IMGT-KG: A Knowledge Graph for Immunogenetics

Gaoussou Sanou[1,2(✉)], Véronique Giudicelli[1], Nika Abdollahi[1], Sofia Kossida[1], Konstantin Todorov[2], and Patrice Duroux[1]

[1] IGH/University of Montpellier/CNRS, Montpellier, France
{gaoussou.sanou,veronique.giudicelli,nika.abdollahi,sofia.kossida,patrice.duroux}@igh.cnrs.fr
[2] LIRMM/University of Montpellier/CNRS, Montpellier, France
{gaoussou.sanou,konstantin.todorov}@lirmm.fr

Abstract. Knowledge graphs are emerging as one of the most popular means for data federation, transformation, integration and sharing, promising to improve data visibility and reusability. Immunogenetics is the branch of life sciences that studies the genetics of the immune system. Although the complexity and the connected nature of immunogenetics data make knowledge graphs a prominent choice to represent and describe immunogenetics entities and relations, hence enabling a plethora of applications, little effort has been directed towards building and using such knowledge graphs so far. In this work, we present the IMGT Knowledge Graph (IMGT-KG), the first of its kind FAIR knowledge graph in immunogenetics. IMGT-KG acquires and integrates data from different immunogenetics databases, hence creating links between them. Consequently, IMGT-KG provides access to 79 670 110 triplets with 10 430 268 entities, 673 concepts and 173 properties. IMGT-KG reuses many existing terms from domain ontologies or vocabularies and provides external links to other resources of the same domain, as well as a set of rules to guide inference on nucleotide sequence positions by applying Allen Interval Algebra. Such inference allows, for example, reasoning about genomics sequence positions. IMGT-KG fills in the gap between genomics and protein sequences and opens a perspective to effective queries and integrative immuno-omics analyses. We make openly and freely available IMGT-KG with detailed documentation and a Web interface for access and exploration.

Keywords: Immunogenetics and immunoinformatics · Ontology · Knowledge graphs · SPARQL endpoint · Reasoning rules · Semantic web

1 Introduction

Immunogenetics has the mission to decrypt the genetics of the immune system and immune responses. To take an example, immunogenetics plays a crucial role

U. Sattler et al. (Eds.): ISWC 2022, LNCS 13489, pp. 628–642, 2022.
https://doi.org/10.1007/978-3-031-19433-7_36

in the current context marked by the COVID-19 pandemic. The genetic basis of the immune response in COVID-19 cases may explain the inter-individual disease variability and provide a way to classify patients in different severity profiles according to the presence or absence of genetic variants [16]. In addition, the understanding of such genetics bases contributes to the rapid development of vaccines.

IMGT®, the *International ImMunoGeneTics Information System* ®, is an international data reference in the immunogenetics field, particularly in the management of the adaptive immune response data [13,14]. Over the past 30 years, IMGT elaborated several high-quality databases, web resources and tools for understanding and cracking the adaptive immune response, now considered as a reference in the field. IMGT® offers a knowledge management system, that allows for a standardised annotation of immunogenetics entities from genomic to protein data by using a formal vocabulary: the IMGT-ONTOLOGY [12]. Based on the type of the immunogenetics entity (genomic or protein), IMGT® provides five different databases: two genomic databases (IMGT/LIGM-DB, IMGT/GENE-DB) and three protein databases (IMGT/2Dstructure-DB, IMGT/3Dstructure-DB and IMGT/mAb-DB), described in the following section in more detail. These databases are freely accessible via different query form-like interfaces[1] [13,14]. According to the connected nature of immunogenetics information, federating and integrating different entities in a central knowledge base will not only give a way to make integrative analyses (via expressive, complete and rich queries like, for example, *"find all proteins with their gene and alleles with a particular genomic reference sequence"*), but will also provide a way to discover new facts like, for example, the particular genomic sequence associated to a protein structure.

To fill in this gap, we introduce IMGT-KG, the first Findable, Accessible, Interoperable and Reusable (FAIR) knowledge graph (KG) in immunogenetics, which provides access to structured immunogenetics data based on the IMGT® resources. IMGT-KG is built and published following the W3C recommendations and best practices [4]. The data model of IMGT-KG is an extended version of the IMGT-ONTOLOGY [12]. For interoperability purposes, IMGT-KG also reuses terms from various biomedical resources. To generate the IMGT-KG, we collect and lift data from the IMGT databases and instantiate the data model, applying a reasoner to check its consistency and to enrich it by inferring new facts. In addition, we use a set of rules based on Allen's interval algebra [1], to infer the different spatial relations between sequence features.[2] Hence, IMGT-KG enables advanced exploration of immunogenetics data via queries such as *"Find all protein chains, domains, the associated allele and its genomic reference sequence"* or *"find the epitopes on the Immune Epitope Database (IEDB) which are in interaction with a particular structure of IMGT-KG"* or *"Find the structure that interacts with the COVID-19 spike and their associated chains,*

[1] https://www.imgt.org/.

