A unique approach to SARS-CoV-2 data and knowledge ingestion, integration and querying

Anna Bernasconi[1], Ruba Al Khalaf[1], Tommaso Alfonsi[1], Arif Canakoglu[1], Luca Cilibrasi[1], Andrea Gulino[1], Pietro Pinoli[1], and Stefano Ceri[1]

Politecnico di Milano, Milan, Italy, firstname.lastname@polimi.it

Abstract. The outbreak of the COVID-19 epidemic has concentrated unique attention on the genetic mechanisms underlying viral infection and the related disease. We focused on the collection, integration and analysis of SARS-CoV-2 data, which informs on the structure of the virus, its ability to spread, mutate, and evolve its behavior with related effects. At the beginning of the pandemic, we promptly addressed the domain by conducting an extensive requirements investigation, leading to the precise modeling of data types and knowledge types. Then, we employed a big data-driven approach for building ingestion, integration, and processing pipelines. Big optimized databases are now made available by means of several specialized endpoints, abiding by the FAIR principles. We present our position in the viral genomics data management as a solid support to the fight against SARS-CoV-2 spread and similar perspective phenomena.

Keywords: COVID-19 · viral genomics · SARS-CoV-2 · knowledge management · data integration

Introduction

The outbreak of COVID-19 has presented novel challenges to the research community, pushed by the intent of rapidly mitigating the pandemic effects. During these times, we observed the production of an exorbitant amount of data; the total number of sequences of SARS-CoV-2 available worldwide went from few hundreds in March 2020, up to about one hundred thousand in August 2020, and more than 3.5 millions in September 2021.

Inspired by our work on genomic data integration [5,8], we searched for effective ways to help investigate the new phenomenon with our contribution. For understanding areas of interest (including types of data and user needs), we started a broad requirements analysis activity, involving interviews to several domain experts. We then produced concise models to understand and organize data as the basis for building search, visualization and analysis systems. In this paper, we state our current position in the COVID-19 data modeling and tool development community and motivate our present and future commitment to FAIR systems in this domain.

Collecting requirements

As a first activity, we identified potential areas of interest for what concerns available data types (e.g., sequences, mutations, and epitopes) and user needs (e.g., immunology, surveillance, veterinary virology). We then interviewed experts in various domains about specific issues (see Fig. 1 for exemplary paths of investigation). Several research areas were covered, including host genetics (e.g., which genes can explain a specific COVID-19 disease onset), immunology (i.e., the binding mechanism of the virus to the host immune system), phylogenetics (e.g., how the virus *evolves* in time/space), epidemiology (e.g., how the virus *spread* in time/space).

The requirements collection process, targeted around 30 experts from both university and private companies; it involved professionals from Italy, Europe, and in the World (Singapore, Hong Kong, USA), as described in [2]. The long conversations, prototypes, and focus groups, led to the design and development of several models and systems that are outlined in the bottom row of Fig. 1 and summarized in the following.

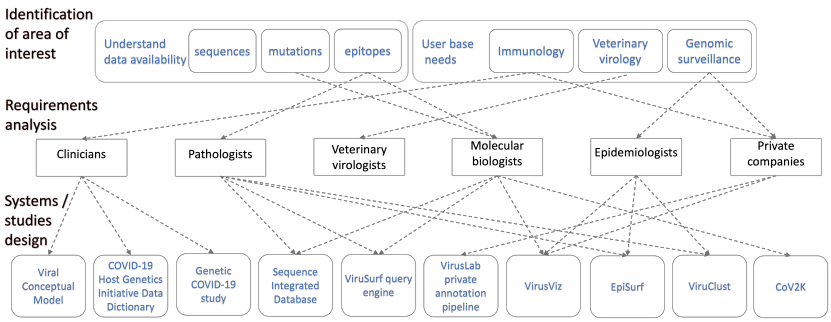

Fig. 1. Overview of areas of interest explored during requirements analysis sessions for the design of our tool-suite modules.

The most unexpected aspect of our interview process was the availability and enthusiasm of the experts that we interviewed. Experts from all fields were excited in sharing their knowledge with us and actually available above our expectations.

