Empowering Virus Sequence Research through Conceptual Modeling

Anna Bernasconi(✉)[0000−0001−8016−5750], Arif Canakoglu[0000−0003−4528−6586], Pietro Pinoli[0000−0001−9786−2851], and Stefano Ceri[0000−0003−0671−2415]

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio 34/5, 20133, Milano, Italy
{first.last}@polimi.it

Abstract. The pandemic outbreak of the coronavirus disease has attracted attention towards the genetic mechanisms of viruses. We hereby present the Viral Conceptual Model (VCM), centered on the virus sequence and described from four perspectives: biological (virus type and hosts/sample), analytical (annotations, nucleotide and amino acid variants), organizational (sequencing project) and technical (experimental technology).
VCM is inspired by GCM, our previously developed Genomic Conceptual Model, but it introduces many novel concepts, as viral sequences significantly differ from human genomes. When applied to SARS-CoV-2 virus, complex conceptual queries upon VCM are able to replicate the search results of recent articles, hence demonstrating huge potential in supporting virology research.
Our effort is part of a broad vision: availability of conceptual models for both human genomics and viruses will provide important opportunities for research, especially if interconnected by the same human being, playing the role of virus host as well as provider of genomic and phenotype information.

Keywords: Conceptual Model · Open Data · SARS-CoV-2 · Viral Genomics · Biological Research

1 Introduction

Despite the advances in drug and vaccine research, diseases caused by viral infection pose serious threats to public health, both as emerging epidemics (e.g., Zika virus, Middle East Respiratory Syndrome Coronavirus, Measles virus, or Ebola virus) and as globally well-established epidemics (such as Human Immunodeficiency Virus, Dengue virus, Hepatitis C virus). The pandemic outbreak of the coronavirus disease COVID-19, caused by the "Severe acute respiratory syndrome coronavirus 2" virus species SARS-CoV-2 (according to the GenBank [41] acronym[1]), has brought unprecedented attention towards the genetic mechanisms of coronaviruses.

[1] SARS-CoV-2 is generally identified by the NCBI taxonomy [17] ID 2697049.

Thus, understanding viruses from a conceptual modeling perspective is very important. The sequence of the virus is the central information, along with its annotated parts (known genes, coding and untranslated regions...) and the nucleotide/amino acids variants, computed with respect to the reference sequence chosen for the species. Each sequence is characterized by a *strain name*, which belongs to a specific virus species. Viruses have complex taxonomies (as discussed in [26]): a species belongs to a genus, to a sub-family, and finally to a family (e.g., Coronaviridae). Other important aspects include the host organisms and isolation sources from which viral materials are extracted, the sequencing project, the scientific and medical publications related to the discovery of sequences; virus strains may be searched and compared intra- and cross-species. Luckily, all these data are made available publicly by various resources, from which they can be downloaded and re-distributed.

Our recent work is focused on data-driven genomic computing, providing contributions in the area of modeling, integration, search and query answering. We had previously proposed a conceptual model focused on human genomics [8], which was based on a central entity ITEM, representing files of genomic regions. The simple schema evolved into a knowledge graph [5], including ontological representation of many relevant attributes (e.g., diseases, cell lines, tissue types...). The approach was validated through the implementation of the integration pipeline META-BASE [6], which feeds an integrated database, searchable through the GenoSurf interface [10]. On the basis of this experience, we are already developing the ViruSurf interface[2] for inspecting the content of a database for virus sequences, constructed by using VCM as reference conceptual schema. Based on these considerations, in this paper we contribute as follows:

- We propose a new **Viral Conceptual Model (VCM)**, a general conceptual model for describing viral sequences, organized along specific dimensions that highlight a conceptual schema similar to GCM [8];
- We provide a list of **interesting queries** replicating newly released literature on infectious diseases; these can be easily answered by using VCM as reference conceptual schema.

The manuscript is organized as follows: Section 2 overviews current technologies available for virus sequence data management. Section 3 proposes our VCM, describing the central entity SEQUENCE and the dimensions that characterize it. We show examples of applications in Section 4 and review related works in Section 5. Section 6 discloses our vision for future developments.

2 Current scenario

The landscape of relevant resources and initiatives dedicated to data collection, retrieval and analysis of virus sequences is shown in Fig. 1. We partitioned the space of contributors by considering: institutions that host data sequences, primary sequence deposition databases, tools provided for directly querying and

[2] GenoSurf: `http://gmql.eu/genosurf/`; ViruSurf: `http://gmql.eu/virusurf/`

searching them, and then organizations and tools hosting secondary data analysis interfaces that also connect to viral sequence databases. White rectangles represent resources identified using their logo. Relations between them are of three kinds: institutions *host* deposition databases, retrieval tools *query* deposition databases, secondary databases/interfaces *include data from* deposition databases, typically adding other features.

