Deep Learning in Bioinformatics

The emergence of deep learning models has revitalized the broad field of artificial intelligence in the past 20 years. The foundation of most modern deep learning models is artificial neural networks. They are composed of multiple inter-connected layers, each of which consists of separate simple computational units called neurons. The input information flows through the network as follows: each layer receives input data for each of its neurons, each neuron then executes a simple user-defined function, and then the output of the neuron is transmitted as input to neurons in the next layer. Two neurons are said to be connected if a neuron in one layer sends output to the other neuron in the next layer. The connections are weighted, reflecting the contribution to the prediction. The learning process of a neural network is the updating of these connection weights, based on prediction errors made with training data. By composing the numerous simple functions executed by each neuron in a network structure, complex relationships between inputs and their relevance to the output can be learned. The term "*deep learning*" was introduced to refer to the use of many layers of neural networks to progressively generate features from the original input data. Networks with more layers can learn more complex functions, thus explaining the power of deep learning. However, since the input features are sent through a complex network of functions, it is difficult to pinpoint the most informative features. A more powerful family of neural networks, which is able to automatically extract and select features from raw inputs such as images and sentences, have been developed and achieved remarkable performance in various applications. The most successful architectures are convolutional neural networks (CNNs) and recurrent neural network (RNN), which are now the cornerstone of all leading methods in image classification and natural language processing, respectively. However, despite the success of deep learning, it is often criticized for the risk of overfitting, lack of model interpretability, and large numbers of model parameters. Nevertheless, deep learning models have achieved record-breaking results in the fields of image classification, natural language processing, speech recognition, and more recently in bioinformatics. In this special issue, we report representative projects that exemplify this recent success.

We begin by providing an exoteric introduction of deep learning and concrete examples, and implementations of its representative applications in bioinformatics. [Li *et. al.*] starts from the recent achievements of deep learning in the bioinformatics field, pointing out the problems which are suitable to use deep learning. It then introduces deep learning in an easy-to-understand fashion, from shallow neural networks to legendary convolutional neural networks, legendary recurrent neural networks, graph neural networks, generative adversarial networks, variational autoencoder, and the most recent state-of-the-art architectures. After that, eight examples are provided, covering five bioinformatics research directions and all the four kinds of data types, with the implementation written in Tensorflow and Keras. Finally, it discusses some common issues, such as overfitting and interpretability, that users will encounter when adopting deep learning methods and provides corresponding suggestions.

Detecting cancer-related genes and their interactions is a crucial task in cancer research. In [Su *et. al.*], the authors proposed a deep unsupervised biclustering method, to detect coding genes, microRNAs (miRNAs), and their interactions related to a particular cancer or a cancer subtype using their expression data from the same set of samples. Firstly, biclusters specific to a particular type of cancer are detected based on rectified factor networks and ranked according to their associations with general cancers. Secondly, coding genes and miRNAs in each bicluster are prioritized by considering their differential expression and differential correlation values, protein-protein interaction data, and potential cancer markers. Finally, a rank fusion process is used to obtain the final comprehensive rank by combining multiple ranking results. Results on breast cancer datasets show that this method outperforms other methods in detecting breast cancer-related coding genes and miRNAs. Furthermore, the method is very efficient in computing time, which can handle tens of thousands of genes/miRNAs and hundreds of patients on a desktop. This work may aid researchers in studying the genetic architecture of complex diseases and improving the accuracy of diagnosis.

