International Journal of
Molecular Sciences

MDPI

Editorial

Editorial of Special Issue "Deep Learning and Machine Learning in Bioinformatics"

Mingon Kang [1] and Jung Hun Oh [2,*]

[1] Department of Computer Science, University of Nevada, Las Vegas, NV 89154, USA; mingon.kang@unlv.edu
[2] Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
* Correspondence: ohj@mskcc.org

In recent years, deep learning has emerged as a highly active research field, achieving great success in various machine learning areas, including image processing, speech recognition, and natural language processing, and now rapidly becoming a dominant tool in biomedicine [1]. In particular, a dramatically increasing number of deep learning-based approaches have been proposed in biomedical image analysis and biosignal processing, as well as medical prediction modeling. However, the application of deep learning to genomics and bioinformatics has been rather limited, perhaps due to the combined difficulties of interpretation as well as steep data requirements.

One of the major challenges is that many approaches in deep learning and traditional machine learning are based on the assumption that the number of samples is huge in order to train models with a vast number of features. The situation in medicine is often reversed by necessity: the number of features desired to be analyzed is often one or two orders of magnitude greater than the number of samples. Researchers must contend with this fundamental issue, and in the end must be content with models that are consistent with the data.

In this Special Issue entitled "Deep Learning and Machine Learning in Bioinformatics", submissions address the application of deep learning and novel machine learning methods to diverse bioinformatic problems and provide practical guidance. These methods include useful approaches that may improve predictive performance and separately enhance our understanding of biological mechanisms of target diseases.

Among the 55 submissions reviewed, 21 were accepted, including 17 research articles and 4 reviews, with 124 contributors. The contributions were global, for the accepted papers originating from 12 countries, including Australia (2), China, France, Italy (3), Japan (2), Poland, South Korea (2), Spain, Sweden, Taiwan, Thailand, and the United States (5). Figure 1 shows the map of countries with the symbol ★ for the first or corresponding authors of the accepted papers.

Figure 1. A map of countries with the symbol ★ for the first or corresponding authors of the accepted papers.

Ten research papers demonstrated the application of deep learning to various kinds of biological data. Le et al. proposed an ensemble neural network to identify essential genes via word embedding features from genomic sequences [2]. Persson Hodén et al. developed

Citation: Kang, M.; Oh, J.H. Editorial of Special Issue "Deep Learning and Machine Learning in Bioinformatics". *Int. J. Mol. Sci.* **2022**, *23*, 6610. https://doi.org/10.3390/ijms23126610

Received: 7 June 2022
Accepted: 10 June 2022
Published: 14 June 2022

a convolutional neural network (CNN) model capable of efficiently identifying true mRNA cleavage sites, which was implemented as an R package called smartPARE [3]. Nosi et al. proposed a neural network method to detect MET exon 14 skipping events using RNAseq data from The Cancer Genome Atlas (TCGA) archive for lung cancer [4]. Alessandri et al. developed a new autoencoder model, called Sparsely Connected Autoencoders, to improve the traditional decoder model for better identifying biological features from single cell data [5]. Al Mamun et al. developed a multi-run concrete autoencoder to identify a stable set of features which was applied to TCGA genome-wide lncRNA expression profiles in 12 cancers, resulting in the identification of key lncRNAs [6]. Lee et al. introduced a peptide data augmentation method, which was employed to predict spider neurotoxic peptides, showing improved predictive power when coupled with a CNN model [7]. Madani et al. developed a novel deep learning sequence-based solubility predictor, called DSResSol, for fast, reliable, and inexpensive prediction of protein solubility [8]. Zulfiqar et al. developed a 1D CNN-based model, named Deep-4mCGP, to identify 4mC sites in Geobacter pickeringii [9]. Roethel et al. developed a deep learning architecture for a holistic sequential and structural analysis of biomolecules [10].

Hazra et al. employed generative adversarial networks (GAN) to create synthetic nucleic acid sequences of the cat genome [11].

