Development of data infrastructure and in silico prediction method to promote genomic medicine

Mayumi Kamada[1]

[1]Graduate School of Medicine, Kyoto University

Abstract. In genome medicine, which is now being implemented in medical care, variants detected by genome analysis such as next-generation sequencers are clinically interpreted to determine the diagnosis and treatment plan. The clinical interpretation is performed based on the detailed clinical background and the information from journal papers and public databases, such as frequencies in the population and their relationship to the disease. A large amount of genomic data has been accumulated so far, and many genomic variant databases related to diseases have been developed, including ClinVar. On the other hand, the genes and variants involved in diseases are different between populations with different genetic backgrounds. Furthermore, it has been reported that there is a racial bias in the information shared in current public databases, which affects clinical interpretation. Therefore, increasing the diversity of genomic variant data has become an important issue worldwide. In Japan, the Japan Agency for Medical Research and Development (AMED) launched a project to develop an integrated clinical genome information database in 2016. This project targeted "Cancer," "Rare/Intractable diseases," "Infectious diseases," "Dementia," and "Hearing loss", and in collaboration with research institutes that provide genomic medicine in Japan, we developed an integrated database named MGeND (Medical Genomics Japan Database). The MGeND is a freely accessible database, which provides disease-related genomic information detected from the Japanese population. The MGeND widely collects variant data for monogenic diseases represented by rare diseases and polygenic diseases such as dementia and infectious disease. The genome variant data are integrated by genomic position for these diseases and can be searched across diseases. The useful genome analysis methods differ depending on the disease area. Therefore, in addition to "SNV, short indel, SV, and CNV" data handled by ClinVar, MGeND includes GWAS (Genome-Wide Association Study) data, which is widely used in studies of polygenic diseases, and HLA (Human Leukemia Virus) allele frequency data, which is used in immune-related diseases such as infectious diseases. As of September 2021, more than 150,000 variants have been registered in MGeND, and 60,000 unique variants have been made public. Of these variants, about 70% were variants registered only in MGeND and not registered in ClinVar. This fact shows the importance of the efforts to collect genomic information by each ethnic group. On the other hands, many variants have not been annotated with any clinical interpretation because the effects on molecular function and the mechanisms of disease are not clear at this time. These variants of uncertain significance (VUS) are a bottleneck for genomic medicine because they cannot be used for diagnosis or treatment selection. The evaluation of VUS requires detailed experimental validation and a vast amount of knowledge integration, which is costly. In order to understand the molecular function and disease relevance of VUS and to enable optimal drug selection, we have been developing a machine learning-based method for predicting the pathogenicity of variants and a computational platform for estimating the effect of variants on drug sensitivity. Many methods for predicting the pathogenicity of genomic variants using machine learning have been developed. Most of them use the conservation of amino acid or nucleotide sequences among closely related species, physicochemical properties of proteins as features for prediction. There are also many prediction methods based on ensemble learning that aggregate the predicted scores by existing tools. These approaches focus on individual genes and variants and evaluate their effects. However, in many diseases, multiple molecules play a complex role in the pathogenesis of the disease. In other words, to assess the pathological significance of variants more accurately, it is necessary to consider the molecular association. Therefore, we constructed a knowledge graph based on molecular networks, genomic variants, and predicted scores by existing methods and proposed a prediction model using Graph Convolutional Network (GCN). The prediction performance evaluation using a benchmark set showed that the GCN-based method outperformed existing methods. It is known that variants can affect the interaction between a molecule and a drug. For optimal drug selection, it is necessary to clarify the effect of the variant on drug affinity. It is time-consuming and costly to perform experiments on a large number of VUSs. Our previous studies show that molecular dynamics calculation can evaluate the affinity between mutants and drugs energetically and estimate with high accuracy. We are currently working on a project to estimate the effects of a large number of VUSs using the supercomputer Fugaku. To realize calculations for many VUS in this project, we are developing a data platform for seamlessly performing molecular dynamics simulation from genome information. Moreover, we are constructing a database to publish calculation results and their outcomes for contributing a selection of optimal drugs. In the presentation, I will introduce the development of the databases and prediction methods to improve the efficiency of genomic medicine.