From a Conceptual Model to a Knowledge Graph for Genomic Datasets

Anna Bernasconi(✉), Arif Canakoglu, and Stefano Ceri

Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Milano, Italy
{anna.bernasconi,arif.canakoglu,stefano.ceri}@polimi.it

Abstract. Data access at genomic repositories is problematic, as data is described by heterogeneous and hardly comparable metadata. We previously introduced a unified conceptual schema, collected metadata in a single repository and provided classical search methods upon them. We here propose a new paradigm to support semantic search of integrated genomic metadata, based on the Genomic Knowledge Graph, a semantic graph of genomic terms and concepts, which combines the original information provided by each source with curated terminological content from specialized ontologies.

Commercial knowledge-assisted search is designed for transparently supporting keyword-based search without explaining inferences; in biology, inference understanding is instead critical. For this reason, we propose a graph-based visual search for data exploration; some expert users can navigate the semantic graph along the conceptual schema, enriched with simple forms of homonyms and term hierarchies, thus understanding the semantic reasoning behind query results.

Keywords: Knowledge Graph · Semantic Search · Conceptual model · Data integration · Genomics · Next Generation Sequencing · Open data

1 Introduction

Next-Generation Sequencing (NGS) technologies and data processing pipelines are supplying high-quality sequencing data at unprecedented pace [16]. Many international consortia provide open access to an increasing number of valuable datasets [6,14,8]. Use of integrated data produced at the various sources is fueling modern biological and clinical research. While the provided sequencing data is generally of high quality, their metadata are not properly standardized and normalized, some of them have missing values, and they are organized differently, with no interoperability support across data sources. To alleviate these problems, we developed the Genomic Conceptual Model (GCM, [1]), covering 8 entities and 37 attributes which describe the most important and complex data sources, including The Cancer Genome Atlas and Genomic Data Commons [6], the Encyclopedia of DNA Elements [14], Roadmap Epigenomics [8], and others. We currently import 40 million metadata key-value pairs from 8 sources, which describe about 240k genomic items.

In our ongoing effort to provide the genomics community with useful concepts and tools, our next challenge is to make metadata semantically searchable and explorable. Along with GCM, we implemented a multi-ontology semantic knowledge base of genomic terms and concepts, called Genomic Knowledge Graph (GKG). We selected ten attributes from GCM; their values were semantically enriched by using the respective best ontologies, after a careful domain-specific selection process. For each associated ontological term, we described synonyms and other syntactic or semantic variants. We then provided a hierarchy of hypernyms and hyponyms. The focus of this paper is not on the GKG construction, discussed elsewhere [2], but rather in its use for supporting a domain-specific semantic search.

Semantic search technology, which is fueling the main search engines developed by Google, Microsoft, Facebook and Amazon, is empowered by the use of large knowledge graphs, supporting search at the semantic level. In these systems, when the query string can be reliably associated to a given entity, other similar instances associated with that entity are also retrieved and displayed together with the entity properties. Inspired by the successful exploitation of knowledge graph in search engines, we envisioned a semantic search approach empowered by our Genomic Knowledge Graph. However, our approach to semantic search differs from the paradigm used by the main search engines; our semantic search is focused only on *domain specific* outputs, and takes into account the fact that users must check semantic inferences, as they are typically ill-defined and error-prone due to the use of external ontologies. Since some expert users may be willing to spend additional effort on search, we expose to them the structure of the knowledge graph, by offering *exploration capabilities* for accessing entities, relationships and hierarchies, e.g., by navigating from given experiments to the cell lines or tissues of provenance, to the donors with their demography and phenotypes, and to the extraction process with the used technology and device.

This paper is organized as follows. In Section 2, we briefly recall the data preparation pipeline to generate the Conceptual Model (GCM) and Knowledge Graph (GKG). Section 3 shows how advanced users can query the knowledge graph according to significant patterns of interaction; we briefly discuss the Neo4j data format to allow exploration queries on GKG. Sections 4 and 5 present related work and conclusions.

2 Building the Genomic Knowledge Graph

The construction of the Genomic Knowledge Graph is performed at the end of a process of data preparation which downloads, transforms, and cleans metadata from original sources, then integrates them in the GCM, performing normalization and enrichment on a number of selected attributes. Such process uses an ETL procedure, which stores data within relational tables; the enrichment process is assisted by tools that minimize the integration designers' efforts.

