Med2RDF: Semantic Biomedical Knowledge-base and APIs for the Clinical Genome Medicine

Mayumi Kamada[1], Toshiaki Katayama[2], Shuichi Kawashima[2], Ryosuke Kojima[1], Masahiko Nakatsui[1], Yasushi Okuno[1]

[1] Kyoto University, 54 Shogoin, Sakyo-ku, 606-8397, Kyoto, Japan
[2] Database Center for Life Science, 178-4-4 Wakashiba, Kashiwa-shi, 277-0871 Chiba, Japan
mkamada@kuhp.kyoto-u.ac.jp

Abstract. For clinical interpretation of genomic variants, it is necessary to aggregate knowledge from public databases and literatures. To construct an integrated knowledge-base for interpretation, we have developed RDF versions of major biomedical databases in the Med2RDF project. This resource uses the originally developed med2rdf ontology covering core concepts ranging from genomes, genes, transcripts, variations, diseases, to evidence, common to the supported databases. We currently provide converters for 19 public databases that are required to interpret disease relevance. We stored most of the resulting RDF data in our SPARQL endpoint and are currently developing APIs to utilize the RDF data for accelerating application development for genomic medicine.

Keywords: Med2RDF, Database integration, APIs, clinical genome medicine.

1 Introduction

Genomic medicine aims to provide an appropriate medical treatment policy based on individual genetic background. However, many of genomic variants identified by genome sequence analysis are unclear in relation to mechanism of disease and often do not lead to clinical determination. These variants are called variants of uncertain significance (VUS) and the interpretation of these variants is a bottleneck of genomic medicine. To clarify the disease relevance of VUS, in addition to specialized knowledge in each disease domain, comprehensive interpretation of enormous amounts of information in the literature and public databases is needed. Thus, in Med2RDF project, we have tackled to integrate knowledge required to the clinical interpretation utilizing Resource Description Framework (RDF).

To date, major life science databases have been developed and provided as RDF data thanks to the community efforts [1]. Our Med2RDF is an addition of biomedical databases to this collaboration. We provide converters for MedGen, HGNC, ClinVar, dbSNP, dbVar ExAC, gnomAD, dbNSFP, dbscSNV, HiNT, INstruct, ICGC, TCGA,

CIViC, COSMIC, CCLE, GDSC, OpenTG-Gates and DGIdb at the Med2RDF GitHub repository[1] and have stored the resulting RDF datasets at our SPARQL endpoint[2].

2 Med2RDF ontology and API development

Along with the development of RDF data, we have developed the med2rdf ontology covering core concepts ranging from genomes, genes, transcripts, variations, diseases, to evidence, common to the supported databases (**Fig 1**).

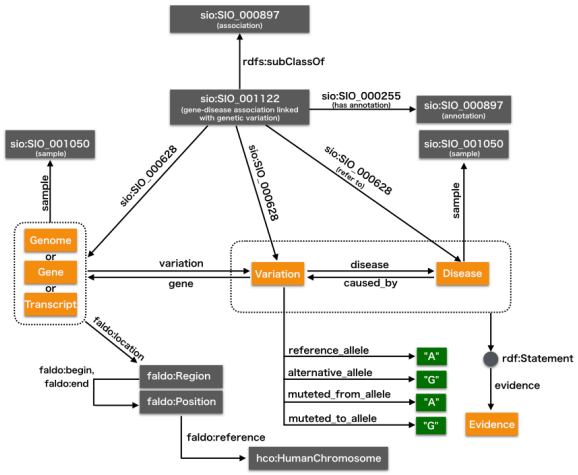

Fig 1. A schematic representation of the Med2RDF ontology commonly used in the Med2RDF datasets to improve interoperability.

This ontology enables us to integrate heterogeneous datasets by improving the interoperability, and one can utilize any combination of data in a standardized manner. Moreover, we are currently developing APIs that encapsulate the SPARQL query with the help of the SPARQList[3] with which users can develop applications for the clinical genome medicine with ease. This will also help researchers to apply machine learning methods to Med2RDF data for the clinical interpretation of VUS obtained from clinical sequencing.

Acknowledgements. This research is supported by the Program for an Integrated Database of Clinical and Genomic Information from Japan Agency for Medical Research and development, AMED.

Reference. 1. Katayama T, Kawashima S, Micklem G *et al.*: BioHackathon series in 2013 and 2014, *F1000Research* 2019, 8:1677[4]

[1] https://github.com/med2rdf
[2] http://sparql.med2rdf.org/
[3] https://github.com/dbcls/sparqlist
[4] https://doi.org/10.12688/f1000research.18238.1