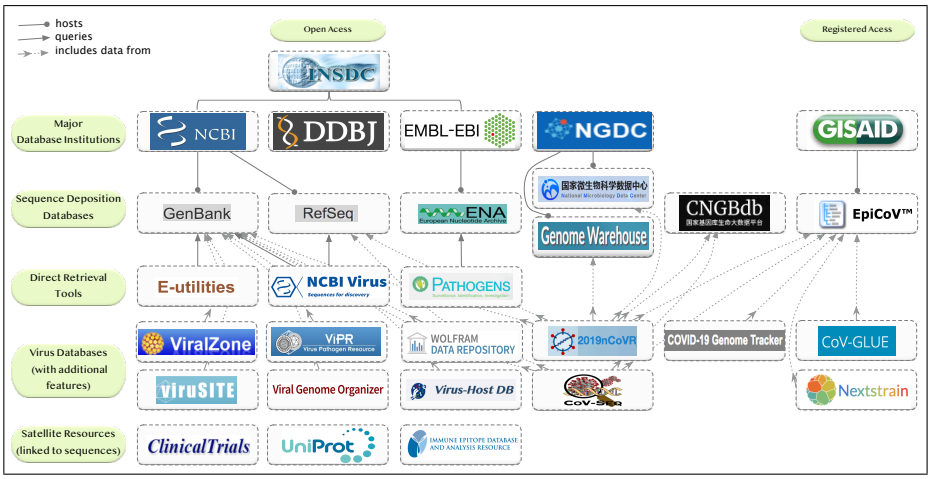

Fig. 1. Current relevant resources and initiatives dedicated to data collection, retrieval and analysis of virus sequences, divided by open and registered access.

The three main organizations providing open-source viral sequences are NCBI (US), DDBJ (Japan), and EMBL-EBI (Europe); they operate within the broader contexts provided by the International Nucleotide Sequence Database Collaboration[3]. NCBI hosts the two, so far, most relevant open viral sequence databases: RefSeq [33] provides a stable reference for genome annotation and gene identification/characterization; GenBank [41] contains an annotated collection of publicly available DNA/RNA sequences. It is continuously updated thanks to the abundant sharing of multiple laboratories and data contributors around the world (note that SARS-CoV-2 nucleotide sequences have increased from about 300 around the end of March 2020, to 13,314 as of August 1st, 2020). EMBL-EBI hosts the European Nucleotide Archive [1], which accepts submissions of nucleotide sequencing information, including raw sequencing data, sequence assembly information and functional annotations.

Several tools are available for querying and searching these databases; E-utilities [40], NCBI Virus [23], and Pathogens[4] are tools and portals directly provided by the INSDC institutions for supporting the access to their viral resources, however lacking the possibility of querying based on annotations and

[3] http://www.insdc.org/
[4] https://www.ebi.ac.uk/ena/pathogens/

variants. A number of databases and data analysis tools refer to these viral sequences databases: ViralZone [24] by the SIB Swiss Institute of Bioinformatics, which provides access to SARS-CoV-2 proteome data as well as cross-links to complementary resources; the Virus Pathogen Database and Analysis Resource (ViPR, [36]), an integrated repository of data and analysis tools for multiple virus families, supported by the Bioinformatics Resource Centers program; viruSITE [46], an integrated database for viral genomics; the Viral Genome Organizer[5], implemented by the Canadian Viral Bioinformatics Research Centre, focusing on search for sub-sequences within genomes.

Another cluster of resources[6] is connected to the Chinese National Genomics Data Center (at the Beijing Institute of Genomics) and the China National GeneBank; these include the National Microbiology Data Center and the Genome Warehouse, as well as other virus database retrieval tools. Note that not all such resources have a related webpage in English, therefore can be difficult to use.

While the INSDC consortium provides full open access to sequences, the GISAID Initiative [43] was created in 2008 with the explicit purpose of offering an alternative to traditional public-domain data archives, as many scientists hesitated to share influenza data due to their legitimate concern about not being properly acknowledged, among others. GISAID hosts EpiFlu^{TM}, a large sequence database, which started its mission for influenza data and is now expanding with EpiCoV^{TM} having a particular focus on the SARS-CoV-2 pandemic (75,509 sequences for SARS-CoV-2 on August 1st, 2020). Some interesting portals have become interfaces to GISAID data with particular focuses: NextStrain [22] overviews emergent viral outbreaks based on the visualization of sequence data integrated with geographic information, serology, and host species; CoV-GLUE [44], contains a database of replacements, insertions and deletions observed in sequences sampled from the pandemic.

Many other accessory resources link to viral sequence data, including: drug databases, particularly interesting as they provide information about clinical studies (see ClinicalTrials[7]), protein sequences databases (e.g., UniProtKB/Swiss-Prot [38]), and cell lines databases (e.g., Cellosaurus [3]).