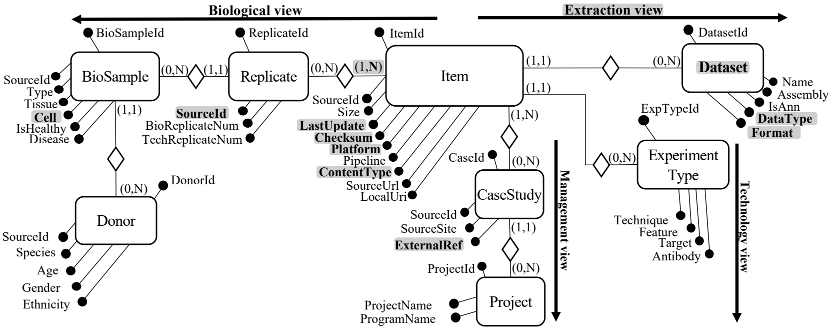


Fig. 1. The Genomic Conceptual Model.

Original Metadata. Metadata are directly downloaded from the original sources and transformed into key-value pairs. In some cases, information is already exposed in this semi-structured format; in other cases, pairs are obtained after flattening hierarchical structures such as JSON or XML.

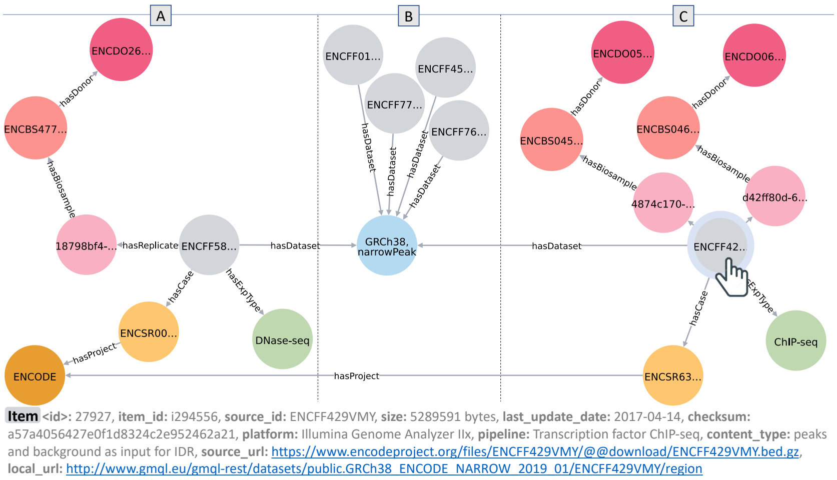
Genomic Conceptual Model. GCM is an entity-relation schema whose main objective is to recognize a common organization for a limited set of supported by most data sources, although with very different names and formats [1]. In Fig. 1 we show GCM in its current state; additions of new attributes, highlighted with grey background and bold font, are due to the practical experience we gained in the field. The schema is organized as a four-pointed star, centered on the Item entity, which represents an elementary experimental unit: a single file of genomic regions and their attributes. Dimensions (or *views*) respectively describe: (1) the biological phenomena observed in the experiment: the sequenced replicated sample, the biological material and its preparation, its donor; (2) the management aspects of the experiment: the case studies and projects/organizations behind its production; (3) the technological process used for the production of the experimental item; (4) the extraction parameters used for internal selection and organization of items, based on a partitioning strategy acting on different parameter values used in programmatic calls towards the sources.

Ontological Terms. As result of a normalization and enrichment phase, we associate specific values of the GCM with controlled terms. Out of all GCM attributes, we selected ten of them as worthy of enrichment. Then, we selected one or two preferred bio-ontologies for each attribute, and performed an enrichment process. The ontological terms information has been retrieved by using the Ontology Lookup Service [7] "search term" API. We save vocabulary terms with their preferred labels, synonyms (or other semantic variants), iri, descriptions and external references (i.e., identifiers of equivalent terms in alternative ontologies). The details of the annotation process are documented in [2].

Ontological Hierarchy. As a further ontological enrichment, we materialize subsets of the aforementioned ontologies which are relevant to annotate our data (typically these range up to five hierarchical levels). The terms are linked through relationships which represent subsumption (*IS_A*), thus including hypernyms and hyponyms of the stored terms, and containment (*PART_OF*), thus including their holonyms and meronyms.