[2] A feature is a region in a sequence—a succession of nucleotide or amino acids—with coordinates (start and end value) and a label.

genes, alleles and genomic reference sequences". Currently, IMGT-KG contains data from IMGT/LIGM-DB, IMGT/GENE-DB, IMGT/2Dstructure-DB, IMGT/3Dstructure-DB. It provides access to 79 670 110 triplets with 10 430 268 entities, 673 concepts and 173 properties. We make openly and freely available IMGT-KG with adequate documentation.[3]

In Sect. 2, we provide basic background in biology. We describe the IMGT® resources and databases used to generate IMGT-KG in Sect. 3. In Sect. 4, we detail the construction of the KG, while Sect. 5 gives examples of use-cases. We present related resources in Sect. 6, before we conclude in Sect. 7.

2 Background

This section lays down some fundamentals in molecular biology, needed for the understanding of our work.

Our body is constituted of tissues and tissues are made of cells. Every cell has a kernel that contains chromosomes. Each chromosome contains deoxyribonucleic acid (DNA), which is coded with nucleotides represented by four letters: A,T,C,G. The DNA has a double helix structure and can be considered as the cell manual. The information contained in one DNA helix is transcribed to RNA (ribonucleic acid), what is known as a transcription process. The RNA is then translated to a protein chain or polypeptide,[4] known as a translation process. The protein chains will fold to create a 3D conformation: protein structure. The processes of transcription and translation are depicted in Fig. 1. A protein is coded by one or more genes and is made with a succession of residues or amino acids. A protein can have different units called chains and a chain can be constituted by domains and regions. A gene is a DNA sequence (genomics level) that can be potentially transcribed and/or translated (protein level). A gene can have multiple versions marked by mutations[5] called alleles. A gene is localised in a particular place on a chromosome called a locus.

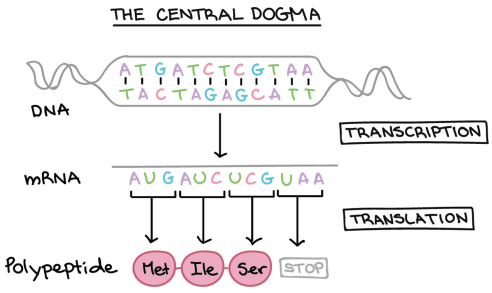

Fig. 1. Central dogma of molecular biology (from biocore).

[3] https://www.imgt.org/imgt-kg/, gives access to the entire IMGT®database.
[4] Successions of amino acids.
[5] Either an insertion of nucleotide, either a deletion of nucleotide or substitution of nucleotide.

When an organism that can produce a disease, known as a pathogen, enters our body, a defensive strategy is put in place: our immune response produces proteins thanks to the B cells, called immunoglobulins (IG) or antibodies and thanks to the T cells, called T cell receptors (TR). These proteins recognise the pathogen thanks to their motifs (epitopes) and trigger their destruction. Each produced antibody will be specific to the pathogen. This is the so-called adaptive immune response.

3 IMGT®: A Knowledge Management System for Immunogenetics Data

IMGT® is an information system specialised in the management of the diversity and complexity of the immunoglobulins or antibodies, T cell receptors, major histocompatibility (MH) and superfamilies of IG (IgSF) of MH (MhSF) and immune system proteins (RPI) [13,14]. It manages immunogenetics data through 3 axes:

- the first axis aims to decipher the IG and TR loci, genes and alleles in the genome of jawed vertebrates.
- the second axis concerns the exploration and analysis of the expressed IG and TR repertoires based on comparison with IMGT reference directories in normal and pathological situations.
- the third axis aims to analyse the amino acid changes and functions of 2D and 3D structures of engineered antibody and TR.

IMGT® provides a standard way to represent immunogenetics data based on the IMGT-ONTOLOGY [12] a vocabulary that describes immunogenetics data from genomics (nucleotide) data level to protein (three-dimensional structure) level. This vocabulary allows IMGT®, to build a rich knowledge system with 7 databases in total, 17 tools and more than 20,000 web pages and documents [13, 14]. The vocabulary terms allow for the identification, description, classification, localisation, orientation, acquisition and numbering of immunogenetics data. The databases of interest for this study are:

- IMGT/LIGM-DB[6] [10] and IMGT/GENE-DB[7] [11]. The former (246951 entries) provides standardised terms to annotate immunogenetics data including IG, TR and MH nucleotide sequences from human and other vertebrate species. The latter (9089 entries) stores the IG and TR genes curated with all IMGT identified alleles.
- IMGT/3Dstructure-DB and IMGT/2Dstructure-DB,[8] [8] and the monoclonal antibodies database IMGT/mAb-DB.[9] The structure databases (8260 entries) provide an access to protein structures and their related chains and domains. IMGT/mAb-DB is a monoclonal antibody database (1257 entries) for therapeutic purposes.