3 Conceptual modeling for viral genomics

We previously proposed the Genomic Conceptual Model (GCM, [8]), an Entity-Relationship diagram that recognizes a common organization for a limited set of concepts supported by most genomic data sources, although with different names and formats. The model is centered on the ITEM entity, representing an elementary experimental file of genomic regions and their attributes. Four views depart from the central entity, recalling a classic star-schema organization that is typical of data warehouses [9]; they respectively describe: i) the *biological* elements involved in the experiment: the sequenced sample and its preparation, the

[5] `https://4virology.net/virology-ca-tools/vgo/`
[6] NGDC: `https://bigd.big.ac.cn/`; CNGB: `https://db.cngb.org/`
[7] `http://clinicaltrials.gov/`

donor or patient; ii) the *technology* used in the experiment, including a specific assay (i.e., technique); iii) the *management* aspects: the projects/organizations involved in the preparation and production; iv) the *extraction* parameters used for internal selection and organization of items.

GCM is employed as a driver of integration pipelines for genomic datasets, fueling user search-interfaces (such as GenoSurf [10]). Lessons learnt from that experience include the benefits of having: a central *fact* entity that helps structuring the search; a number of surrounding *dimensions* capturing organization, biological and experimental conditions to describe the facts; a direct representation of a data structure suitable for conceptually organizing genomic elements and their describing information. a data layout that is easy to learn for first-time users and that helps the answering of practical questions (demonstrated in [4]).

We hereby propose the Viral Conceptual Model (VCM), which is influenced by our past experience with human genomes, with the comparable goal of providing a simple means of integration between heterogeneous sources. However, there are significant differences between the two conceptual models. The human DNA sequence is long (3 billions of base pairs) and has been understood in terms of *reference genomes* (named `h19` and `GRCh38`) to which all other information is referred, including genetic and epigenetic signals. Instead, viruses are many, their sequences are short (order of thousands of base pairs) and each virus has its own reference sequence; moreover, virus sequences are associated to a host sample of another species.

With a bird's eye view, the VCM conceptual model is centered on the Sequence entity that describes individual virus sequences; sequences are analyzed from the *biological* perspective (HostSample and Virus), the *technological* perspective (ExperimentType), and the *organizational* perspective (SequencingProject). Three other entities, NucleotideVariant, Annotation, and AminoacidVariant represent an *analytical* perspective of the sequence, allowing to analyze its characteristics, its sub-parts, and the differences with respect to reference sequences for the specific virus species. We next illustrate the central entity and the four perspectives.

Central entity. A viral Sequence can regard DNA or RNA; in either cases, databases of sequencing data write the sequence as a DNA *NucleotideSequence*: possible characters include guanine (G), adenine (A), cytosine (C), and thymine (T)[8], but also eleven "ambiguity" characters associated with all the possible combinations of the four DNA bases [37]. The sequence has a specific *Strand* (positive or negative), *Length* (ranging from hundreds to millions, depending on the virus), and a percentage of read G and C bases (*GC%*). As quality of sequences is very relevant to virologists, we also include the percentage of ambiguous bases (i.e., *N%*) to give a more complete information on reliability of the sequencing process. Each sequence is uniquely identified by an *AccessionID*, which is retrieved directly from the source database (GenBank's are usually formed by two capital letters, followed by six digits, GISAID by the string "EPI_ISL_" and six digits). Sequences can be complete or partial (as encoded by the Boolean flag

[8] In RNA sequencing databases uracil (U) is replaced with thymine (T).

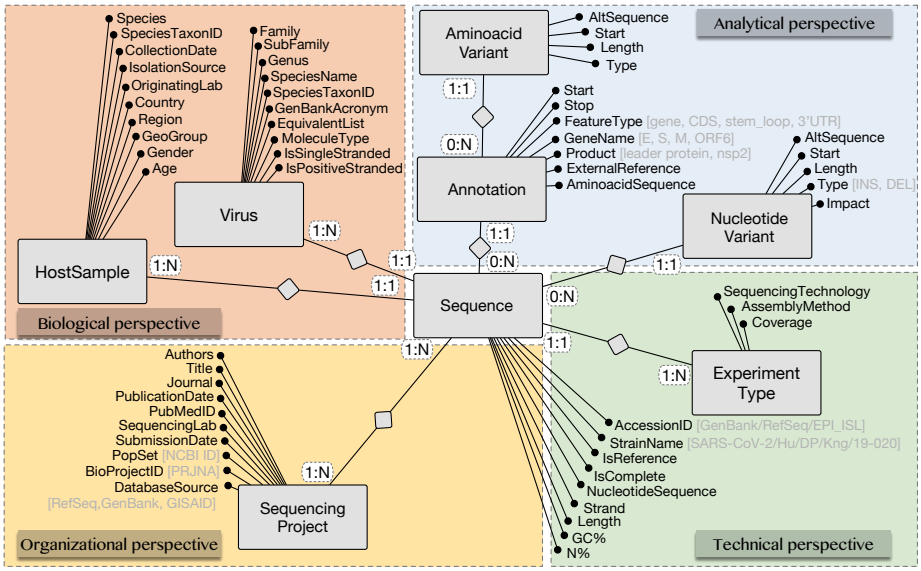

Fig. 2. The Viral Conceptual Model: the central fact SEQUENCE is described by four different perspectives (biological, technical, organizational and analytical).

