Guest Editorial: Deep Learning For Genomics

Barbara Di Camillo and Giuseppe Nicosia

1 Introduction

Thanks to the development of high-throughput technologies, a huge amount of omics data is being produced relative to DNA and RNA sequences and (and also) abundance at individual subject or even at individual cell level. In particular, the genomics field is rich in data thanks to the rapid reduction in the cost of genetic sequencing. On the other hand, deep learning is transforming the field of many machine learning applications, such as computer vision and natural language processing, by effectively leveraging on big amount of data and is now emerging as a promising approach for many genomics modeling tasks.

In particular, recurrent neural networks have been used to predict methylation status, alternative splicing, DNA and RNA binding, and protein structure prediction since their framework is suitable to deal with sequential data, such as genomics sequences. On the other hand, convolutional neural networks, which are traditionally applied to image processing and biomedical imaging, have been also used to predict, for example, gene expression from histone modification data, or to classify samples based on gene expression. Autoencoders have been used for protein function prediction and, more recently, to the task of dimensionality reduction and zero imputation in single cell transcriptomic.

The scope of this special section is to discuss novel algorithms, methodologies and applications of deep learning to genomic studies with focus on their potentialities and challenges. After a rigorous review process, six articles were selected for publication in this special section. They are briefly discussed in the following.

2 A Brief Overview of the Papers in This Special Section

The authors proposed different solutions to cope with DNA sequence data, data representation and data integration, mainly based on attention heads, convolutional neural networks, graph neural network, and autoencoders.

A first paper [1] addresses biological sequences annotation for the automatic detection of transcription start sites, translation initiation sites, methylation sites, etc. using the deep learning transformer-based architectures recently proposed for natural language processing, allowing for fast processing of long DNA sequences. The model processes the genome in sequential segments of nucleotides processed through multiple layers with the use of multiple attention heads. The authors also implement and discuss the use of a convolutional layer in the attention heads to improve the predictive capabilities of the model.

A second contribution [2] copes with genomic data high dimensionality and paucity of samples adopting a drop-feature technique to classify tumor subtypes/stages. The main idea is to iteratively determine an "optimal" feature subset by codifying a 0-1 drop-feature vector integrated in the input layer.

Cristovao *et al.* [3] also deal with cancer subtyping. In particular, they focus on Breast Cancer and on the use of semi-supervised learning to overcome, at least partially, the lack of samples in the field when compared to the number of features. In particular, the authors investigate the use of both feed-forward neural networks trained with different tumor types but then used only to predict breast cancer subtypes and Variational Autoencoders to first learn embedded representations of data and then, based on the compressed representation, train the classifier.

Autoencoders are adopted also in [4] to combine different types of omics data, such as RNA sequencing, micro RNAs and methylation data for the identification of subtypes of diseases. In particular, the authors compare Standard Deep Autoencoder, where the different data sources are stacked together in input, and a Disjointed Deep Autoencoder, where each source of omics data source has its own representation prior to having a combined representation as dimensionality reduction steps before clustering.

Interestingly, Nguyen *et al.* [5] propose a graph convolutional network for drug response prediction on different cell lines. The authors codify chemical information of drugs as a multi-dimensional binary feature vector describing drug properties and interactions between pairs of atoms as an adjacency matrix. Following a graph neural network, a fully connected layer is used to convert the result to 128 dimensions that are combined with another 128 dimensions vector obtained after applying convolutional neural network layers to the genomic features of cell lines, i.e., mutations, represented in one-hot encoding. Then, the combination of drug's features and cell line's features are combined and put through two fully-connected layers to predict the half-maximal inhibitory concentrations of each drug for each specific cell line.

Finally, Mahapatra *et al.* [6] address protein-protein interaction prediction combining deep neural network and

- *Barbara Di Camillo is with the Department of Information Engineering, Department of Comparative Biomedicine and Food Science, University of Padova, 35122 Padova, Italy. E-mail: barbara.dicamillo@unipd.it.*
- *Giuseppe Nicosia is with the Department of Biochemistry, Cambridge Systems Biology Centre, University of Cambridge, CB2 1TN Cambridge, U.K., and also with the Department of Biomedical & Biotechnological Sciences, School of Medicine, University of Catania, 95124 Catania, Italy. E-mail: gn263@cam.ac.uk.*

Digital Object Identifier no. 10.1109/TCBB.2021.3080094

extreme gradient boosting classifier. They use a fusion of three amino acid sequence-based features as input and deep learning to extract the protein features that are then given as input to the extreme gradient boosting classifier.