[6] https://www.imgt.org/ligmdb/.
[7] http://www.imgt.org/genedb/.
[8] https://www.imgt.org/3Dstructure-DB/.
[9] https://www.imgt.org/mAb-DB/.

4 IMGT-KG Construction

IMGT-KG is constructed by using an extended version of the IMGT-ONTOLOGY [12] as data model and the IMGT® genomic and protein databases as data sources. The construction comprises three steps, as illustrated in Fig. 2: i) defining a data model based on the IMGT-ONTOLOGY by reusing existing terms when possible and linking equivalent terms with the sameAs property, ii) instantiating the model and generating the KG, iii) checking the consistency of the KG by the help of a reasoning engine and completing the KG with newly inferred facts.

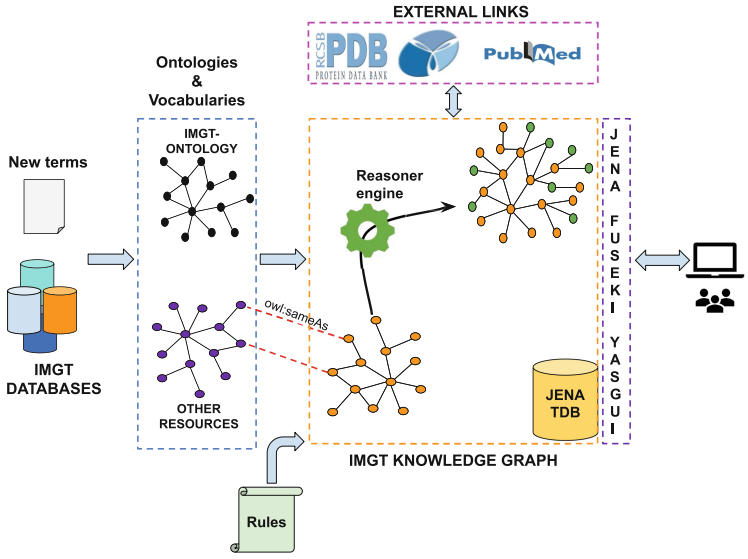

Fig. 2. Pipeline of IMGT-KG construction

The KG is built by using W3C best practices and standards.[10] We generate URIs[11] for the IMGT-KG by using the existing URI pattern of the IMGT-ONTOLOGY: http://www.imgt.org/imgt-ontology#. The implementation of IMGT-KG is made by the means of the Apache Jena framework.[12] To take advantage of existing terms, we use OntoFox, a web-based application, that allows the extraction of terms from an ontology by keeping their related properties, annotations and classes [19].[13] It provides also a means to serialise the extracted terms in the W3C recommendation format. To interact with the

[10] RDF, RDFS, and OWL.
[11] Uniform Resource Identifier.
[12] https://jena.apache.org/.
[13] http://ontofox.hegroup.org.

IMGT® databases, we use an Object-Relational Mapping (ORM) model to map the relational database to an object model, then we access to the data with a Java Persistence API (JPA). In order to check the consistency of the KG, we use Pellet[14] an OWL2 DL (Description Logic) reasoner. Additionally, we make OWL2 DL inference with the latter and use the Jena rule engine to compute deductions of a defined rule set.

4.1 IMGT-KG Data Model's Definition and (Re-)used Ontologies and Vocabularies

The IMGT-KG data model provides the necessary vocabulary and axioms to describe immunogenetics entities and their relations. We updated the previous version of the IMGT-ONTOLOGY and defined new terms (identification, description, classification, localisation, orientation, acquisition and numbering) for structuring immunogenetics data following the W3C recommendations. In order to enhance interoperability and link our KG with other external resources [3,5], we reuse existing terms when it is possible and make equivalence links with certain Sequence Ontology (SO) terms.[15] Major parts of these terms come from the OBO foundry ontologies (see Figs. 3 and 4):[16]

- Relation Ontology (RO) [18] provides more than 400 terms for defining relations across a variety of domains.
- Feature Annotation Location Description Ontology (FALDO) [6] provides terms to describe a sequence based on location (position/coordinates) of its different features. This is particularly useful to annotate a position or coordinates of a feature.
- Genotype Ontology (GENO) is an ontology that provides terms covering genotype description and genetic variations in model organisms.[17]
- NCI Thesaurus (NCIt) is a reference terminology and core biomedical ontology, providing 120,000 key biomedical concepts with a rich set of terms, codes, 115,000 textual definitions, and over 400,000 inter-concept relationships.[18]
- NCBI Taxonomy provides terminology that covers classification and organisms nomenclature.[19]
- Sequence Ontology (SO) [9] provides a controlled and standardised vocabulary for sequence annotation, aiming to unify all sequence annotations. SO uses the concept of feature in sequence annotation, and provides links that point to some IMGT labels [12]. In fact, 64 terms of SO have synonyms in IMGT labels.