IsComplete) and they can be a reference sequence (stored in RefSeq) or a regular one (encoded by *IsReference*). Sequences have a corresponding *StrainName* (or isolate) assigned by the sequencing laboratory, somehow hard-coding relevant information (e.g., hCoV-19/Nepal/61/2020 or 2019-nCoV_PH_nCOV_20_026).

Technological perspective. The sequence derives from one experiment or assay, described in the EXPERIMENTTYPE entity (cardinality is 1:N from the dimension towards the fact). It is performed on biological material analyzed with a given *SequencingTechnology* platform (e.g., Illumina Miseq) and an *AssemblyMethod*, collecting algorithms that have been applied to obtain the final sequence, for example: BWA-MEM, to align sequence reads against a large reference genome; BCFtools, to manipulate variant calls; Megahit, to assemble NGS reads. Another technical measure is captured by *Coverage* (e.g., 100x or 77000x).

Biological perspective. Each sequence belongs to a specific VIRUS, which is described by a complex taxonomy. The most precise definition is the *SpeciesName* (e.g., Severe acute respiratory syndrome coronavirus 2), corresponding to a *SpeciesTaxonID* (e.g., 2697049, according to the NCBI Taxonomy [17]), related to a simpler *GenBankAcronym* (e.g., SARS-CoV-2) and to many comparable forms, contained in the *EquivalentList* (e.g., 2019-nCoV, COVID-19, SARS-CoV2, SARS2, Wuhan coronavirus, Wuhan seafood market pneumonia virus, ...). The species belongs to a *Genus* (e.g., Betacoronavirus), part of a *SubFamily* (e.g., Orthocoronavirinae), finally falling under the most general category of *Family* (e.g., Coronaviridae). Each virus species corresponds to a specific

MoleculeType (e.g., genomic RNA, viral cRNA, unassigned DNA), which has either double- or single-stranded structure; in the second case the strand may be either positive or negative. These possibilities are encoded within the *IsSingleStranded* and *IsPositiveStranded* Boolean variables. An assay is performed on a tissue extracted from an organism that has hosted the virus for an amount of time; this information is collected in the HostSample entity. The host is defined by a *Species*, corresponding to a *SpeciesTaxonID* (e.g., 9606 for Homo Sapiens, according to the NCBI Taxonomy). The sample is extracted on a *CollectionDate*, from an *IsolationSource* that is a specific host tissue (e.g., nasopharyngeal or oropharyngeal swab, lung), in a certain location identified by the quadruple *OriginatingLab* (when available), *Region*, *Country*, and *GeoGroup* (i.e., continent) – for such attributes ISO standards may be used. In some cases information related to the *Age* and *Gender* of the individual donating the HostSample may also be available. Both entities of this perspective are in 1:N cardinality with the Sequence.

Organizational perspective. The entity SequencingProject describes the management aspects of the production of the sequence. Each sequence is connected to a number of studies, usually represented by a research publication (with *Authors*, *Title*, *Journal*, *PublicationDate* and eventually a *PubMedID* referring to the most important biomedical literature portal[9]). When a study is not available, just the *SequencingLab* and *SubmissionDate* are provided. In rare occasions, a project is associated with a *PopSet* number, which identifies a collection of related sequences derived from population studies (submitted to GenBank), or with a *BioProjectID* (an identifier to the BioProject external database[10]). We also include the name of *DatabaseSource*, denoting the organization that primarily stores the sequence. In this perspective all cardinalities are 1:N as sequences can be part of multiple projects; conversely, sequencing projects contain various sequences.

Analytical perspective. This perspective allows to store information that is useful during the secondary analysis of genomic sequences. The Nucleotide-Variant entity contains sub-parts of the main Sequence that differ from the reference sequence of the same virus species. They can be identified just by using the *AltSequence* (i.e., the nucleotides used in the analyzed sequence at position *Start* for an arbitrary *Length*, typically just equal to 1) and a specific *Type*, which can correspond to insertion (INS), deletion (DEL), substitution (SUB) or others. The content of the attributes of this entity is not retrieved from existing databases; instead it is computed in-house by our procedures. Indeed, we use the well-known dynamic programming algorithm of Needleman-Wunsch [31], that computes the optimal alignment between two sequences. From a technical point of view, we compute the pair-wise alignment of every sequence to the reference sequence of RefSeq (e.g., NC_045512 for SARS-CoV-2); from such alignment we then extract all insertions, deletions, and substitutions that transform (i.e., edit) the reference sequence into the considered sequence. Finally, we include the *Im-*

[9] `https://www.ncbi.nlm.nih.gov/pubmed/`
[10] `https://www.ncbi.nlm.nih.gov/bioproject/`

pact information, an annotation of the variant computed using SNPEff tool [11], which calculates the effect that the variant produces on known genes; a variant may, for example, be irrelevant, silent, produce small changes in the transcript or be deleterious for the transcript.