Overall application of deep learning to genomics data still suffers of some drawbacks, such as the difficulty of application when available data are not sufficient to train the model, though it is sometimes possible to recur to transfer learning or semi-supervised learning. The most promising applications seem indeed related to data integration and representation. However, a main drawback is the ability to explain the model outcomes. To gain biological insight in the field of genomics, high accuracy is as necessary as the understanding of the underlying reasons of the outcome predictions and the inference of casual relationships among variables and with the outcome. In this context, explainable and interpretable artificial intelligence is emerging as a further research direction in genomic applications.

Acknowledgments

The guest editors thank all those who helped make this special section possible, especially the Editor-in-Chief, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* editorial office, and the authors and reviewers of the contributions.

References

[1] J. Clauwaert, and W. Waegeman, "Novel transformer networks for improved sequence labeling in genomics," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 1, pp. 97–106, Jan./Feb. 2022.

[2] Z. Chen, W. Zhang, H. Deng, and K. Zhang, "Effective cancer subtype and stage prediction via dropfeature-DNNs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 1, pp. 107–120, Jan./Feb. 2022.

[3] F. Cristovao *et al.*, "Investigating Deep Learning Based Breast Cancer Subtyping Using Pan-Cancer and Multi-Omic Data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 1, pp. 121–134, Jan./Feb. 2022.

[4] G. Viaud, P. Mayilvahanan, and P.-H. Cournède, "Representation learning for the clustering of multi-omics data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 1, pp. 135–145, Jan./Feb. 2022.

[5] T. Nguyen, G. T. T. Nguyen, T. Nguyen, and D.-H. Le, "Graph convolutional networks for drug response prediction" *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 1, pp. 146–154, Jan./Feb. 2022.

[6] S. Mahapatra, V. R. Gupta, S. S. Sahu, and G. Panda, "Deep neural network and extreme gradient boosting based hybrid classifier for improved prediction of protein-protein interaction," *IEEE/ACM Trans. Computat. Biol. Bioinf.*, vol. 19, no. 1, pp. 155–165, Jan./Feb. 2022.

Barbara Di Camillo received the master's degree in electronic engineering from the University of Padova, in 2000, and the PhD degree (Dottorato di Ricerca) in biomedical engineering from the University of Padova, in 2004. She is full professor in computer science with the Department of Information Engineering, University of Padova. She was a visiting scientist with Mayo Clinic (Rochester, MN), "endocrinology research group" led by Dr. S. Nair, in 2002 and at the "Bioinformatics Center of University of Technology of Graz" (Graz, Austria), led by Prof. Trajanoski in 2003. Her research activity is centered in the development and application of advanced modeling, data mining and machine learning methods for high-throughput biological data analysis in the field of Bioinformatics and Systems Biology. In particular, she has developed and applied different methods for robust biomarker discovery, predictive modeling and clustering of next generation sequencing (NGS) data. She has also a great expertise in the development and application of differential equation based models, Boolean and Bayesian Networks for modeling of transcriptional networks and signaling pathways.

Giuseppe Nicosia received the PhD degree in computer science and synthetic biology. He is associate professor in bioengineering with the Department of Biomedical & Biotechnological Sciences, School of Medicine, University of Catania. From 2017 he is visiting professor with the University of Cambridge, U.K. He has got the National Scientific Qualification for full professor withn the area of Bioengineering and Computer Science. He is currently involved in the design and development of machine learning algorithms and artificial intelligence methods for systems biology, synthetic biology, and bioengineering. He is author and co-author of more than 150 papers in international journals and conference proceedings, book chapters and 20 edited books. He has chaired several international conferences, summer schools, advanced courses and workshops in the research areas of artificial intelligence, machine learning, data science, computational biology, synthetic biology, and bioengineering. He has founded and chaired the Advanced Course on Data Science and Machine Learning – ACDL (2018–2021), the Synthetic and Systems Biology Summer School – SSBSS (2014–2019), and the Conference on machine Learning, Optimization and Data Science – LOD (2015–2021). He received several research grants as PI in the area of Bioengineering, Synthetic Biology, Systems Biology, Artificial Intelligence, Machine Learning and Data Science. His research has received awards at several premier conferences and journals. Giuseppe Nicosia is regularly serving as area chair or senior program committee member for several top Conferences. He has been visiting professor and scholar with the Massachusetts Institute of Technology (MIT), University of Cambridge, University of Florida, Nicolaus Copernicus University and New York University.