Seven research papers used traditional (non-deep learning) machine learning approaches to analyze biological data. Two computational methods were introduced, PUP-Fuse [12] and PredNTS [13], for the prediction of pupylation sites and nitrotyrosine sites, respectively, by integrating multiple sequence representations coupled with a random forest approach. Rodin et al. proposed a novel computational pipeline to dissect the response to cancer immunotherapy, employing systems biology and Bayesian network techniques on flow cytometry data [14]. Campos et al. employed machine learning approaches to identify essential genes common to both Caenorhabditis elegans and Drosophila melanogaster [15]. Charoenkwan et al. developed a sequence-based predictor, named iBitter-Fuse, to identify bitter peptides by fusing multi-view features [16]. Jabeen et al. adopted a random forest model to identify novel high activity agonists of human ectopic olfactory receptors [17]. Pouryahya et al. proposed a network-based clustering method coupled with optimal mass transport theory to predict cell line-drug sensitivity, and showed that random forest modeling conducted on the resulting cell line-drug clusters outperformed alternative computational methods in predicting in vitro drug responses [18].

Four papers reviewed the use of deep learning or machine learning approaches to biological data analysis. Auslander et al. reviewed machine learning/deep learning approaches incorporated to establish bioinformatics and computational biology frameworks in the areas of molecular evolution, protein structure analysis, systems biology, and disease genomics [19]. Del Giudice et al. comprehensively reviewed machine learning/deep learning solutions for computational problems in bulk and single-cell RNA-sequencing data analysis [20]. Banegas-Luna et al. discussed the interpretability of machine learning/deep learning methods in cancer research [21]. Defresne et al. reviewed deep learning methods used for protein design [22].

In summary, the articles in this Special Issue provide a range of reviews and updates to the use of deep learning and machine learning in bioinformatics.