ANNOTATIONs include a number of sub-sequences, each representing a segment (defined by *Start* and *Stop* coordinates) of the original sequence, with a particular *FeatureType* (e.g., gene, peptide, coding DNA region, or untranslated region, molecule patterns such as stem loops and so on), the recognized *GeneName* to which it belongs (e.g., gene "E", gene "S" or open reading frame genes such as "ORF1ab"), the *Product* it concurs to produce (e.g., leader protein, nsp2 protein, RNA-dependent RNA polymerase, membrane glycoprotein, envelope protein...), and eventually related *ExternalReference* when the protein is present in a separate database such as UniProtKB. Additionally, for each ANNOTATION we also store the corresponding *AminoacidSequence*(encoded according to the notation of the International Union of Pure and Applied Chemistry[11]). Example codes are A (Alanine), D (Aspartic Acid), F (Phenylalanine).

The AMINOACIDVARIANT entity contains sub-parts of the *AminoacidSequence* stored in the specific ANNOTATION, which differ from the reference amino acids of the same virus species. These variants are calculated similarly to the NUCLEOTIDEVARIANTs (a comparable approach is used within CoV-GLUE. Also here we include the *AltSequence*, the *Start* position, the *Length*, and a specific *Type* (SUB, INS, DEL...).

4 Answering complex biological queries

In addition to very general questions that can be easily asked through our conceptual model (e.g., retrieve all viruses with given characteristics), in the following we propose a list of interesting application studies that could be backed by the use of our conceptual model. In particular, they refer to the SARS-CoV-2 virus, as it is receiving most of the attention of the scientific community. Fig. 3 represents the reference sequence of SARS-CoV-2[12], highlighting the major structural sub-sequences that are relevant for the encoding of proteins and other functions. It has 56 region ANNOTATIONS, of which Fig. 3 represents only the 11 genes (ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N, ORF10) plus the RNA-dependent RNA polymerase enzyme, with approximate indication of the corresponding coordinates. We next describe biological queries supported by VCM, from the easy to complex ones, typically suggested by existing studies.

Q1. The most common variants found in SARS-CoV-2 sequences can be selected for US patients; the query can be performed on entire sequences or only on specific genes.

[11] https://en.wikipedia.org/wiki/Nucleic_acid_notation#IUPAC_notation

[12] It represents the positive-sense, single-stranded RNA virus (from 0 to the 29903^{th} base) of NC_045512 RefSeq staff-curated complete sequence (*StrainName* "Wuhan-Hu-1"), collected in China from a "Homo Sapiens" HOSTSAMPLE in December 2019.

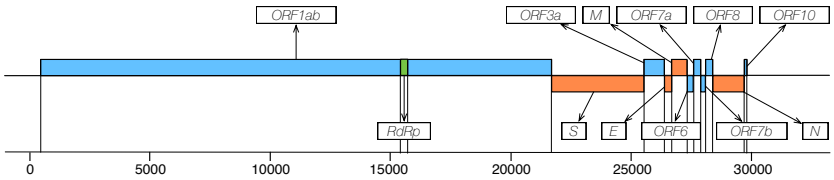

Fig. 3. Location of major structural protein-encoding genes (as red boxes: S = Spike glycoprotein, E = Envelope protein, M = Membrane glycoprotein, N = Nucleocapsid phospoprotein), accessory protein ORFs = Open Reading Frames (as blue boxes), and RNA-dependent RNA polymerase (RdRp) on the sequence of the SARS-CoV-2.

Q2. COVID-19 European patients affected by a SARS-CoV-2 virus can be selected when they have a specific one-base variant on the first gene (ORF1ab), indicated by using the triple <start, reference_allele, alternative_allele>. Patients can be distributed according to their country of origin. This conceptual query is illustrated in Fig. 4, where selected attribute values are specified in red, in place of attribute names in the ER model; values in NucleotideVariant show one possible example. *Country* is in blue as samples will be distributed according to such field.

Q3. According to [13], E and RdRp genes are highly mutated and thus crucial in diagnosing COVID-19 disease; first-line screening tools of 2019-nCoV should perform an E gene assay, followed by confirmatory testing with the RdRp gene assay. Conceptual queries are concerned with retrieving all sequences with mutations within genes E or RdRp and relating them to given hosts, e.g. humans affected in China.

Q4. Tang *et al.* [48] claim that there are two clearly definable "major types" (S and L) of SARS-CoV-2 in this outbreak, that can be differentiated by transmission rates. Intriguingly, the S and L types can be clearly distinguished by just two tightly linked SNPs (Single Nucleotide Polymorphisms, i.e., a specific kind of variant) at positions 8,782 (within the ORF1ab gene from C to T) and 28,144 (within ORF8 from T to C). Then, queries can correlate these SNPs to other variants or the outbreak of COVID-19 in specific countries (e.g., [20]).

