Construction of Knowledge-base for Clinical Interpretation of Genomic Variants

Mayumi Kamada[1], Toshiaki Katayama[2], Shuichi Kawashima[2], Fumie Ono[1], 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.
Clinical interpretation for variants of uncertain significance is important to provide appropriate medical treatment. However, enormous effort and specialized knowledge are required to give a clinical interpretation to variants. To reduce the burden, it is necessary to develop an automated estimation system of clinical significance using aggregated knowledge from public databases and literature for interpretation. We are constructing a database that collects disease-related variants in Japanese population in order to improve interpretation of Japanese variants. In this work, we carry out RDF conversion of public databases that are needed to interpret variants, and integration of them to apply to the estimation system using a machine learning method.

Keywords: Clinical interpretation of variants, Estimation of clinical significance, Integrated knowledge base.

1 Introduction

The improvement of genome sequencing technology enables us to apply clinical sequence using next generation sequencer on clinical diagnosis. The purpose of clinical sequence is to provide an appropriate medical treatment policy, based on individual genetic background. However, many of the detected sequence variants are unclear in relation to mechanism of disease and often do not lead to clinical determination. These variants are called as variants of uncertain significance (VUS) which is one of the problem to obstruct precision medicine. In order to clarify the disease relevance of VUS, it is needed to obtain 1) specialized knowledge in each disease domain, 2) comprehensive interpretation of enormous information in literature, and 3) the clinical background of individual patients. The aggregation of such knowledge leads us to the realization of a system that can automatically estimate the disease relevance.

The Database Center for Life Science (DBCLS) in Japan has been working on integration of public databases using Resource Description Framework (RDF) by promoting international collaborations on standardization of semantics in life sciences and biomedical domains[1]. We have been developing a machine learning method to estimate clinical significance for each variant, for which graph-structured data is used as learning

[1] https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-5-5

data. In order to fertilize the learning data, we carry out RDF conversion of databases that are required to interpret disease relevance.

2 Target variants and concept of integrated knowledge-base

The goal of our project is to construct a database to give appropriate interpretation for Japanese variants. It is well known that disease association is affected by the genomic background difference in a population. For Japanese population, we have been constructing a disease-related genomic information database. The database is going to store variants and clinical data collected from the fields of "cancer", "rare disease", "infectious disease", "dementia", "hearing loss".

The germline variant is a mutation in a reproductive cell (egg or sperm), which induces single-gene disorders in many cases. Interpretation of germline variants is often done with a guideline developed by the American College of Medical Genetics and Genomics (ACMG)[2]. Based on the ACMG guideline for making medical treatment decisions, we have been converting the following databases into RDF and also integrating them with the existing RDF datasets on reference genome and protein annotations.

- ClinVar[3], COSMIC[4] (Pathogenicity of variants)
- dbNSFP[5] (Effects by variants)
- dbSNP[6], dbVar[7] (Genetic variants and their frequency in population)
- DGIdb[8] (Drug-gene interaction)
- HINT[9], INstruct[10] (Molecular interaction)

If different terms meaning the same object are used in individual databases, we cannot use them in an integrated manner. Thus, we promote the unification of terms by ontology development, and the standardization of URI by using the same prefix (http://identifiers.org). In fact, for each database, the converter from database dump files (csv, tsv etc.) and RDB to RDF will be released as the Docker[11] containers. We also plan to provide our database of Japanese disease-related genomic data as RDF so that it can be seamlessly integrated with the constructed knowledge graph in the above.

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).

---

2 Richards S. et al., Genet. Med., 17(5):405-24, 2015.
3 https://www.ncbi.nlm.nih.gov/clinvar/
4 http://cancer.sanger.ac.uk/cosmic
5 https://sites.google.com/site/jpopgen/dbNSFP
6 https://www.ncbi.nlm.nih.gov/projects/SNP/
7 https://www.ncbi.nlm.nih.gov/dbvar
8 http://dgidb.org/
9 http://hint.yulab.org/
10 http://instruct.yulab.org/
11 https://www.docker.com/