Author Contributions: Conceptualization, M.K. and J.H.O.; funding acquisition, J.H.O.; writing—original draft preparation, J.H.O.; writing—review and editing, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by an NIH grant (R21 CA234752) and MSK Cancer Center Support grant (P30 CA008748).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Brief. Bioinform.* **2017**, *18*, 851–869. [CrossRef] [PubMed]
2. Le, N.Q.K.; Do, D.T.; Hung, T.N.K.; Lam, L.H.T.; Huynh, T.T.; Nguyen, N.T.K. A Computational Framework Based on Ensemble Deep Neural Networks for Essential Genes Identification. *Int. J. Mol. Sci.* **2020**, *21*, 9070. [CrossRef] [PubMed]
3. Persson Hoden, K.; Hu, X.; Martinez, G.; Dixelius, C. smartPARE: An R Package for Efficient Identification of True mRNA Cleavage Sites. *Int. J. Mol. Sci.* **2021**, *22*, 4267. [CrossRef] [PubMed]
4. Nosi, V.; Luca, A.; Milan, M.; Arigoni, M.; Benvenuti, S.; Cacchiarelli, D.; Cesana, M.; Riccardo, S.; Di Filippo, L.; Cordero, F.; et al. MET Exon 14 Skipping: A Case Study for the Detection of Genetic Variants in Cancer Driver Genes by Deep Learning. *Int. J. Mol. Sci.* **2021**, *22*, 4217. [CrossRef]
5. Alessandri, L.; Ratto, M.L.; Contaldo, S.G.; Beccuti, M.; Cordero, F.; Arigoni, M.; Calogero, R.A. Sparsely Connected Autoencoders: A Multi-Purpose Tool for Single Cell omics Analysis. *Int. J. Mol. Sci.* **2021**, *22*, 12755. [CrossRef]
6. Al Mamun, A.; Tanvir, R.B.; Sobhan, M.; Mathee, K.; Narasimhan, G.; Holt, G.E.; Mondal, A.M. Multi-Run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers. *Int. J. Mol. Sci.* **2021**, *22*, 11919. [CrossRef]
7. Lee, B.; Shin, M.K.; Hwang, I.W.; Jung, J.; Shim, Y.J.; Kim, G.W.; Kim, S.T.; Jang, W.; Sung, J.S. A Deep Learning Approach with Data Augmentation to Predict Novel Spider Neurotoxic Peptides. *Int. J. Mol. Sci.* **2021**, *22*, 12291. [CrossRef]
8. Madani, M.; Lin, K.; Tarakanova, A. DSResSol: A Sequence-Based Solubility Predictor Created with Dilated Squeeze Excitation Residual Networks. *Int. J. Mol. Sci.* **2021**, *22*, 13555. [CrossRef]
9. Zulfiqar, H.; Huang, Q.L.; Lv, H.; Sun, Z.J.; Dao, F.Y.; Lin, H. Deep-4mCGP: A Deep Learning Approach to Predict 4mC Sites in Geobacter pickeringii by Using Correlation-Based Feature Selection Technique. *Int. J. Mol. Sci.* **2022**, *23*, 1251. [CrossRef]
10. Roethel, A.; Bilinski, P.; Ishikawa, T. BioS2Net: Holistic Structural and Sequential Analysis of Biomolecules Using a Deep Neural Network. *Int. J. Mol. Sci.* **2022**, *23*, 2966. [CrossRef]
11. Hazra, D.; Kim, M.R.; Byun, Y.C. Generative Adversarial Networks for Creating Synthetic Nucleic Acid Sequences of Cat Genome. *Int. J. Mol. Sci.* **2022**, *23*, 3701. [CrossRef]
12. Auliah, F.N.; Nilamyani, A.N.; Shoombuatong, W.; Alam, M.A.; Hasan, M.M.; Kurata, H. PUP-Fuse: Prediction of Protein Pupylation Sites by Integrating Multiple Sequence Representations. *Int. J. Mol. Sci.* **2021**, *22*, 2120. [CrossRef]
13. Nilamyani, A.N.; Auliah, F.N.; Moni, M.A.; Shoombuatong, W.; Hasan, M.M.; Kurata, H. PredNTS: Improved and Robust Prediction of Nitrotyrosine Sites by Integrating Multiple Sequence Features. *Int. J. Mol. Sci.* **2021**, *22*, 2704. [CrossRef]
14. Rodin, A.S.; Gogoshin, G.; Hilliard, S.; Wang, L.; Egelston, C.; Rockne, R.C.; Chao, J.; Lee, P.P. Dissecting Response to Cancer Immunotherapy by Applying Bayesian Network Analysis to Flow Cytometry Data. *Int. J. Mol. Sci.* **2021**, *22*, 2316. [CrossRef]
15. Campos, T.L.; Korhonen, P.K.; Young, N.D. Cross-Predicting Essential Genes between Two Model Eukaryotic Species Using Machine Learning. *Int. J. Mol. Sci.* **2021**, *22*, 5056. [CrossRef]
16. Charoenkwan, P.; Nantasenamat, C.; Hasan, M.M.; Moni, M.A.; Lio, P.; Shoombuatong, W. iBitter-Fuse: A Novel Sequence-Based Bitter Peptide Predictor by Fusing Multi-View Features. *Int. J. Mol. Sci.* **2021**, *22*, 8958. [CrossRef]
17. Jabeen, A.; de March, C.A.; Matsunami, H.; Ranganathan, S. Machine Learning Assisted Approach for Finding Novel High Activity Agonists of Human Ectopic Olfactory Receptors. *Int. J. Mol. Sci.* **2021**, *22*, 11546. [CrossRef]
18. Pouryahya, M.; Oh, J.H.; Mathews, J.C.; Belkhatir, Z.; Moosmuller, C.; Deasy, J.O.; Tannenbaum, A.R. Pan-Cancer Prediction of Cell-Line Drug Sensitivity Using Network-Based Methods. *Int. J. Mol. Sci.* **2022**, *23*, 1074. [CrossRef]
19. Auslander, N.; Gussow, A.B.; Koonin, E.V. Incorporating Machine Learning into Established Bioinformatics Frameworks. *Int. J. Mol. Sci.* **2021**, *22*, 2903. [CrossRef]
20. Del Giudice, M.; Peirone, S.; Perrone, S.; Priante, F.; Varese, F.; Tirtei, E.; Fagioli, F.; Cereda, M. Artificial Intelligence in Bulk and Single-Cell RNA-Sequencing Data to Foster Precision Oncology. *Int. J. Mol. Sci.* **2021**, *22*, 4563. [CrossRef]
21. Banegas-Luna, A.J.; Pena-Garcia, J.; Iftene, A.; Guadagni, F.; Ferroni, P.; Scarpato, N.; Zanzotto, F.M.; Bueno-Crespo, A.; Perez-Sanchez, H. Towards the Interpretability of Machine Learning Predictions for Medical Applications Targeting Personalised Therapies: A Cancer Case Survey. *Int. J. Mol. Sci.* **2021**, *22*, 4394. [CrossRef]
22. Defresne, M.; Barbe, S.; Schiex, T. Protein Design with Deep Learning. *Int. J. Mol. Sci.* **2021**, *22*, 11741. [CrossRef]