Q5. To inform SARS-CoV-2 vaccine design efforts, it may be needed to track antigenic diversity. Typically, pathogen genetic diversity is categorized into distinct *clades* (i.e., a monophyletic group on a phylogenetic tree). These clades may refer to 'subtypes', 'genotypes', or 'groups', depending on the taxonomic level under investigation. In [20], specific sequence variants are used to define clades/haplogroups (e.g., the *A group* is characterized by the 20,229 and 13,064 nucleotides, originally C mutated to T, by the 18,483 nucleotide T mutated to C, and by the 8,017, from A to G). VCM supports all the information required to replicate the definition of SARS-CoV-2 clades requested in the study. Fig. 5 illustrates the conjunctive selection of sequences with all four variants corresponding to the *A clade group* defined in [20].

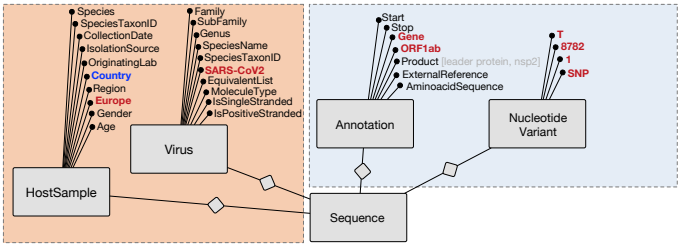

Fig. 4. Visual representation of query **Q2**.

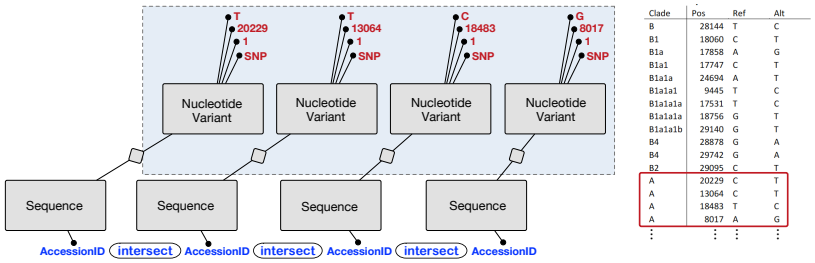

Fig. 5. Illustration of the selection predicate for the *A clade* [20], used in query **Q5**.

Q6. Morais Junior at al. [25] propose a subdivision of the global SARS-CoV-2 population into sixteen subtypes, defined using "widely shared polymorphisms" identified in nonstructural (nsp3, nsp4, nsp6, 27 nsp12, nsp13 and nsp14) cistrons, structural (spike and nucleocapsid), and accessory (ORF8) genes. VCM supports all the information required to replicate the definition of all such subtypes.

5 Related work

The genomics community has always made great use of specialized ontologies (see the collective OBO Foundry [45], including the Gene Ontology [12]). In addition, the use of conceptual modeling to describe genomics databases dates back to the late nineties, including a functional model for DNA databases named "associative information structure" [32], a model representing genomic sequences [30], and a set of data models for describing transcription/translation processes [35]. Later, a stream of works developed conceptual modeling-based data warehouses, including: the GEDAW UML Conceptual schema [21] (for a gene-centric data warehouse), the Genomics Unified Schema [2], the Genome Information Management System [14] (a genome-centric data warehouse), and the GeneMapper Warehouse [16] (integrating expression data from genomic sources).

More recently, there has been a solid stream of works dedicated to data quality-oriented conceptual modeling: [18] presents an ontological approach, [39] introduces the Human Genome Conceptual Model and [34] applies it to uncover relevant information hidden in genomics data lakes. Conceptual modeling has

been mainly concerned with aspects of the *human* genome, even when more general approaches were adopted; in [8] we presented GCM, describing metadata associated with genomic experimental datasets available for model organisms.

In the variety of types of genomic databases [15], aside the resources dedicated to humans [7], several ones are devoted to viruses [42]; however, very few works relate to conceptual data modeling. Among them, [47] considers host information and normalized geographical location, and [28] focuses on influenza A viruses. The closest work to ours, described in [44], is a flexible software system for querying virus sequences; it includes a basic conceptual schema[13]. In comparison, VCM covers more dimensions and attributes, which are very useful for supporting research queries on virus sequences.

6 Discussion and future developments

This paper responds to an urgent need: understanding the conceptual properties of SARS-CoV-2 so as to facilitate research studies. The model applies to any type of virus and can be used as a basis for the development of search systems. In the past, we first presented the conceptual model for human genomics [8], then we developed the Web-based search system GenoSurf [10]. Inspired by our previous experience, we practically employed the VCM by directly translating it into a logical schema and then into a solid relational database implementation that supports the ViruSurf search interface (`http://gmql.eu/virusurf`) in the back end. We started by focusing on the sequences of SARS-CoV-2; we include them from five sources, i.e., GenBank, RefSeq, COG-UK, GISAID, and NMDC. After SARS-CoV-2, we are progressively adding sequences of other virus species that could provide relevant comparative information for dealing with the COVID-19 pandemic, e.g., for vaccine and drug development.