Modeling SARS-CoV-2 data and knowledge

Understanding viruses from a conceptual modeling perspective is very important. In April 2020 we designed the Viral Conceptual Model (VCM, [4]): the sequence of the virus is the central entity, described by three "views" regarding i) the information on the virus and on the infected host; ii) details on the technology and process used for extracting the sequence; iii) metadata on the project and laboratory managing the sampling, sequencing and analysis pipelines. Additionally, we model the sequence's annotated parts (known genes, coding and

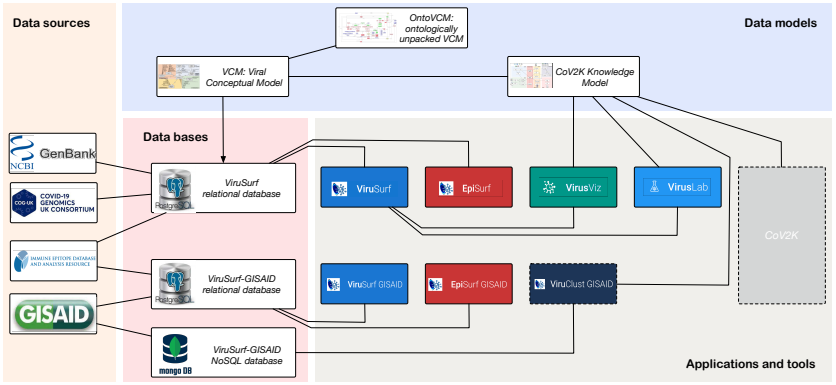

Fig. 2. Comprehensive data sources, models, bases, and tools overview.

untranslated regions...) and their nucleotide/amino acids mutations, computed with respect to the reference sequence of the the viral species. An enriched model – resulting from the ontological unpacking of the initial VCM – has also been proposed [12].

We then described a Knowledge Model that allows to represent both data (thus embedding the VCM) and the external knowledge that is being collected about SARS-CoV-2. This includes notions on variants; their effects (in terms of disease severity, transmissibility, vaccine escape, etc.); their composition – in terms of sets of mutations; the peculiarities of mutations due to their original and alternative nucleotide or amino acid residue; and the definition of particular regions of the genome with given functions. The proposal is preliminarily sketched in CoV2K [1], but we are working towards a much broader definition, allowing to perform targeted search that interconnects data and knowledge, with potential related statistical tests.

The information about the virus can be effectively connected to the data about the host genotype and phenotype, as we proposed in [3], where we described a conceptual model for the phenotype of COVID-19 patients. The model was exploited within the COVID-19 Host Genetics Initiative [10], an open community that gathered thousands of researchers to produce, share, and analyze data to learn the genetic determinants of COVID-19 susceptibility, severity, and outcomes; among several studies, we collaborated to [11].

Building data-driven systems

After our modeling effort, we built solid pipelines for extracting data from the original deposition portals and integrating them within our global models. In doing so, the most difficult aspect we had to consider is the growth of data (which reached 3.5 million genomes as of September 2021, in only 1.5 years). Such data continuously needs to be mastered by increasingly powerful computing resources with several logical and physical optimizations.

ViruSurf [9] is our first system, designed for collecting sequences from the two biggest – completely open – SARS-CoV-2 data sources, e.g., GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) and COG-UK (`https://www.cogconsortium.uk/`). We also implemented the dual system ViruSurf-GISAID, storing sequences from GISAID (`https://www.gisaid.org/`), currently hosting the major deposition database; accessing GISAID data is subject to accepting a Data Agreement that is typically granted to users from research and academy. ViruSurf offers a practical interface where each drop-down menu is a metadata attribute describing viral sequences. Possible values are paired with the number of available sequences in the database. Different conditions can be built on the presence of specific mutational patterns (predicating on either nucleotides or amino acid residues). EpiSurf and EpiSurf-GISAID [6] are companion systems for analysing sequences mutations in the context of specific viral genomic regions, i.e., epitopes. Epitopes, extracted from the The Immune Epitope Database (`https://www.iedb.org/`), are strings of amino acid residues from the virus protein that can be recognized by antibodies or other host's receptors. In the mentioned interfaces, results are produced as browsable tables of sequences and epitopes, described by their metadata. They can be downloaded as textual files that are easily embedded in bioinformatic pipelines. A more visual and interactive support is provided by VirusViz [7], which can be opened directly from sequences sets defined within ViruSurf or EpiSurf, as well as from a user-input file of sequences. VirusViz allows to partition the population of interest into groups and visualize comparatively their mutation distributions, with several options for highlighting positions, mutational patterns, and regions of interests. VirusViz has a dual close-source solution, i.e., VirusLab [14], commercialized by the Quantia Consulting S.R.L. company and developed in collaboration with our group at Politecnico di Milano, within the EIT project N. 20663. Lately, we have been working towards the expansion of our systems' corpus by proposing ViruClust, a tool for comparing SARS-CoV-2 genomic sequences and lineages in space and time without any computational background, and CoV2K-API, a flexible API for exploring the interplay between SARS-CoV-2 data and knowledge—an ever growing source of information for variants, their effects and genetic characteristics.