3 Exploration of the Genomic Knowledge Graph

The Genomic Knowledge Graph connections can be visually explored by users who understand the entities and relationships of GCM, as well as their linking to the vocabulary, and then to navigate the generalization *IS_A* and the containment *PART_OF* relationships. The user exploration may start from GCM entities or from the vocabulary terms. We next explain 4 typical patterns of exploration: finding items of a given dataset, of a given patient, of a given case study and associated with a given term.



Item **<id>:** 27927, **item_id:** i294556, **source_id:** ENCFF429VMY, **size:** 5289591 bytes, **last_update_date:** 2017-04-14, **checksum:** a57a4056427e0f1d8324c2e952462a21, **platform:** Illumina Genome Analyzer IIx, **pipeline:** Transcription factor ChIP-seq, **content_type:** peaks and background as input for IDR, **source_url:** https://www.encodeproject.org/files/ENCFF429VMY/@@download/ENCFF429VMY.bed.gz, **local_url:** http://www.gmql.eu/gmql-rest/datasets/public.GRCh38_ENCODE_NARROW_2019_01/ENCFF429VMY/region

Fig. 2. Sequential interaction, from panel (A)—centered on Item ENCFF58—to panel (B)—centered on GRCh38_narrowPeak Dataset—to panel (C)—centered on Item ENCFF42. Note that the items in (A) and (C) share the same Project, ENCODE.

Finding other items from the same datasets. A typical three-step exploratory interaction from an Item to a different Item of the same Dataset is shown in Fig. 2. Entity instances are represented as circles which include the value of entity identifiers or some relevant properties; directed edges, carrying the relationship names, connect entity instances. At all times, one of the entity instances is the *navigation handler*, and its attributes can be (on request) extensively represented in a box presented below the diagram. The end of the navigation is shown in Fig. 2 (C), where the navigation handler points to entity Item ENCFF42, but the navigation starts from Item ENCFF58 in Fig. 2 (A).

We use Fig. 2 (A) to illustrate the typical organization of a GCM instance, centred of the Item ENCFF58 (gray color, in the center), connected to the other entities Replicate, BioSample, Donor (colors from pink to dark red, along the biological view), to CaseStudy and Project (yellow colors, along the management view) and to ExperimentType (green color, along the technology view). In

Fig. 2 (B) we show that the user navigates to the Dataset entity (blue color, along the extraction view), where several other Item instances of the same Dataset are illustrated; then, Fig. 2 (C) shows the end of the navigation. Navigation progressively occurs by double-clicking on entity instances, while attributes of a given entity instance (in this case, of Item) are displayed by single-clicking.

Finding all the datasets of a given patient. Another typical search query asks for all data types pertaining to a specific cancer patient; associating the same patient with heterogeneous data types is highly valuable in order to understand the possible research questions that can be asked to the underlying data repository. However, this query must be explored patient by patient, as each patient may be associated to a highly variable number of data types.

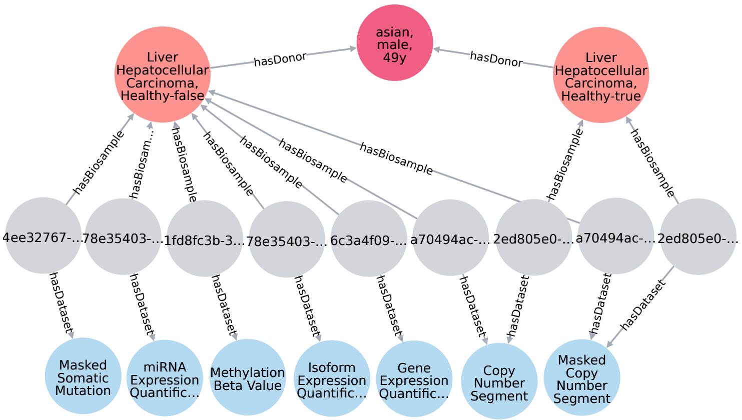


Fig. 3. Exploration starting from a Donor, providing tumoral and normal tissues, which are used to provide Items belonging to different Datasets. Note that here we omit Replicate nodes for space reasons; they have 1:1 correspondence with BioSamples.

As shown in Fig. 3, we represent Donors through their ethnicity, gender, and age (in this specific case through values [asian, male, 49y]). The database stores two biological samples extracted from this patient, who is affected by “Liver Hepatocellular Carcinoma”. One sample is tumoral and the other one is healthy (i.e., a control). By further expanding the nodes, the user reaches the Item level, thereby extracting 9 data Items which belong to 7 different Datasets, each showing the type of data described in the region files (e.g., mutations, methylation levels, copy number variations, and RNA or miRNA gene expression).