[14] https://github.com/stardog-union/pellet.
[15] http://www.sequenceontology.org/.
[16] https://obofoundry.org/.
[17] https://github.com/monarch-initiative/GENO-ontology.
[18] https://ncit.nci.nih.gov/ncitbrowser/.
[19] https://www.ncbi.nlm.nih.gov/taxonomy.

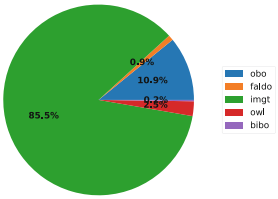

Fig. 3. Concepts in IMGT-KG (Color figure online)

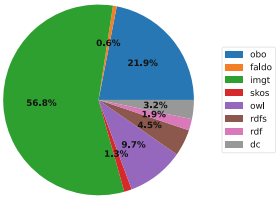

Fig. 4. Properties in IMGT-KG (Color figure online)

Figure 5 describes the IMGT-KG data model. We use the following colour code in the figure representing the model:

- The light blue colour represents the Gene and its associated knowledge. In fact, a Gene can be a member of (obo:RO_0002350) IMGT Group, SubGroup and/or Clan. A Clan or a SubGroup can also be a member of a Group. A concept of Gene is associated with a type (variable, diversity, joining, constant, conventional) and a structure type. A Gene has at least one Allele (obo:GENO_0000413) and an Allele can have a functionality type that states its functionality (functional, Open Reading Frame (ORF) or pseudogene). An allele is associated to a coding region (faldo:Region), this region can be a reference sequence or a sequence from the literature. Each Gene is ordered in a Locus and belongs to a taxon (obo:NCBITaxon_1).
- At the locus level (orange), a Locus has a location type (major locus, orphon set etc.), it is also member of a given chromosome (obo:SO_0000340) for a given taxon. This chromosome is member of a given assembly (obo:SO_0001248) and the assembly has a version number (obo:SWO_0004000), data origin (obo:NCIT_C103167) and belongs to a taxon.
- Light green colour represents the sequence features and their related description. A feature is a sequence Region (faldo:Region) with a location (faldo:ExactPosition) and an IMGT label. Every location has a start value (obo:GENO_0000894) and an end value (obo:GENO_0000895). A gene feature part of (obo:BFO_0000050) a genomic sequence (obo:GENO_0000960) with an accession number (obo:NCIT_C25402). The feature with associated IMGT prototype label (e.g. V-GENE) contains other small features thanks to the imgt:isInPrototype relation and is related to a genomics sequence with the imgt:isPrototypeInSeq. Also, the feature with the IMGT cluster label contains the features with IMGT prototype labels thanks to imgt:isInCluster relation and is related to a genomics sequence with the imgt:isClusterInSeq.
- The pink colour introduces to the protein level and characterises the protein Chain and its related properties. A chain (obo:NCIT_C41207) can have protein domains (obo:NCIT_C13303) with a domain type. It has also regions and residues (obo:NCIT_C48795) with the associated amino acid (obo:CHEBI_33709) and an IMGT numbering. Every protein chain domain has an IMGT label and a location. The associated region of a Chain is the

reference sequence of an Allele with a similarity score. Chains belong to a taxon and a structure (obo:NCIT_C13303).

– The white colour represents the protein Structure and its associated description. A Structure (obo:NCIT_C13303) that can belong to a complex (NCIT_C19398), having IMGT label and a molecular component. A Structure is attached to an entry of an amino-acid sequence (obo:GENO_0000720). This sequence has an accession number, a related bibliographic reference and an acquisition experiment.

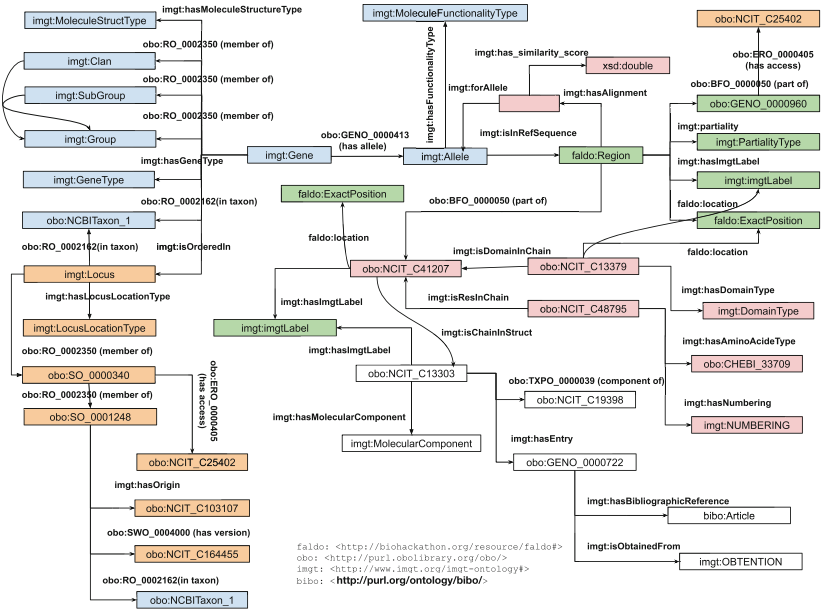

Fig. 5. An overview of the IMGT-KG data model without annotation properties. Details are given on the KG interface: https://www.imgt.org/imgt-kg/kgdescription.html.