While the need for data is pressing, there is also a need of conceptually well-organized information. In our broad vision, the availability of conceptual models for both human genomics and viruses will provide important opportunities for research, amplified to the maximum when human and viral sequences will be interconnected by the same human being, playing the role of host of a given virus sequence as well as provider of genomic and phenotype information.

In this direction, we are participating to the COVID-19 Host Genetics Initiative[14], aiming at *bringing together the human genetics community to generate, share and analyze data to learn the genetic determinants of COVID-19 susceptibility, severity and outcomes.* We are coordinating the production of a data dictionary for the phenotype definition, which is now being used as a reference by participating institutions, hosted by EGA [19], the European Genome-phenome Archive of EMBL-EBI[15]. When both phenotype and viral sequences datasets will be accessible, other more powerful studies will be possible. Some early findings

[13] `http://glue-tools.cvr.gla.ac.uk/images/projectModel.png`

[14] `https://www.covid19hg.org/`

[15] We coordinated about 50 active participants and released the "Freeze 1" version of the data dictionary on April 16, 2020 (`http://gmql.eu/phenotype/`).

have been already published connecting virus sequences with phenotypes, so far with very small datasets (e.g., [27] with only 5 patients, [29] with 9 patients, and [48] with 103 sequenced SARS-CoV-2 genomes). As reaffirmed by these works, there is need for additional comprehensive studies linking the viral sequences of SARS-CoV-2 to the phenotype of patients affected by COVID-19; such studies will be produced in the near future as result of ongoing clinical protocols.

In the future we will continue our modeling and integration efforts for virus genetics in the context of humans, by interacting with the community of scholars who study viruses. We may add more discovery-oriented entities to the model, that could be of use in a future scenario, e.g., a new pandemic offspring. We will expand our schema in several directions: 1) we will add both validated and predicted epitopes (i.e., antigen parts to which antibodies attach) with their sequence, lineage, host, evidence, reference or algorithm, type of response; 2) we will link entities to specific external/ontological knowledge, which is being discovered nowadays, e.g., each variant to COVID-19 morbidities and each epitope to the specific strain and geographic population it refers to; 3) we will also link sequences to complete tree-structured taxonomies of viruses and host organisms.

In this way, we will be able to cover a wider spectrum of domain specific queries. A user researching on diagnosis could ask, for example, what sequence patterns are unique to the whole or sub-part of the database (i.e., do not appear in viruses within the database). Whereas, a user working on vaccine development could be interested in what are the epitopes that cover the whole database or a partition of it, for MHC types prevalent in different infected humans. Possibly, other dimensions will be necessary, such as drug resistance information and drug resistance-associated mutations.

Acknowledgements. This research is funded by the ERC Advanced Grant 693174 GeCo (Data-Driven Genomic Computing), 2016-2021. The authors would like to thank Ilaria Capua, Luca Ferretti, Alice Fusaro, Susanna Lamers, Francesca Mari, Carla Mavian, Alessandra Renieri, Stephen Tsui, and Limsoon Wong for their precious contributions during the phase of requirements elicitation and for their inspiration towards future developments of this research.

References

1. Amid, C., et al.: The European Nucleotide Archive in 2019. Nucleic acids research **48**(D1), D70–D76 (2020)
2. Babenko, V., et al.: GUS the genomics unified schema a platform for genomics databases, `http://www.gusdb.org/`, (August 1st, 2020, date last accessed)
3. Bairoch, A.: The Cellosaurus, a cell-line knowledge resource. Journal of biomolecular techniques: JBT (2018)
4. Bernasconi, A., et al.: Exploiting conceptual modeling for searching genomic metadata: A quantitative and qualitative empirical study. In: Guizzardi, G., et al. (eds.) Advances in Conceptual Modeling. pp. 83–94. Springer International Publishing, Cham (2019)
5. Bernasconi, A., et al.: From a conceptual model to a knowledge graph for genomic datasets. In: Laender, A.H.F., et al. (eds.) Conceptual Modeling. pp. 352–360. Springer International Publishing, Cham (2019)