Exploring the organization of a given case study. Fig. 4 shows another typical exploration. Assume that a user is not aware of what constitutes a case of study in the ENCODE data source and wants to discover it. Thus, she starts with a given CaseStudy entity ENCSR63, shown at the bottom of the figure. This entity represents a set of Items that are gathered together, because they contribute to the same research objective. The interaction first allows to visualize the group of eight Items associated with this case study, belonging to the hg19_narrowPeak and GRCh38_narrowPeak Datasets (respectively having cardinality five and three). Then, the underlying biological views are revealed, by showing that all the Items are associated with chains originating from two distinct Donors.

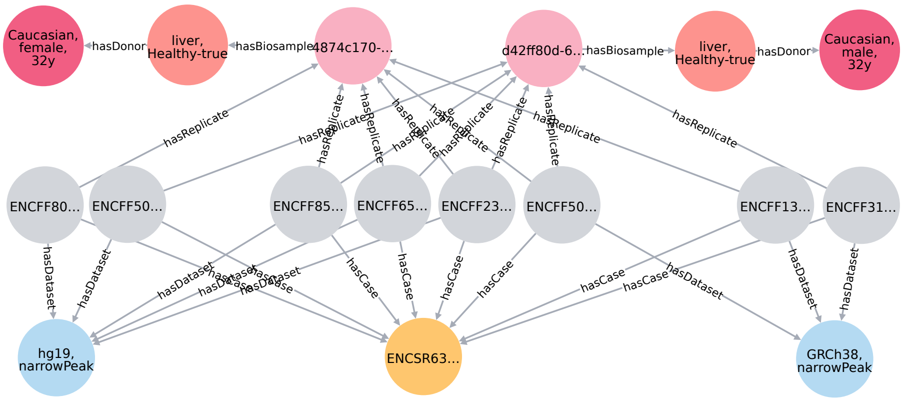


Fig. 4. From bottom to top: a CaseStudy contains multiple Items, which derive from two different Replicates/BioSamples/Donors and are contained in two Datasets based on the reference assembly of the genome.

Ontological exploration. By starting from terms, the user may see how each term is connected to different entities, thereby typically exploring the hierarchical structure of ontological terms. Fig. 5 shows how multiple Items (grey nodes on the right) can be retrieved by using different graph paths starting from the same hierarchical ancestor, ⟨brain⟩. A typical search may start from this entity, which already has a number of connected BioSamples (i.e., samples which have been *annotated* as related to brain concept) and progressively discover all its sub-concepts up to the level where terms annotate other BioSamples. Then, the exploration connects BioSamples to their Replicates and eventually to Items. Note that, in the figure, ⟨brain⟩ directly annotates a BioSample and is an indirect hypernym of ⟨pons⟩ and ⟨globus pallidus⟩, each connected to two BioSamples. Note also that five BioSamples give rise to six Replicates and then to seven Items, and also note that some Items are associated with two Replicates. Once Items are reached, the user may be interested in understanding from which datasets or experiment types they derive; this is possible by further exploring from the Item nodes, using the first pattern of exploration discussed in this Section.

Implementation using Neo4j. For supporting the exploration of GKG, we converted the relational database describing GKG content [2] into a graph database; among many available graph databases (e.g., Neptune or Titan[1]), we have chosen Neo4j (`https://neo4j.com/`), currently the leading open source graph database, used by several companies also in the bioinformatics domain (e.g., EBI, Intermine[2]). We map to Cypher (Neo4j's query language) exploration queries which are progressively built by our query interface.

[1] `https://aws.amazon.com/neptune/`, `http://titan.thinkaurelius.com/`

[2] `https://www.ebi.ac.uk/ols/docs/neo4j-schema`, `https://github.com/intermine/neo4j`

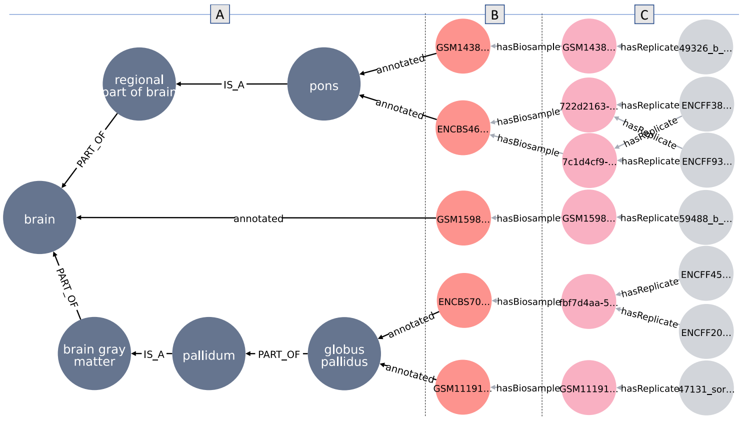


Fig. 5. Search starting from ontological terms. Essentially, (A) contains the ontological terms, (B) contains annotated BioSamples, and (C) the Replicates (pink) and derived Items (grey).