4.2 IMGT-KG Data Model's Population

Once we have the schema of our knowledge graph, we proceed to populate it, i.e. add facts to the graph by using data from the IMGT databases described above thanks to the ORM model and the JPA. In the genomic level, we do JPQL queries in IMGT/GENE-DB in order to retrieve information about genes including their classification (group, subgroup, clan, allele), their localisation (locus, chromosome, assembly) and their identification (gene type, molecule type etc.) and we associate the description (IMGT labels) of their genomic nucleotide sequence thanks to the information from IMGT/LIGM-DB.[20] On protein level,

[20] Java Persistence Query Language.

we query the Structure databases in order to retrieve information about structures, chains, domains, regions, residues, alleles and other related properties and we associate the alleles of structure databases to alleles in IMGT/GENE-DB. Thanks to the JPA, we harvest the result of these queries and instantiate the data model with Jena Ontology API, then we serialise the data in the turtle format.

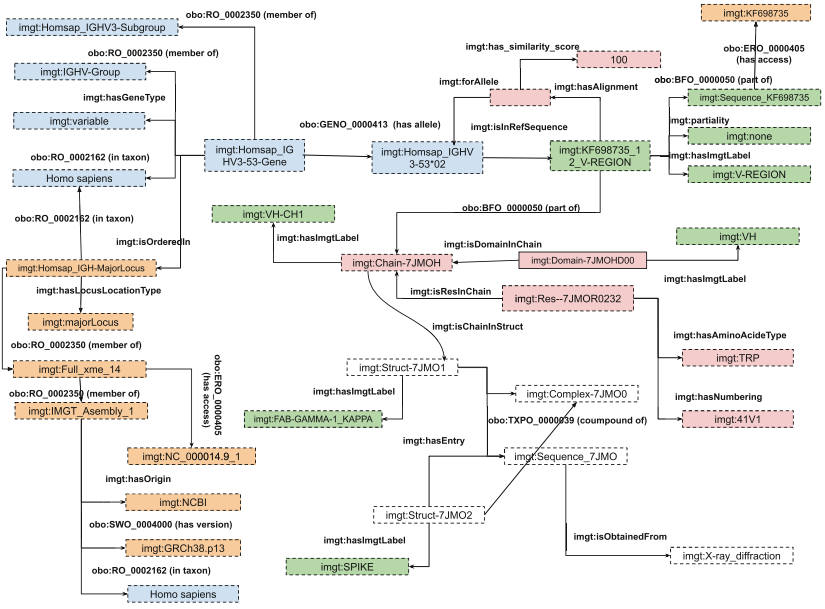

Fig. 6. Instances in the IMGT-KG related to the COVID-19 SPIKE protein of the structure 7JMO

As an example, Fig. 6 shows the representation of the SPIKE protein in the COVID-19 case. We have the structure 7JMO2 annotated with the SPIKE IMGT label; it is a component of the 7JMO0 complex and is associated to the 7JMO amino acid sequence obtained from X-ray diffraction. We have also the 7JMO1 structure which belongs to the same complex and has the same amino acid. The 7JMO1 labelled with FAB-GAMMA-1_KAPPA, contains the 7JMOH chain (VH-CH1), this chain has 7JMOHD00 domain (VH), a tryptophan residue at position 41 and a coding region KF698735_12_V-REGION. This region is not partial and is part of the genomic sequence KF698735. The region is also a reference sequence of an allele Homsap_IGHV3-53*02 with an alignment similarity score of 100. The allele is a variant of a joining gene Homsap_IGHV3-53-Gene which is member of the IGHV-Group and the Homsap_IGHV3-Subgroup. This gene belongs to *Homo sapiens (human)* taxon and is ordered in the major locus Homsap_IGH-MajorLocus. This locus belongs to the same taxon and is member of the chromosome 14. The chromosome comes from a NCBI assembly of the same human taxon.

We generated a KG with 79 670 110 triplets, 10 430 268 entities, 15 848 105 distinct subjects, 21 861 727 distinct objects, 673 distinct concepts or classes and 171 distinct properties or relations. The top 10 instantiated concepts and properties are presented in Fig. 7, 8.

