Oral Presentation 12

日時： 2020 年 10 月 30 日 14:45-16:15
チャンネル： 3

口頭発表12
『バイオインフォマティックス』
Bioinformatics

座長：　清水 祐吾（慶應義塾大学）
Yugo Shimizu (Keio University)

O3-12
"Development of machine learning based prediction model for hypertensive disorders of pregnancy (HDP)"
Satoshi Mizuno
(Tohoku Medical Megabank Organization, Tohoku University)

O3-13
"Selecting the genes related to COVID-19 with PCA-based unsupervised feature extraction"
Kota Fujisawa
(Graduate school of Engineering and Science, University of the Ryukyus)

O3-12

Development of machine learning based prediction model for hypertensive disorders of pregnancy (HDP)

Satoshi Mizuno[1] samizuno@med.tohoku.ac.jp
Satoshi Nagaie[1] satoshi.nagaie@gmail.com
Gen Tamiya[2] gtamiya@genetix-h.com
Shinichi Kuriyama[3] kuriyama@med.tohoku.ac.jp
Hiroshi Tanaka[4] tanaka@cim.tmd.ac.jp
Nobuo Yaegashi[5] nobuo.yaegashi@gmail.com
Junichi Sugawara[6] jsugawara@med.tohoku.ac.jp
Soichi Ogishima[1] ogishima@megabank.tohoku.ac.jp

[1] Department of Informatics for Genomic Medicine, Group of Integrated Database Systems, Tohoku Medical Megabank Organization, Tohoku University, Miyagi, Japan
[2] Department of Statistical Genetics and Genomics, Group of Disease Risk Prediction, Tohoku Medical Megabank Organization, Tohoku University, Miyagi, Japan
[3] Department of Molecular Epidemiology, Group of the Birth and Three-Generation Cohort Study, Tohoku Medical Megabank Organization, Tohoku University, Miyagi, Japan
[4] Department of Bioclinical Informatics, Group of Integrated Database Systems, Tohoku Medical Megabank Organization, Tohoku University, Miyagi, Japan
[5] Department of Gynecology and Obstetrics, Tohoku University Graduate School of Medicine, Tohoku University, Miyagi, Japan
[6] Department of Feto-Maternal Medical Science, Group of Community Medical Supports, Tohoku Medical Megabank Organization, Tohoku University, Miyagi, Japan

Keywords: Machine learning, hypertensive disorders of pregnancy, Phenotyping

In the resent years, machine learning (ML) is widely applied to clinical tasks including detection of tumors from medical images [1] and predict clinical events from electronic health records（EHRs）[2]. ML is also expected to be applied to early prediction of common diseases such as diabetes and heart failure from various data including genetic factors and exposures. As major limitations underlying current effort to early prediction of common diseases from big data, static and informatic issues including multi-modality, high dimension and variety of data remain to be addressed.

In this study, we developed prediction model of hypertensive disorders of pregnancy (HDP) with 22,256 pregnancy women in the BirThree cohort study [3]. Time-series exposures, laboratory tests and medical records were used as input data. To address static and informatic issue, we compared predictive power of several types of ML models and preprocessing methods in combination. Both training and test labels were identified by previously developed precise phenotyping algorism (PPV = 0.94). Evaluation of predictive powers were performed based on ten-fold cross validation.

The predictive power of developed ML model was up-to 0.95 of F1 score. Among the developed models, interpretable models show high importance for blood pressure around the mean of onset, eating habit and lifestyle.

Our developed ML models enable us not only to conduct risk prediction but also knowledge acquisition for drug development of HDP.

[1] S Kudo *et al, Artificial Intelligence-assisted System Improves Endoscopic Identification of Colorectal Neoplasms*, Clin Gastroenterol Hepatol, 18(8):1874-1881.e2, 2020
[2] A Rajkomar *et al, Scalable and accurate deep learning with electronic health Records*, NPJ Digit Med, 8;1:18, 2018.
[3] S Kuriyama *et al*, *The Tohoku Medical Megabank Project: Design and Misson*, J Epidemiol, 26(9):493-511, 2016.

O3-13

Selecting the genes related to COVID-19 with PCA-based unsupervised feature extraction

Kota Fujisawa[1] k198422@eve.u-ryukyu.ac.jp
Mamoru Shimo[2] e175238@eve.u-ryukyu.ac.jp
Yoshihiro Taguchi[3] tag@granular.com
Shinya Ikematu[4] ikematsu@okinawa-ct.ac.jp
Ryota Miyata[2] miyata26@tec.u-ryukyu.ac.jp

[1] Graduate school of engineering and science, University of the Ryukyus, 1 Senbaru, Nishihara, Nakagami, Okinawa 903-0213, Japan
[2] Faculty of Engineering, University of the Ryukyus, 1 Senbaru, Nishihara, Nakagami, Okinawa 903-0213, Japan
[3] Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo, Tokyo 112-8551,Japan
[4] Department of Bioresources Engineering, National Institute of Technology, Okinawa College, 905 Henoko, Nago, Okinawa 905-2192, Japan

Keywords: COVID-19, SARS-CoV-2, machine learning, gene selection

Coronavirus disease 2019 (COVID-19) is raging all over the world. This potentially fatal infectious disease is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). However, the mechanism of COVID-19 is not well understood. Therefore, we analyze the gene expression profiles of COVID-19-infected patients to identify the disease-related genes using an innovative machine learning method, which allows us to perform gene selection from a dataset with small samples and many candidates based on a data-driven strategy.

First, we applied Principal-components-analysis-based unsupervised feature extraction (PCAUFE [1]) to the mRNA expression profiles of 17 patients and 17 healthy controls (GSE152418 [4]), identifying 123 genes as critical for COVID-19 progression from 60,683 candidate genes. Second, we also applied PCAUFE to GSE1739 [2], a dataset of SARS, which was the other bat-origin coronavirus disease and was caused by SARS-CoV. An integrated analysis of the two datasets revealed 83 genes uniquely selected from the COVID-19 dataset, such as B2M, EIF4G2, and HLA-DPA1. Moreover, we also found 40 genes commonly selected from both the datasets such as CD74, HLA-DRA, and HLA-DRB1.

Finally, to investigate the biological reliability of these selected genes, we uploaded them an enrichment analysis sever called GeneSetDB [3]. Both the 83 genes unique to the COVID-19 dataset and the 40 genes common to these zoonotic-coronavirus datasets mainly included immune-related genes. These results suggest that PCAUFE could successfully identify a biologically feasible set of COVID-19-related genes.

[1] Taguchi, Y., *Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach*, Springer, 2019.
[2] Reghunathan, R., *et al.*, Expression profile of immune response genes in patients with Severe Acute Respiratory Syndrome, *BMC Immunol*, 6:2, 2005.
[3] https://genesetdb.auckland.ac.nz/haeremai.html
[4] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152418