Editorial

Introduction to the special issue on Applications of Artificial Intelligence in Biomarker Research

Karin D. Rodland[a,b], Bobbie-Jo Webb-Robertson[a,c] and Sudhir Srivastava[d]
[a]*Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA*
[b]*Cell, Developmental, and Cancer Biology, Oregon Health and Science University, Portland, OR, USA*
[c]*Biomedical Engineering, Oregon Health and Science University, Portland, OR, USA*
[d]*Editor-in-Chief, Cancer Biomarkers*

Recent advances in high throughput analytical technologies and artificial intelligence (AI) are changing the paradigm of biomarker discovery. This ability to analyze large numbers of clinical samples is providing unprecedented amounts of imaging and molecular-based biomarker data ranging from the scale of GBs for genome, transcriptome, metabolome and proteome data to TBs for imaging data [1] beyond the ability of traditional statistical analyses and tools. The availability of big data has resulted in the development of new computational tools, including new machine (ML) and deep learning (DL) algorithms to provide levels of accuracy expected of meaningful clinical interpretations. For example, the exciting results by AlphaFold in the recent Critical Assessment of Protein Structure Prediction (CASP) was driven by a deep neural network trained on over 170,000 protein structures allowing the team to accurately model 24 of the 43 domains with the next method far behind at 14 of the 43 domains [2,3]. However, as noted by the authors of AlphaFold, deep neural networks are complex and understanding how the network reaches the residue-based distance predictions is not clear. This is one of the most challenging areas for AI in application to biomedical problems, building highly accurate models that are both explainable and trustworthy. For biomarker discovery recent advances in integrating multi-omics data with advanced feature selection and ranking are leading to better understanding of mechanisms driving disease [4,5]. Further, integrating constraints into the models based on biological

domain knowledge in a principled manner can improve both the accuracy and interpretation of models being applied to predict health outcomes, stratify patients for treatment, understand the underlying molecular mechanisms of disease, and many other tasks needed to realize the goals of personalized medicine [6].

This special issue of *Cancer Biomarkers* is devoted to the application of artificial intelligence and machine learning approaches to the unique challenges of biomarker discovery. The first article in the series, Mikdadi et al, presents a comprehensive review of the literature regarding the use of AI approaches to identify biomarkers for ovarian and pancreatic cancer. Although the examples chosen focus on ovarian and pancreatic cancer, the underlying principles are universal, and the analysis of gaps and challenges applies to the field as a whole.

Two of the articles in this special issue focus on the development of new computational tools to facilitate the application of AI to biomarker identification. Yoon et al. describes the development of an approach to develop specialized vocabularies for AI and DL studies, without jeopardizing the subjects' personally identifiable information. This issue is of increasing concern, as the resolution of AI approaches to feature selection continues to improve. Tayob et al. is focused on the development of novel Bayesian algorithms to leverage longitudinal measurements of the same features over time in a given patient to facilitate the early detection of cancer, using hepatocellular carcinoma as the model system. This is

an extremely interesting approach which may provide a useful strategy for dealing with human heterogeneity in the baseline levels of potential biomarkers.

The remaining five articles focus on specific applications of AI and DL to the problem of biomarker identification and refinement, often with a minor focus on tool optimization. Li et al. focuses on the use of pre-diagnostic Computed Tomagraphy (CT) scans to stratify patients for risk of pancreatic cancer. This article is interesting in that the sole data input is radiometric images, and the clinical objective is the identification of patients at highest risk of pancreatic cancer for more intensive monitoring. Smith et al. also takes an imaging-focused approach, in this case using whole cell imaging and cyclic multiplexed immunofluorescence to identify immune features in the tumor microenvironment of pancreatic cancer that may provide prognostic information, particularly in regard to the response to immunotherapy. Gazouli et al. used microRNAs as the exclusive data input, and applied machine learning to identify a miRNA profile associated with gastrointestinal stromal tumors. Song et al. combined radiomics data from pre-operative PET and CT images in patients with early stage uterine cervical squamous cell carcinoma and used random survival forest algorithms to develop a prognostic signature capable of predicting disease free survival. Wang et al. used hierarchical clustering of combined multi-omic datasets from TCGA and CPTAC to identify an antitumor immune signature in patients with colon cancer.

The articles in this special issue of *Cancer Biomarkers* are only a small sampling of the various approaches to using AI and machine learning to improve the accuracy and predictive power of biomarkers for cancer and other diseases, collated into one issue. There is a continuing urgent need for more effective strategies for improving the early detection of cancers, particularly those that are missed due to erroneous interpretation of pathological- and radiological images, due to either unclear images or operator-based error in reading the images. Cutting-edge artificial intelligence (AI) systems have been shown to improve sensitivity and specificity of interpretations of both imaging and non-image data for breast, lung, prostate, and cervical cancers. In this special issue of *Cancer Biomarkers* we have highlighted some of the most relevant contributions to this challenge. Many similar articles have been published in *Cancer Biomarkers* as part of the regular journal series.

References

[1] S. Chen, Z. He, X. Han, X. He, R. Li, H. Zhu, D. Zhao, C. Dai, Y. Zhang, Z. Lu, X. Chi and B. Niu, How big data and high-performance computing drive brain science, *Genomics Proteomics Bioinformatics* **17** (2019), 381–392. doi: 10.1016/j.gpb.2019.09.003.

[2] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W.R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu and D. Hassabis, Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13), *Proteins* **87** (2019), 1141–1148. doi: 10.1002/prot.25834.

[3] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W.R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu and D. Hassabis, Improved protein structure prediction using potentials from deep learning, *Nature* **577** (2020), 706–710. doi: 10.1038/s41586-019-1923-7.

[4] M. Leclercq, B. Vittrant, M.L Martin-Magniette, M.P. Scott Boyer, O. Perin, A. Bergeron, Y. Fradet and A. Droit, Large-scale automatic feature selection for biomarker discovery in high-dimensional OMICs data, *Front Genet* **16** (2019), 452.

[5] B.-J.M. Webb-Robertson, L.M. Bramer, B.A. Stanfill, S.M. Reehl, E.S. Nakayasu, T.O. Metz, B.I. Frohnert, J.M. Norris, R.K. Johnson, S.S. Rich and M.J. Rewers, Prediction of the development of islet autoantibodies through integration of environmental, genetic, and metabolic markers, *J Diabetes* **13** (2021), 143–153.

[6] C.-Y. Cheng, Y. Li, K. Varala, J. Bubert, J. Huang, G.J. Kim, J. Halim, J. Arp, H.-J.S. Shih, G. Levinson, S.H. Park, H.Y. Cho, S.P. Moose and G.M. Coruzzi, Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships, *Nat Commun* **12** (2021), 5627.