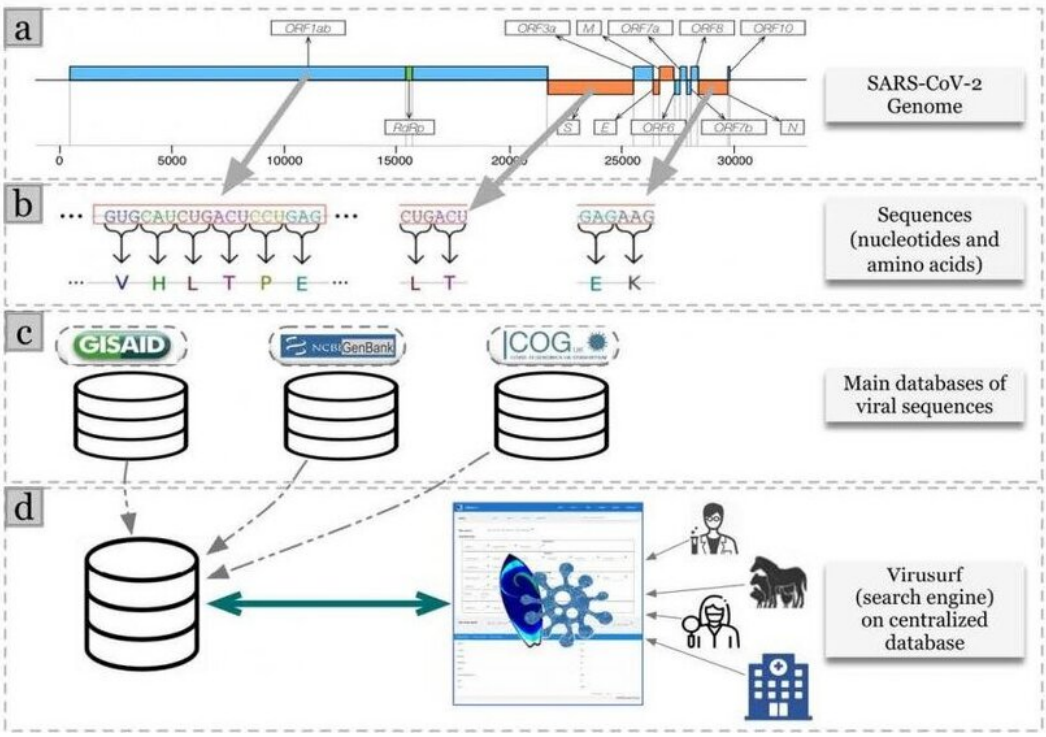
Discovering the secrets of viral sequences in COVID-19

November 23 2020



From the SARS-CoV-2 genome (a) sequences of nucleotides and amino acids are extracted (b); sequences are then deposited to world-wide open repositories: GENBANK, GISAID, COG-UK (c), and imported to the centralized database at Politecnico, where the search engine ViruSurf is accessed (d). Credit: Politecnico di Milano

