N	501	Y	binding affinity host rec.	H	Santos et al.	https://doi.org/10.1101/2020.12.29.424708	prep
Spike	SUB	N	439	K	sensitivity to conv. sera	L	Qianqian et al.	https://doi.org/10.1016/j.cell.2020.07.012	publ
Spike	SUB	N	439	K	sensitivity to mAbs	L	Qianqian et al.	https://doi.org/10.1016/j.cell.2020.07.012	publ

COVID-19 (thus providing passive immunization [1]), as it is assumed that convalescent plasma donors may have developed an effective immune response to the offending pathogen. See N439K in Table 1.

- *Sensitivity to neutralizing mAbs*, measuring the sensitivity of the variants towards monoclonal antibodies – the mechanism in which a subset of antibodies blocks the viral infection is called neutralization. This kind of sensitivity has a crucial role in vaccine development. See N439K in Table 1.
- *Binding affinity to host receptor*. SARS-CoV-2 is entering the host cells by binding its receptor-binding domain (RBD), in the spike protein, to a cell receptor called angiotensin-converting enzyme 2 (ACE2). Modifying the binding affinity could lead to a change in the efficacy of cell entering. Hence, binding affinity potentially affects cell infectivity and immune evasion [13]. See N501Y in Table 1.

Subcategories may be associated to higher, lower and null levels:

- Higher (H): the variant's presence leads to an increase of a specific effect.
- Lower (L): the variant's presence leads to a decrease of a specific effect.
- Null (N): the variant's presence does not change a specific effect (after testing).

All categories and sub-categories are flexible and will be extended according to the newly studied variants and their timely reported impact. Table 1 represents an excerpt of the current state of the knowledge base, with a few examples of variants' impacts.

3 Integration with sequence data

ViruSurf is a large integrated database of viral sequences of SARS-CoV-2 (and similar viruses), hosted at `http://gmql.eu/virusurf/`; it stores all the sequences that have been deposited to GenBank and COG-UK, whereas a similar database, hosted at `http://gmql.eu/virusurf_gisaid/`, stores only a subset of the data and metadata from the GISAID repository, which however includes amino acid variants, the most important information from the knowledge

base perspective. An incremental pipeline can be frequently initiated (e.g. on a weekly or bi-weekly basis) in order to add new deposited sequences to the ViruSurf databases; the pipeline applies a variant calling algorithm extracting mutations on both the nucleotide and the amino acid levels, and includes a search for overlapping sequences, since many sequences deposited in GenBank and COG-UK overlap with those deposited in GISAID. As of January 31st, 2021 the databases contain about 500K non-overlapping sequences, with a significant monthly growth rate (15-25%).

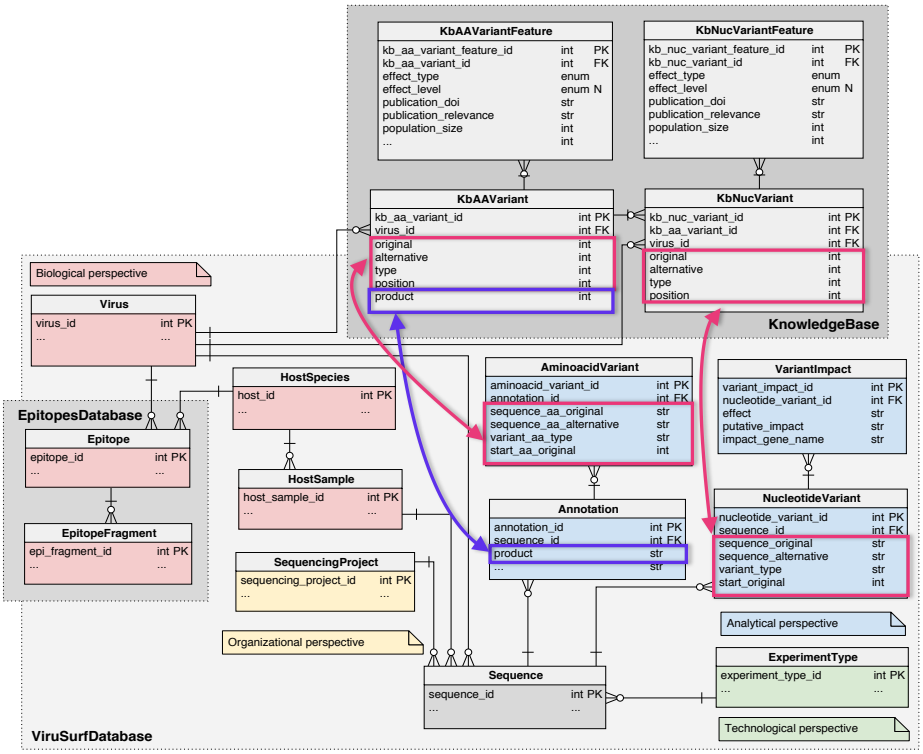

Fig. 1. Schema of the knowledge base connected to the ViruSurf database

Fig. 1 represents the logical schema of ViruSurf databases. While the complete ViruSurf schema can be appreciated in [5] based on the Viral Conceptual Model [3], we here report the essential aspects: it is centered on the Sequence table, connected to SequencingProject, Virus, HostSample/HostSpecies, and ExperimentType. The "analytical" perspective of the schema contains information about Annotations (characterization of sub-parts of the sequence), NucleotideVariants and AminoacidVariants.