4 Related Works

Some recent works employ conceptual models' expressive power to explain biological entities and their interactions [15,11], or to characterize the processes and objects during related analysis workflows [13]. The GKG is instead based on a CM [1] that drives the data integration process and exposes the unified view resulting from this effort. A classic work [5] proposed a Genomics Ontology, while a more recent one [4] promotes the use of foundational ontologies to avoid errors while creating and curating genomic domain models for personalized medicine. We instead use ontologies to find a common ground between the descriptions and terminologies used in different sources.

Among a number of integrated databases in the bioinformatics domain that employ graph-based paradigms, we cite: BioGraphDB [10], a resource to query, visualize and analyze biological data belonging to several online available sources (focused on genes, proteins, miRNAs, pathways); Bio4j [12], a platform integrating semantically rich biological data (focused on proteins, functional annotations); ncRNA-DB [3], integrating associations among non-coding RNAs and other functional elements.

5 Conclusions

We built an exploration mechanism for supporting semantic queries upon our Genomic Knowledge Graph; we demonstrated the effectiveness of our approach through four examples which are representative of the use of our query interface. Our repository is already storing data coming from eight data sources of genomic data, including datasets relevant for epigenomics, gene expression data, mutation data, deployed in conjunction with an advanced genomic data manager [9], available at `http://gmql.eu/gmql-rest/`).

Acknowledgement. This research is funded by the ERC Advanced Grant 693174 GeCo (Data-Driven Genomic Computing), 2016-2021.

References

1. A. Bernasconi et al. Conceptual modeling for genomics: Building an integrated repository of open data. In *International Conference on Conceptual Modeling*, pages 325–339. Springer, 2017.
2. A. Bernasconi et al. Ontology-driven metadata enrichment for genomic datasets. In *International Conference on Semantic Web Applications and Tools for Life Sciences*, volume 2275. CEUR-WS, 2018.
3. V. Bonnici et al. Comprehensive reconstruction and visualization of non-coding regulatory networks in human. *Frontiers in bioengineering and biotechnology*, 2:69, 2014.
4. A. M. M. Ferrandis et al. Applying the principles of an ontology-based approach to a conceptual schema of human genome. In *International Conference on Conceptual Modeling*, pages 471–478. Springer, 2013.
5. J. Hammer and M. Schneider. The GenAlg project: developing a new integrating data model, language, and tool for managing and querying genomic information. *ACM SIGMOD Record*, 33(2):45–50, 2004.
6. M. A. Jensen et al. The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, 130(4):453–459, 2017.
7. S. Jupp et al. A new Ontology Lookup Service at EMBL-EBI. In J. Malone et al., editors, *International Conference on Semantic Web Applications and Tools for Life Sciences*, volume 1546, pages 118–119. CEUR-WS, 2015.
8. A. Kundaje et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.
9. M. Masseroli et al. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. *Bioinformatics*, page bty688, 2018.
10. A. Messina et al. BioGraph: a web application and a graph database for querying and analyzing bioinformatics resources. *BMC systems biology*, 12(5):98, 2018.
11. A. L. Palacio et al. A method to identify relevant genome data: Conceptual modeling for the medicine of precision. In *International Conference on Conceptual Modeling*, pages 597–609. Springer, 2018.
12. P. Pareja-Tobes et al. Bio4j: a high-performance cloud-enabled graph-based data platform. *BioRxiv*, page 016758, 2015.
13. G. Rambold et al. Meta-omics data and collection objects (mod-co): a conceptual schema and data model for processing sample data in meta-omics research. *Database*, 2019, 2019.
14. Consortium ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.
15. J. F. R. Román et al. Applying conceptual modeling to better understand the human genome. In *International Conference on Conceptual Modeling*, pages 404–412. Springer, 2016.
16. Z. D. Stephens et al. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.