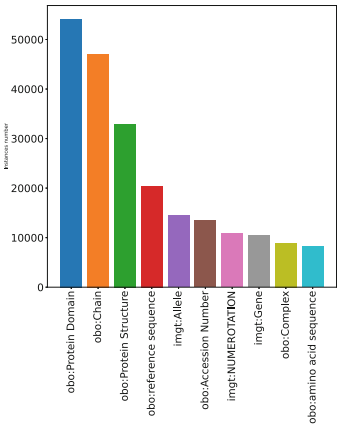

Fig. 7. Top 10 instantiated concepts in IMGT-KG

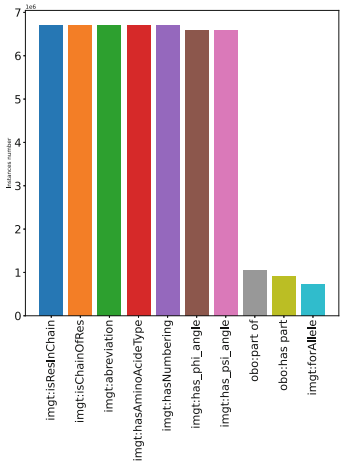

Fig. 8. Top 10 instantiated properties in IMGT-KG

4.3 IMGT-KG Enrichment: Rules, Consistency and Inference

After the KG is generated, we make sure that no data model violation is encountered in the KG using Pellet for consistency check. We formalise a set of rules in order to make deduction on the spatial position of sequence by the means of Jena rule engine API, then we apply Pellet reasoner in order to deduce new facts in the KG.

Allen's Interval Algebra and Rules. In 1989, Allen et al. provided a way to formalise and reason over the time interval [1]. Called Allen's Logic Interval, this powerful tool allows reasoning over time events. For example, suppose the start of event A is the end of another event B, then event A meets event B. Thus, 13 decidable relations were formalised by Allen to describe all types of events that may occur over time interval [1]. Similarly, we transpose this logic on genomic sequence positions (Fig. 9). In fact, the genomic sequence spatial position can be considered as interval with a start and end point, consequently Allen's Logic Interval turns out to be the most suitable to make automatic deductions in our genomic sequence position. Hence, we formalise a set of rules in order to reason over genomic sequence position.

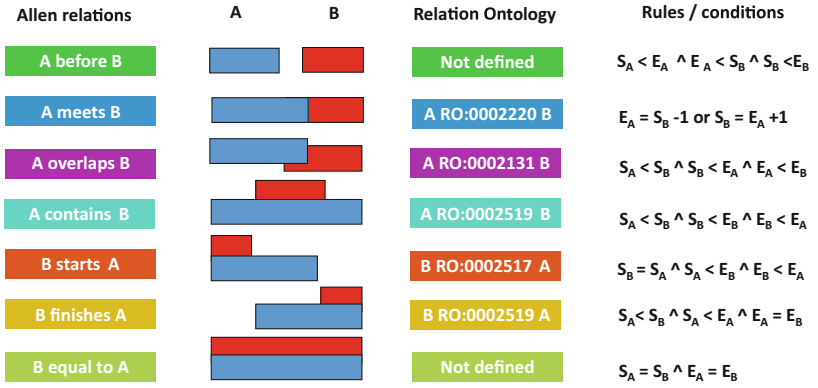

Fig. 9. Allen's interval rules adaptation, S = start and E = end.

For example, the Fig. 10 shows an example of a prototype (V-GENE) which is a topological model to describe a nucleotide sequence architecture[21]. The prototype allows the annotation of IMGT® genomic data. The application of the defined rules in Fig. 9 on the V-GENE sequence allows to infer or deduce that the L-PART2 meets (obo:RO_0002220) the V-INTRON and FR1-IMGT and it starts (obo:RO_0002517) the V-EXON and finishes (obo:RO_0002519) the L-INTRON-L.

Being able to make these deductions allows not only to do spatial reasoning over genomic sequences but also to verify if all features in a genomic sequence are well positioned in order to detect annotation errors.

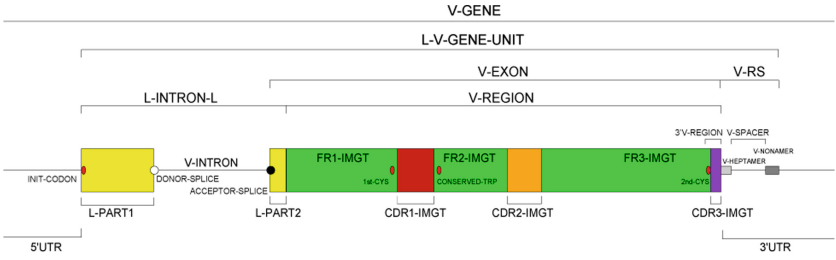

Fig. 10. A V-GENE prototype [12].

Reasoning on IMGT-KG. Once we formalised Allen rule set, we inject them in the KG and compute automatic deductions on genomic sequence positions by the means of the Jena rule engine API. Then, we apply Pellet reasoner to not only check the consistency of the KG but also to infer or complete the KG with inferred fact.

[21] A nucleotide sequence consists of many features with a position and IMGT label.

We built IMGT-KG, an enriched and FAIR KG that integrates a high-quality immunogenetics data harvested from five IMGT® databases. We use Jena triplestore TDB2[22] to store our triplets and we provide an access endpoint thanks to a SPARQL server: Fuseki2.[23] IMGT-KG provides access to over 79 million of triplets without inferences and more than 97 million with inferences.