All mentioned systems are modules of the broad architecture illustrated in Fig. 2, showing different areas: data sources (orange background); data bases (pink); data models (blue); and systems (gray), such as web applications and tools. The tool suite will be progressively extended with additional support to new data types and functionalities.

Discussion

We are aware of multiple initiatives that aim at systematizing researchers work on data production and curation on COVID-19 (e.g., the VODAN network, `https://www.go-fair.org/overview/vodan` and the CIDO ontology [13]). Here, we present our very proactive approach, embracing a broad vision: availability of conceptual models, related databases and search systems for SARS-CoV-2. While we already count three published models and three published tools, we

are already facing new challenges: developing new tools for data analysis that can be used for detailed tracing of mutation prevalence in time and space, with the objective of helping the study and control of the viral spreading.

Our proposal takes full advantage of the FAIRness of SARS-CoV-2 data extracted from external data sources and further commits to the improvement of Accessibility and Findability by offering foundational metadata models for enabling flexible and fast search of data. Interoperability is achieved thanks to the global VCM schema allowing integration of heterogeneous schemata and vocabularies. Reusability is corroborated by the ease of continuously adding modules that complete our whole proposal. FAIR data-driven approaches to genomics and virology will provide important opportunities for research, especially counting on an ever growing corpus of data depositions from labs around the world.

Acknowledgements. Project supported by the ERC AdG 693174 GeCo.

References

1. Al Khalaf, R., et al.: CoV2K: A Knowledge Base of SARS-CoV-2 Variant Impacts. In: Research Challenges in Information Science. pp. 274–282. Springer (2021)
2. Bernasconi, A.: Extreme Requirements Elicitation: Lessons Learnt from the COVID-19 Case Study. In: Joint Proceedings of REFSQ 2021. CEUR Workshop Proceedings, vol. 2857 (2021)
3. Bernasconi, A., et al.: A review on viral data sources and search systems for perspective mitigation of COVID-19. Brief. Bioinform. **22**(2), 664–675 (2021)
4. Bernasconi, A., et al.: Empowering virus sequence research through conceptual modeling. In: Int. Conf. on Conceptual Modeling. pp. 388–402. Springer (2020)
5. Bernasconi, A., et al.: Conceptual modeling for genomics: building an integrated repository of open data. In: Int. Conf. on Conceptual Modeling. pp. 325–339. Springer (2017)
6. Bernasconi, A., et al.: EpiSurf: metadata-driven search server for analyzing amino acid changes within epitopes of SARS-CoV-2 and other viral species. Database **2021** (2021)
7. Bernasconi, A., et al.: VirusViz: Comparative analysis and effective visualization of viral nucleotide and amino acid variants. Nucleic Acids Res. **49**(15), e90 (2021)
8. Canakoglu, A., et al.: GenoSurf: metadata driven semantic search system for integrated genomic datasets. Database **2019** (2019)
9. Canakoglu, A., et al.: ViruSurf: an integrated database to investigate viral sequences. Nucleic Acids Res. **49**(D1), D817–D824 (2021)
10. COVID-19 Host Genetics Initiative: Mapping the human genetic architecture of COVID-19. Nature **600**, 472—477 (2021)
11. Daga, S., et al.: Employing a systematic approach to biobanking and analyzing clinical and genetic data for advancing COVID-19 research. European Journal of Human Genetics **29**(5), 745–759 (2021)
12. Guizzardi, G., et al.: Ontological Unpacking as Explanation: The Case of the Viral Conceptual Model. In: Int. Conf. on Conceptual Modeling. Springer (2021), 356–366
13. He, Y., et al.: CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific data **7**(1), 1–5 (2020)
14. Pinoli, P., et al.: VirusLab: A Tool for Customized SARS-CoV-2 Data Analysis. BioTech **10**(4), 27 (2021)