第２回 ASHBi 数理生物研究集会 プログラム

日 時: 2020 年 9 月 4 日（金） 9:30-17:00
場 所: Zoom オンラインミーティングにて開催

9:30-10:30 **川上 英良**（理化学研究所）

“Data-driven medical research using machine learning”

10:45-11:45 **小井土 大**（東京大学）

“Learn genomics from AI for interpreting the roles of non-coding RNAs in complex traits”

(休憩 11:45–13:30)

13:30-14:30 **玉田 嘉紀**（京都大学）

“Personalized gene network analysis with large scale Bayesian network estimation”

14:45-15:45 **中戸 隆一郎**（東京大学）

“Large-scale analysis of multi-omics data toward the elucidation of epigenomic diversity”

16:00-17:00 フリーディスカッション

主 催：
京都大学高等研究院 ヒト生物学高等研究拠点（WPI-ASHBi）
田﨑 創平、井元 佑介、平岡 裕章、岩見 真吾、斎藤 通紀

川上 英良（理化学研究所）

タイトル:
Data-driven medical research using machine learning

要旨:
In recent years, artificial intelligence technology and data science have rapidly spread in medical research. In medical research, it is often difficult to directly apply a basic biological model due to differences in time scale and hierarchy. Instead, a data-driven approach has been employed, which observe target disease without specific hypothesis. Machine learning, which can extract potential patterns and make highly accurate predictions while considering various types of variables and complex dependencies between variables, is a major tool in the data-driven medical research. We conduct research using machine learning for the purpose of stratifying and visualizing health and disease states and predicting the onset and progression of disease. In this seminar, I would like to introduce an explainable scheme for stratifying and visualizing health and disease states using machine learning and discuss the prospects for preventive and personalized medicine.

小井土 大（東京大学）

タイトル:

Learn genomics from AI for interpreting the roles of non-coding RNAs in complex traits

要旨:

Genome sequence data can be handled in a similar way as image data by considering the four bases (ATGC) as the three primary colors (RGB). From the combination of local features in the image (e.g., eyes, ears, etc.), we recognize an object (e.g., dog or cat), whose mechanisms can be modeled by the deep convolutional neural networks (deep CNN). Similarly, we may know the meaning of long genome sequence patterns from the combination of motif sequences. In fact, the deep CNN has made it possible to predict tissue-specific gene expression levels from the surrounding kb-order genome sequences alone. I extended this concept into the transcription of non-coding RNAs (ncRNAs) and developed mutation effect prediction on ncRNA transcription (MENTR), a machine learning framework reliably connecting genetic associations with expression of ncRNAs (Koido et al., bioRxiv 2020). MENTR-predicted mutation effects on ncRNA transcription were concordant with estimates from previous genetic studies in a cell type-dependent manner. MENTR proposed 7,775 enhancers and 3,548 long-ncRNAs as complex trait-associated ncRNAs in 348 major human primary cells and tissues, including plausible enhancer-mediated functional alterations. In this workshop, I want to discuss how we can use machine learning methods for interpreting the results of statistical genetic studies.

玉田 嘉紀（京都大学）

タイトル:
Personalized gene network analysis with large scale Bayesian network estimation

要旨:
A Bayesian network is a graphical model that can be used to estimate cause-and-effect relationships among variables by considering conditional independencies between them. It is an explainable artificial intelligence model since the relationships among variables are represented explicitly by mathematical equations that are learned automatically from data. We are using this model for a long time to analyze gene-to-gene regulatory relationships with gene expression data. It enables us to model and analyze the entire transcripts including over 20,000 thousand genes in human cells. The difficulty in Bayesian network estimation is the structure search of the network because it is known to be an NP-hard problem, i.e., the search space of possible structures increases exponentially as the number of variables increases. Also, the network estimation results in more than 100,000 edges. Thus, it is hard to extract meaningful results from such a huge network and to interpret them. In this talk, we begin with a brief history of gene network estimation, and then introduce our history of straggling against this difficult problem. The current focus of our research group is on the estimation of personalized individual networks from limited samples. We present our latest progress that overcomes this problem.

中戸 隆一郎（東京大学）

タイトル:
Large-scale analysis of multi-omics data toward the elucidation of epigenomic diversity

要旨:
Large-scale epigenomic analysis that compares hundreds of samples is an essential approach to reveal the high-dimensional interrelationship for regulatory elements and also annotate novel functional genomic regions de novo. As a part of the International Human Epigenome Consortium (IHEC) project, we collected histone modification and gene expression data to profile and characterize the epigenomic landscape of human endothelial cells (ECs) and implemented comprehensive analysis for genome-wide data. To avoid the batch effect and individual differences among samples, we developed a computational pipeline for epigenome analysis combined with chromatin interaction data. We found that most of the deferentially expressed genes and enhancer sites were cooperatively enriched in more than one EC type, suggesting that the distinct combinations of multiple genes play key roles in the diverse phenotypes across EC types. Notably, many homeobox genes were differentially expressed across EC types, and their expression was correlated with the relative position of each organ in the body. Lastly, I discuss several future prospects including the single-cell analysis.