frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 17 March 2023
DOI 10.3389/fgene.2023.1177996

OPEN ACCESS

EDITED BY
William C. Cho,
QEH, Hong Kong SAR, China

REVIEWED BY
Giulia Tini,
European Institute of Oncology (IEO), Italy

*CORRESPONDENCE
Yun Zheng,
zhengyun5488@gmail.com
Zexuan Zhu,
zhuzx@szu.edu.cn

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 02 March 2023
ACCEPTED 13 March 2023
PUBLISHED 17 March 2023

CITATION
Zheng Y and Zhu Z (2023), Editorial: Retrieving meaningful patterns from big biomedical data with machine learning approaches.
Front. Genet. 14:1177996.
doi: 10.3389/fgene.2023.1177996

Editorial: Retrieving meaningful patterns from big biomedical data with machine learning approaches

Yun Zheng[1]* and Zexuan Zhu[2]*

[1]College of Landscape and Horticulture, Yunnan Agricultural University, Kunming, Yunnan, China, [2]National Engineering Laboratory for Big Data System Computing Technology, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong, China

KEYWORDS

big biomedical data, machine learning, classification, regression, prognosis, software tool

Editorial on the Research Topic
Retrieving meaningful patterns from big biomedical data with machine learning approaches

In recent years, a huge amount of biomedical data has been generated by high throughput sequencing facilities. These data include, but not limited to, genomic sequences, transcriptome profiles, non-coding RNA profiles, epigenetics profiles, and single cell RNA-Seq profiles. These data sets are normally large and noisy. Therefore, careful analysis with machine learning methods is often needed to extract meaningful patterns from these data. Machine learning problems were normally grouped into classification and regression problems, whose variables under consideration were categorical (qualitative) and numerical (quantitative), respectively (James et al., 2013).

This Research Topic of Frontiers in Genetics features a Research Topic of three articles for identifying meaningful molecular patterns from clinical samples and an article that introduces a web-based tool for visualization of multi-omics microbial data sets.

Early detection of cancer is critical for better outcome and lower mortality (Siegel et al., 2020). The task is very challenging since clinical tumor samples of some cancers are unaccessible. It is therefore very valuable if early detection of cancer could be reliably conducted with accessible clinical samples, such as blood samples. However, the sensitivities of existing methods (Cohen et al., 2018; Liu et al., 2020) are not satisfactory. Qi et al. collected 75 whole-blood transcriptome profiles from 45 patients with various non-blood cancers (breast, esophagus, stomach, thyroid, rectum, colon, and uterus) and 30 normal controls. They first identified 900 differentially expressed genes (DEGs) from a training set with 53 samples (Qi et al.). Then, the support vector machine (SVM) algorithm was used to build models based on these 900 DEGs, 120 very long intergenic non-coding RNAs (vlincRNAs), and 780 non-vlincRNA genes, respectively. Qi et al. showed that these SVM-based models were accurate for pan-cancer detection on the independent testing data with 22 samples. They also demonstrated that vlincRNAs had superior performance when compared to protein-coding mRNAs (Qi et al.).

Zhong et al. constructed a hypoxia-related prognosis model, namely, HPM based on gene signatures and machine leaning methods to predict the survival of acute myeloid leukemia patients. The proposed HPM and the derivative models with clinical risk factors were validated with various experimental studies including Kaplan-Meier survival analysis, time-dependent ROC analysis, clinical characteristics analysis, and hypoxia-related immune and metabolic alterations (Zhong et al.). The HPM and its derivative models were able to effectively predict the survival of AML patients, which might improve risk classification (Zhong et al.).

Based on cuproptosis-related genes reported in (Tsvetkov et al., 2022), Cai et al. identified 956 cuproptosis-related lncRNAs (CRLs). Then, univariate Cox regression was utilized to identify 69 CRLs in the training group and multivariate Cox regression was used to identify 3 CRLs as independent prognostic factors (Cai et al.). And the regression model based on these three CRLs could accurately predict the prognosis of bladder cancer (BC) patients (Cai et al.). They also showed that high-risk BC patients benefited more from immunotherapy and had stronger immune responses, and their overall survival was better (Cai et al.).

Li et al. developed MicrobioSee, a web-based toolkit for visualizing multi-omics data sets of microorganisms. MicrobioSee provided non-expert users friendly interfaces for 17 different analysis tasks of major omics data of microorganisms (Li et al.).

In summary, the work in this Research Topic provided some practical examples of employing machine learning methods in multi-omics profiles of diseases. Interesting patterns were identified in these studies in the contexts of classification and regression problems.

Author contributions

YZ and ZZ: Conceptualisation, formal analysis, funding acquisition, writing—review and editing.

Funding

The research was supported in part by a grant (No. 31460295) of the National Natural Science Foundation of China and an Open Research Fund (No. SKLGE-2107) of State Key Laboratory of Genetic Engineering, Fudan University, China to YZ; and by a grant (No. 61871272) of the National Natural Science Foundation of China to ZZ.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359, 926–930. doi:10.1126/science.aar3247

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. 1 edn. New York: Springer.

Liu, M., Oxnard, G., Klein, E., Swanton, C., Seiden, M., Liu, M. C., et al. (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free dna. *Ann. Oncol.* 31, 745–759. doi:10.1016/j.annonc.2020.02.011

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA A Cancer J. Clin.* 70, 7–30. doi:10.3322/caac.21590

Tsvetkov, P., Coy, S., Petrova, B., Dreishpoon, M., Verma, A., Abdusamad, M., et al. (2022). Copper induces cell death by targeting lipoylated tca cycle proteins. *Science* 375, 1254–1261. doi:10.1126/science.abf0529