5 IMGT-KG in Use

The IMGT® resources are largely adopted by the international immunogenetics scientific community, establishing themselves as the main reference in the field over the past 30 years. Therefore, IMGT-KG, which integrates and builds on top of these resources, also targets this community, enhancing the adoption of semantic web technologies to a field which is currently underrepresented.[24]

The IMGT-KG fills the gap between different IMGT® databases and consequently unifies the genomics and proteins data. Hence, the centralisation of the IMGT® databases in a FAIR KG allows for more query possibilities and enables the discovery of new knowledge. We provide free access to our KG via a well-documented web interface. This interface allows the user to explore different facets of IMGT-KG. The welcome page introduces users to IMGT-KG and provides information about the IMGT-KG team. The IMGT-KG Description page provides details on the KG data model. We also provide a description of the IMGT-KG dataset via the VoID Vocabulary.[25] The IMGT-KG Statistics page provides detailed statistics and chart plots about the KG. The IMGT-KG Data Access page, powered by YASGUI,[26] provides open access to the resource. There, users will find various SPARQL query examples. The IMGT-KG Model documentation and visualisation page provides useful documentation and a taxonomy visualisation of our data model. These pages are generated with Ontospy.[27]

We provide a set of use-case scenarios. The corresponding queries for each of the scenarios can be found on IMGT-KG's webpage: https://www.imgt.org/imgt-kg/kgyasgui.html.

Scenario 1: Assume that a user searches information on the genes/alleles functionality and why they are not functional. She/he must first select the alleles/genes of interest, the associated reference sequence and its belonging entity, the functionality (here P for pseudogene). To explain the absence of some functionality, she/he must check the qualifier associated to the entity with the imgt:has_imgt_qualifier and filter based on the qualifier that contains "pseudo" terms for example.

Scenario 2: Suppose an immunogenetics researcher wants to explore a specific structure and the associated external links, for example find some structures

[22] https://jena.apache.org/documentation/tdb/.
[23] https://jena.apache.org/documentation/fuseki2/index.html.
[24] We plan to communicate our results and resources to the biological community.
[25] https://www.w3.org/TR/void/.
[26] https://yasgui.triply.cc/.
[27] http://lambdamusic.github.io/Ontospy/.

and their related bibliographies with the PUBMED identifier, the associated visualisation of the structure, the same structure in the Protein Database (PDB) and the probable epitope of the structure. For that, the user must first select the entry associated to the structure, the related properties of the entry (pdb_link, jmol_visualisation, bibliography properties), then select the related properties of the structure (epitope, name).

Scenario 3: Suppose a clinician wants to find the structures that interact with the COVID-19 spike and their associated chains, genes, alleles and genomics reference sequences. For that, the user can select the structures having an IMGT label SPIKE, their related properties (complex, IEDB epitope, entry) and the complementary structure of the SPIKE protein which has also the same entry and belong to the same complex. Subsequently, the user can select the related properties of entry, like the PDB link. For the associated gene and allele, the user must select the regions associated to the chain of the complementary structure and, associated to these regions, their respective allele. These alleles are used to find the related reference sequence, the entity of the reference sequence and the gene.

6 Related Resources: The OBO Foundry

Considering the complexity of the biological field, there has been a growing effort to provide structured data and models in the field, mainly driven by the OBO (Open Biological and Biomedical Ontologies) community. OBO fosters research in biological and life sciences by making available ontologies and vocabularies. Although not specialised for the immunogenetics field, some of these resources provide general terms to describe: **proteins:** Protein Ontology PRO allows the representation of protein-related entities: from protein families to proteoforms to complexes [7];[28] **genes:** Gene Ontology GO provides resources to enhance the scientific knowledge about the functions of genes from different organisms [2];[29] **sequences:** Sequence Ontology SO [9] provides a controlled and standardised vocabulary for sequence annotation, aiming to unify all sequence annotations. In addition to being more general than IMGT-KG, to our knowledge, none of the resources developed in OBO provide access to integrated immunogenetics data, where IMGT-KG comes to fill exactly this gap.

7 Conclusion

Given the complexity of dealing with adaptive immune response from genome (set of genes) to proteome (set of proteins), there is a need for knowledge sharing and advanced data access in this field to facilitate future and ongoing research. Nowadays, responding to many health and sanitary challenges requires a combination of different studies in different domains, for example the understanding of

[28] https://lod.proconsortium.org/.
[29] http://geneontology.org/.

the COVID-19 virus and the development of a vaccine to counter its spread are both powered by immunogenetics research such as the genetics basis, the main protein implied in the COVID-19 disease.