The new knowledge base concepts are centered upon two entities, the KbAAVariant and the KbNucVariant, respectively including all the amino acid and nucleotide-level variants captured from our data acquisition process (Section 2). Two one-to-many relationships connect each variant instance to its possibly many effects, respectively described in KbAAVariantFeature and KbNucVariantFeature.

The connection of the knowledge base to the database requires building bridges from the variant characterization attributes of the knowledge base to the variants in the database. The matching with the ViruSurf tables is depicted in Fig. 1: the products need to match with the annotation table, as shown by the purple arrow/boxes; the variant signatures (original, alternative, type, and position) need to match with the corresponding four fields highlighted by pink arrows/boxes.

4 Data analysis

Basic summary statistics about the CoV2K variants' distribution and their impact can be used to describe CoV2K data, answering questions such as: (i) What is the CoV2K variants' density in a specific gene? (ii) Are there conflicts in specific variant's impact's level (e.g., a variant reporting both H and L levels)? Other simple statistics may relate CoV2K variants to their time and space distribution in the ViruSurf databases, such as: (iii) is there a significant relationship between CoV2K variants and specific geographical areas? (iv) How fast do they spread within such areas?

These observations are important to assess the status of the COVID-19 pandemic (e.g., tracing the Brazilian or South African variants) and can be done with limited delay compared to sequence deposition time in publicly available databases from the various countries. For evaluating the significance of such statistics we use standard tests chosen depending on the type of data, e.g. Fisher's exact test or Chi-squared test on contingency tables by means of aggregate queries on ViruSurf databases. Other statistical analysis are currently under design. Among them, due to space limitations, we only report our current approach to study the knowledge base variants' co-occurrence.

Example on co-occurring mutations. We focus on amino acid variants from sequences extracted from ViruSurf (on January 23rd, 2020). For each pair of variants x and y we computed a 2×2 contingency table, accounting for the number of sequences containing i) both x and y, ii) only x, iii) only y, and iv) neither of the two. Then, for each contingency table we applied the Cramer's V test [8], i.e., a modified version of the well-known Pearson's Chi-squared test [10] which is preferable in case of large sample sizes[2]. We then built an $N\times N$ matrix, where N is the number of distinct pairs of CoV2K variants (possibly paired to

[2] The strength of association ranges from 0 (no association) to 1 (perfect association); the value 0.1 is considered a good significance threshold for the relationship between two variables, see `http://www.acastat.com/statbook/chisqassoc.htm`.

their impact) and the elements of the matrix are the results of the Cramer's V test. The goal is to identify pairs of mutations that co-occur in a statistically significant way in the observed population, which also agree/disagree in their impact, as reported in literature and stored in CoV2K; here we considered all publicly available sequences, but we plan to allow the choice of specific sub-populations.

	NS3_Q57H:fr:l	NSP12_P323L:fr:h	NSP12_P323L:vt:h	N_G204R:ps:l	N_R203K:ps:l	Spike_D614G:vt:h	Spike_N501Y:in:h	Spike_P681H	Spike_S982A
NS3_Q57H:fr:l	x								
NSP12_P323L:fr:h	0.14	x							
NSP12_P323L:vt:h	0.14	x	x						
N_G204R:ps:l	0.38	0.17	0.17	x					
N_R203K:ps:l	0.38	0.17	0.17	0.99	x				
Spike_D614G:vt:h	0.14	0.83	0.83	0.18	0.18	x			
Spike_N501Y:in:h	0.12	0.068	0.068	0.33	0.33	0.064	x		
Spike_P681H	0.13	0.068	0.068	0.32	0.32	0.066	0.92	x	
Spike_S982A	0.13	0.06	0.06	0.34	0.34	0.069	0.95	0.94	x

Fig. 2. Result matrix of co-occurrence analysis. Each row/column label represents a variant:effect(first letters):level(L/H).