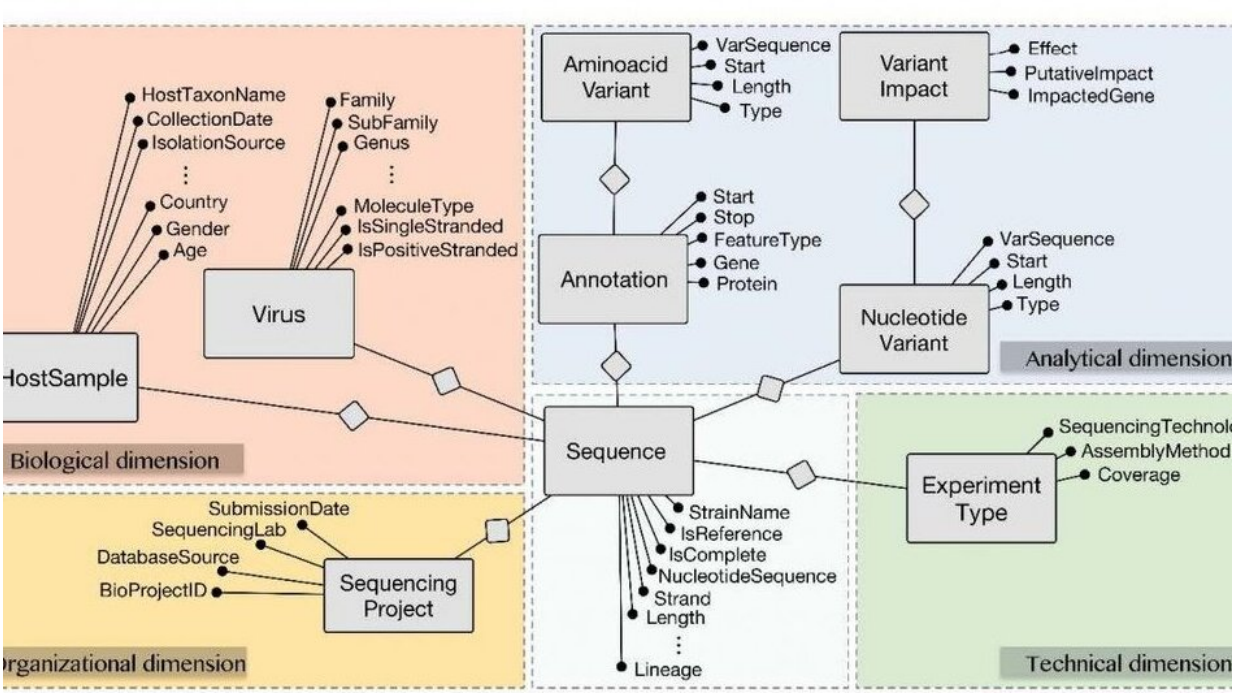
Since the beginning of 2020, labs from all around the world have been

sequencing the material from positive tests of people affected by COVID-19 and then depositing sequences mostly to three points of collection: GenBank, COG-UK, and GISAID. Rapid exploration of this huge amount of data is important for understanding how the genome of the virus is changing. For enabling fast 'surfing' over this data, the research group at Politecnico di Milano led by Prof. Stefano Ceri has developed ViruSurf, a search engine operating on top of a centralized database stored at Politecnico. The database is periodically reloaded from the three sources and as of today contains 200,516 sequences of SARS-CoV-2, the virus causing COVID-19, and 33,256 sequences of other viral species also associated with epidemics affecting humans, such as SARS, MERS, Ebola, and Dengue.

Every sequence is described from four perspectives: the biological features of the virus and the host, the sequencing technology, the project that has produced the original data, the mutations of the whole sequence of nucleotides and of gene-specific amino acids. The advantage provided by ViruSurf is the use of an algorithm for computing viral mutations homogeneously across sources, using cloud computing. The database is optimized for giving quick responses to the search engine surfers.

Among the future developments of ViruSurf, the most important, funded by a six-month-long project by EIT Digital, is a bio-informatic service for ingesting new viral sequences, which highlights the presence of viral mutations associated with enhanced or reduced severity and virulence as they are discovered. Used in clinics, particularly with a less acute pandemics spreading, it will support the addition of critical information to the patient health record; other uses will be possible in the context of animal farming or of the food chain. The system will soon allow the tracing of epitopes—amino acid sequences that are used in vaccine design—for instance to associate epitopes with mutations of the virus that could be present in given countries of the world and that could affect vaccine.

"In the GeCo project, financed by the European Research Council, we had already developed a search engine for datasets describing the human genome, called GenoSurf; at the beginning of the pandemic, there was no such system for viral sequences. To better understand its requirements, we interviewed about twenty expert virologists from all over the world. The result is a user-friendly system: any researcher can connect to it and perform queries, for instance, about when a viral mutation started and how it has spread in the world," says Stefano Ceri, the project leader. The article is published on a high relevance journal, *Nucleic Acids Research*, in the database issue that every year collects the descriptions of the most significant biological databases. The article is authored also by Pietro Pinoli, algorithm designer, Arif Canakoglu, software architect, Anna Bernasconi, data designer, Tommaso Alfonsi, designer of the data loading pipeline, and Damianos P. Melidis from L3S (Hannover), author of some algorithms.



Schema of the integrated database. Viral sequences are described by their biological dimension (the host and the virus), by the sequencing project which has published them, by the sequencing technology, and by the genomic properties (annotations and mutations, both of nucleotide and amino acids). Credit: Politecnico di Milano

More information: Arif Canakoglu et al, ViruSurf: an integrated database to investigate viral sequences, *Nucleic Acids Research* (2020). DOI: 10.1093/nar/gkaa846

Provided by Politecnico di Milano