To face these challenges, we built IMGT-KG, the first FAIR KG in the domain of immunogenetics containing over 79 million triplets. The core model of IMGT-KG is an extended version of the IMGT-ONTOLOGY and the data to populate the KG come from IMGT® databases - established reference data sources for immunogenetics containing both *genomics* and *protein* data. Hence IMGT-KG unifies in a unique manner these two levels of knowledge. This unification gives more query possibilities and opens a way to the discovery of new scientific knowledge. To take an example, among other applications, the IMGT-KG may help to improve knowledge about the coronavirus proteins potentially targeted by the adaptive immune system.

In future work, we will enrich the KG by integrating the IMGT/mAb-DB, the dedicated database to engineered monoclonal antibodies for clinical applications [14], then connect it to related resources like PRO and GO . Subsequently, we will apply representation learning models on the graph in order to predict or discover new links in our data by embedding the KG [15,17]. A named entity recognition system relying on IMGT-KG's entities is currently under construction, aiming to enable the automatic text annotation (e.g. from scientific articles) with IMGT-KG entities.

Resource Availability Statement:

- IMGT-KG web interface: https://www.imgt.org/imgt-kg/
- IMGT-KG fuseki server: https://www.imgt.org/fuseki/#/
- IMGT-KG data model: https://doi.org/10.5281/zenodo.6511279
- Query scenarios: https://doi.org/10.5281/zenodo.6674479
- VoID description: https://www.imgt.org/imgt-kg/kgvoid.html

References

1. Allen, J.F., Hayes, P.J.: Moments and points in an interval-based temporal logic. Comput. Intell. **5**(3), 225–238 (1989). https://doi.org/10.1111/j.1467-8640.1989.tb00329.x
2. Ashburner, M., et al.: Gene ontology: tool for the unification of biology (2000). https://doi.org/10.1038/75556, http://www.flybase.bio.indiana.edu, http://fruitfly.bdgp.berkeley.edu, http://www.genome.stanford.edu, http://www.informatics.jax.org
3. Berners-Lee, T.: Linked Data's rule (2006). https://www.w3.org/DesignIssues/LinkedData.html
4. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 34–43 (2001). https://doi.org/10.1038/scientificamerican0501-34
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semant. Web Inf. Syst. **5**(3), 1–22 (2009). https://doi.org/10.4018/jswis.2009081901
6. Bolleman, J.T., et al.: FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. J. Biomed. Seman. **7**(1), 1–12 (2016). https://doi.org/10.1186/s13326-016-0067-z

7. Chen, C., et al.: Protein ontology on the semantic web for knowledge discovery. Sci. Data **7**(1) (2020). https://doi.org/10.1038/s41597-020-00679-9
8. Ehrenmann, F., Giudicelli, V., Duroux, P., Lefranc, M.P.: IMGT/collier de perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). Cold Spring Harb. Protoc. **6**(6), 726–736 (2011). https://doi.org/10.1101/pdb.prot5635
9. Eilbeck, K., et al.: The sequence ontology: a tool for the unification of genome annotations. Genome Biol. **6**(5) (2005). https://doi.org/10.1186/gb-2005-6-5-r44
10. Giudicelli, V.: IMGT/LIGM-DB, the IMGT(R) comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. **34**(90001), D781–D784 (2006). https://doi.org/10.1093/nar/gkj088
11. Giudicelli, V., Chaume, D., Lefranc, M.P.: IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. **33**(Database Iss.), 256–261 (2005). https://doi.org/10.1093/nar/gki010
12. Giudicelli, V., Lefranc, M.P.: IMGT-Ontology 2012. Front. Genet. **3**(May), 1–16 (2012). https://doi.org/10.3389/fgene.2012.00079
13. Lefranc, M.P., et al.: IMGT R, the international ImMunoGeneTics information system R 25 years on. Nucleic Acids Res. **43**(D1), D413–D422 (2015). https://doi.org/10.1093/nar/gku1056. http://www.imgt.org
14. Manso, T., et al.: IMGT® databases, related tools and web resources through three main axes of research and development. Nucleic Acids Res. **50**(D1), D1262–D1272 (2022). https://doi.org/10.1093/nar/gkab1136
15. Nguyen, D.Q.: A survey of embedding models of entities and relationships for knowledge graph completion. In: Graph-Based Natural Language Processing (TextGraphs 2020), pp. 1–14 (2021). https://doi.org/10.18653/v1/2020.textgraphs-1.1
16. Pojero, F., et al.: The role of immunogenetics in covid-19 (2021). https://doi.org/10.3390/ijms22052636
17. Rossi, A., Barbosa, D., Firmani, D., Matinata, A., Merialdo, P.: Knowledge graph embedding for link prediction: a comparative analysis. ACM Trans. Knowl. Discov. Data **15**(2) (2021). https://doi.org/10.1145/3424672, http://arxiv.org/abs/2002.00819
18. Smith, B., et al.: Relations in biomedical ontologies. Genome Biol. **6**(5) (2005). https://doi.org/10.1186/gb-2005-6-5-r46
19. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. BMC Res. Notes **3** 175 (2010). https://doi.org/10.1186/1756-0500-3-175, http://www.biomedcentral.com/1756-0500/3/175