Fig. 2 reports an excerpt of our results. Empty cells are not computed as they are symmetrical to the lower triangular matrix; the symbol $\times$ indicates positions for the same variants. An explanatory color scheme was used to explain Cramer's V test results:

- `Black` captures pairs that are not significant (lower than 0.1), e.g., the black rectangle indicates that D614G in the Spike does not significantly co-occur with the last three variants of the table.
- `Blue` captures pairs that are significant (higher than 0.1), e.g., the three values marked by a blue triangle in the figure.
- `Green` captures pairs that, in addition to being significant, agree on the same effect and its level (e.g., the Spike protein D614G with the NSP12 protein P323L, and the two N protein substitutions at 203 and 204 positions).
- `Red` captures pairs that, in addition to being significant, agree on the effect but report an opposite level. This is the case of NS3 protein Q57H variant – which is reported as *decreasing* the fatality rate – significantly co-occurring with the NSP12 protein P323L variant – which is instead reported as *increasing* the fatality rate.

Note that the last three rows of the table represent some of the variants that define the B.1.1.7 lineage[3], which correspond to the "UK strain" that raised worldwide attention since December 2020. The Cramer's V test results of these three variants against all the other ones are very similar (see the last three rows

[3] https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563

of the matrix); the co-occurrence of the three variants is also well documented by the UK strain definition [12].

As the matrix entries proved effective in confirming properties, we may use them also for prediction: by looking at entries within the blue triangle and considering that N501Y is reported as increasing the infectivity of the virus, we could predict the same effect also for also P681H and S982A. Such mechanism prompts further investigations of other variant pairs that could be similarly constructed. Referring instead to variants that are not yet described in CoV2K, co-occurrence with variants with a known effect may prompt targeted lab experiments.

5 Conclusion

We reported our initial design of a knowledge base of SARS-CoV-2 variants, whose strength lies both in a well-structured procedure for acquiring and organizing data and in the integration with the ViruSurf databases; the interplay between a large amount of up-to-date sequence information and manually curated consolidated knowledge is very promising, as confirmed by our preliminary data analysis results. In the future, we will refine the breadth and complexity of statistical tests, and fine tune them by means of bias correction methods (e.g., [2]) and by choosing thresholds appropriate for large samples (e.g., [6]).

The effectiveness of CoV2K will be evaluated with the help of domain experts that will also inspire more complex analyses to increase its benefits. CoV2K is by now only a taxonomy, but we will consider building a richer semantic representation of its elements, thereby helping automate reasoning and statistical tests.

Acknowledgment

This research is funded by the ERC Advanced Grant 693174 GeCo (data-driven Genomic Computing).

References

1. Abraham, J.: Passive antibody therapy in COVID-19. Nature Reviews Immunology **20**(7), 401–403 (2020)
2. Bergsma, W.: A bias-correction for Cramér's V and Tschuprow's T. Journal of the Korean Statistical Society **42**(3), 323–328 (2013)
3. Bernasconi, A., Canakoglu, A., Pinoli, P., Ceri, S.: Empowering Virus Sequence Research Through Conceptual Modeling. In: Dobbie, G., Frank, U., Kappel, G., Liddle, S.W., Mayr, H.C. (eds.) Conceptual Modeling. pp. 388–402. Springer International Publishing, Cham (2020)
4. Brinda, K., Vishveshwara, S.: A network representation of protein structures: implications for protein stability. Biophysical journal **89**(6), 4159–4170 (2005)
5. Canakoglu, A., Pinoli, P., Bernasconi, A., Alfonsi, T., Melidis, D.P., Ceri, S.: ViruSurf: an integrated database to investigate viral sequences. Nucleic acids research **49**(D1), D817–D824 (2021)

6. Cao, H., Hripcsak, G., Markatou, M.: A statistical methodology for analyzing co-occurrence data from a large sample. Journal of biomedical informatics **40**(3), 343–352 (2007)
7. Centers for Disease Control and Prevention: Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics. Atlanta, GA: US Dept. of Health and Human Services, Centers for Disease (2006), `https://www.cdc.gov/csels/dsepd/ss1978/ss1978.pdf`; accessed on Jan. 31st, 2021.
8. Cramér, H.: Mathematical methods of statistics, vol. 43. Princeton University Press (1946)
9. He, Y., Yu, H., Ong, E., Wang, Y., Liu, Y., Huffman, A., Huang, H.h., Beverley, J., Hur, J., Yang, X., et al.: CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific Data **7**(1), 1–5 (2020)
10. Pearson, K.: X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **50**(302), 157–175 (1900)
11. Rahimi, A., Mirzazadeh, A., Tavakolpour, S.: Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. Genomics (2020)
12. Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G.: A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology **5**(11), 1403–1407 (2020)
13. Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., Li, F.: Cell entry mechanisms of SARS-CoV-2. Proceedings of the National Academy of Sciences **117**(21), 11727–11734 (2020)
14. Wu, J.T., Leung, K., Bushman, M., Kishore, N., Niehus, R., de Salazar, P.M., Cowling, B.J., Lipsitch, M., Leung, G.M.: Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nature medicine **26**(4), 506–510 (2020)