6. Bernasconi, A., et al.: META-BASE: a novel architecture for large-scale genomic metadata integration. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020)
7. Bernasconi, A., et al.: The road towards data integration in human genomics: players, steps and interactions. Briefings in Bioinformatics (2020)
8. Bernasconi, A., et al.: Conceptual modeling for genomics: Building an integrated repository of open data. In: Mayr, H.C., et al. (eds.) Conceptual Modeling. pp. 325–339. Springer International Publishing, Cham (2017)
9. Bonifati, A., et al.: Designing data marts for data warehouses. ACM Transactions on Software Engineering and Methodology **10**(4), 452–483 (2001)
10. Canakoglu, A., et al.: GenoSurf: metadata driven semantic search system for integrated genomic datasets. Database **2019** (2019)
11. Cingolani, P., et al.: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly **6**(2), 80–92 (2012)
12. Consortium, G.O.: The gene ontology resource: 20 years and still going strong. Nucleic acids research **47**(D1), D330–D338 (2019)
13. Corman, V.M., et al.: Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance **25**(3) (2020)
14. Cornell, M., et al.: GIMS: an integrated data storage and analysis environment for genomic and functional data. Yeast **20**(15), 1291–1306 (2003)
15. De Francesco, E., et al.: A summary of genomic databases: overview and discussion. In: Biomedical Data and Applications, pp. 37–54. Springer (2009)
16. Do, H.H., et al.: Flexible integration of molecular-biological annotation data: The GenMapper approach. In: International Conference on Extending Database Technology. pp. 811–822. Springer (2004)
17. Federhen, S.: The NCBI taxonomy database. Nucleic acids research **40**(D1), D136–D143 (2012)
18. Ferrandis, A.M.M., et al.: Applying the principles of an ontology-based approach to a conceptual schema of human genome. In: International Conference on Conceptual Modeling. pp. 471–478. Springer (2013)
19. Flicek, P., et al.: The European Genotype Archive: Background and implementation [white paper] (2007)
20. Gudbjartsson, D.F., et al.: Spread of SARS-CoV-2 in the icelandic population. New England Journal of Medicine (2020)
21. Guerin, É., et al.: Integrating and warehousing liver gene expression data and related biomedical resources in gedaw. In: International Workshop on Data Integration in the Life Sciences. pp. 158–174. Springer (2005)
22. Hadfield, J., et al.: Nextstrain: real-time tracking of pathogen evolution. Bioinformatics **34**(23), 4121–4123 (2018)
23. Hatcher, E.L., et al.: Virus variation resource–improved response to emergent viral outbreaks. Nucleic acids research **45**(D1), D482–D490 (2017)
24. Hulo, C., et al.: ViralZone: a knowledge resource to understand virus diversity. Nucleic acids research **39**(suppl_1), D576–D582 (2011)
25. Junior, I.J.M., et al.: The global population of SARS-CoV-2 is composed of six major subtypes. bioRxiv (2020)
26. Koonin, E.V., et al.: Global organization and proposed megataxonomy of the virus world. Microbiology and Molecular Biology Reviews **84**(2) (2020)
27. Lescure, F.X., et al.: Clinical and virological data of the first cases of COVID-19 in Europe: a case series. The Lancet Infectious Diseases (2020)

28. Lu, G., et al.: Influenza A virus informatics: genotype-centered database and genotype annotation. In: Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007). pp. 76–83. IEEE (2007)
29. Lu, R., et al.: Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet **395**(10224), 565–574 (2020)
30. Médigue, C., et al.: Imagene: an integrated computer environment for sequence annotation and analysis. Bioinformatics (Oxford, England) **15**(1), 2–15 (1999)
31. Needleman, S.B., et al.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology **48**(3), 443–453 (1970)
32. Okayama, T., et al.: Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library. Bioinformatics (Oxford, England) **14**(6), 472–478 (1998)
33. O'Leary, N.A., et al.: Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research **44**(D1), D733–D745 (2015)
34. Palacio, A.L., et al.: A method to identify relevant genome data: conceptual modeling for the medicine of precision. In: International Conference on Conceptual Modeling. pp. 597–609. Springer (2018)
35. Paton, N.W., et al.: Conceptual modelling of genomic information. Bioinformatics **16**(6), 548–557 (2000)
36. Pickett, B.E., et al.: ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic acids research **40**(D1), D593–D598 (2012)
37. Nomenclature Committee of the International Union of Biochemistry (NC-IUB): Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. Proceedings of the National Academy of Sciences of the United States of America **83**(1), 4–8 (1986)
38. UniProt Consortium: UniProt: a worldwide hub of protein knowledge. Nucleic acids research **47**(D1), D506–D515 (2019)
39. Román, J.F.R., et al.: Applying conceptual modeling to better understand the human genome. In: International Conference on Conceptual Modeling. pp. 404–412. Springer (2016)
40. Sayers, E.: The E-utilities in-depth: parameters, syntax and more. Entrez Programming Utilities Help [Internet] (2009), `https://www.ncbi.nlm.nih.gov/books/NBK25499/`
41. Sayers, E.W., et al.: GenBank. Nucleic acids research **47**(D1), D94–D99 (2019)
42. Sharma, D., et al.: Unraveling the web of viroinformatics: computational tools and databases in virus research. Journal of virology **89**(3), 1489–1501 (2015)
43. Shu, Y., et al.: GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance **22**(13) (2017)
44. Singer, J., et al.: Cov-glue: A web application for tracking sars-cov-2 genomic variation (2020)
45. Smith, B., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology **25**(11), 1251–1255 (2007)
46. Stano, M., et al.: viruSITE-integrated database for viral genomics. Database **2016** (2016)
47. Tahsin, T., et al.: Named entity linking of geospatial and host metadata in genbank for advancing biomedical research. Database **2017** (2017)
48. Tang, X., et al.: On the origin and continuing evolution of SARS-CoV-2. National Science Review (2020)