Polyadenylation signals (PAS) are found in most protein-coding and some non-coding genes in eukaryotes. Their accurate recognition improves understanding gene regulation mechanisms and recognition of the 3'-end of transcribed gene regions where premature or alternate transcription ends may lead to various diseases. Although different methods and tools for in-silico prediction of genomic signals have been proposed, the correct

identification of PAS in genomic DNA remains challenging due to a huge number of non-relevant hexamers identical to PAS hexamers. In [Bajic *et. al.*], the authors developed a novel method for PAS recognition. The method is implemented in a hybrid PAS recognition model (HybPAS), which is based on deep neural networks (DNNs) and logistic regression models (LRMs). One model is developed for each of the 12 most frequent human PAS hexamers. DNN models appear the best for eight PAS types (including the two most frequent PAS hexamers), while LRM appeared best for four PAS types. The new models use different combinations of signal processing-based, statistical, and sequence-based features as input. The results obtained on human genomic data show that HybPAS outperforms the well-tuned state-of-the-art Omni-PolyA models, with Omni-PolyA models being better for two PAS types. On average, HybPAS reduces the error by 30.29%.

Due to the large numbers of **transcription factors (TFs) and cell types**, querying binding profiles of all valid TF/cell type pairs is not experimentally feasible. To address this issue, [Quang *et. al.*] presents a convolutional-recurrent neural network model, called FactorNet, to computationally impute the missing binding data. FactorNet trains on binding data from reference cell types to make predictions on testing cell types by leveraging a variety of features, including genomic sequences, genome annotations, gene expression, and signal data. FactorNet implements several convenient strategies to reduce runtime and memory consumption. FactorNet ranked among the top teams in the ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge, achieving first place on six of the 13 final round evaluation TF/cell type pairs, the most of any competing team.

Enhancer is a DNA sequence of a genome that controls transcription of downstream target genes. Enhancers are known to be associated with certain epigenetic signatures. Traditional machine learning tools, such as CSI-ANN, ChromHMM, and RFECS, were developed for predicting enhancers using various epigenetic features. However, predictions by different tools vary widely and quite a significant portion of enhancer predictions does not agree. Thus, computational methods for enhancer prediction should be further developed. In [Lim *et. al.*] a hybrid neural network called Enhancer-CRNN, a convolutional neural network (CNN) followed by an RNN, was developed and they were used to predict enhancer regions with histone modification marks as input. Hybridization of both neural networks outperformed existing prediction tools in experiments with GM12878, H1hesc, HeLaS3, and HepG2 cell lines. On average, 13 to 17 percent of the enhancers predicted by Enhancer-CRNN were cell type-specific. With the trained model, optimized virtual input histone marks were generated to provide a deeper insight into how histone modification marks can represent enhancer regions in which histone marks indicate active or repressed enhancers.

The **inverse virtual screening** is an important technique in the early stage of drug development. This technique can provide preliminary clues for unknown molecules, which is useful in the following researches. In [Zhang *et. al.*], a dense fully connected neural network (DFCNN) algorithm, IVS2vec, is utilized to build a prediction model. This model is able to perform a binary classification. Given a query molecule, proteins can be classified into two sets. One set includes the potential targets with high possibilities to bind with the query molecule and the other includes the proteins with low possibilities to bind with the query molecule. Alternatively, IVS2vec can output a score reflecting binding possibility of the association between a protein and a molecule. It has been demonstrated that IVS2vec can be used to detect possible therapeutic targets as well as to find targets related to adverse drug reactions. Moreover, IVS2vec is efficient in processing massive compound databases.

Multi-omics integration presents high potentials for diseases-relevant translational discoveries. By analyzing abundance levels of heterogeneous molecules over time, we may uncover biological interactions and networks that were previously unidentifiable. However, to effectively perform integrative analysis of temporal multi-omics, computational methods must account for the heterogeneity and complexity in the data. To this end, [Chung *et. al.*] performed unsupervised classification of proteins and metabolites in mice during cardiac remodeling using two innovative deep learning approaches. First, long short-term memory (LSTM)-based variational autoencoder (LSTM-VAE) was trained on time-series numeric data. The low-dimensional embeddings extracted from LSTM-VAE were then used for clustering. Second, deep convolutional embedded clustering (DCEC) was applied on images of temporal trends. Instead of a two-step procedure, DCEC performs a joint optimization for image reconstruction and cluster assignment. Pathway enrichment analysis using the Reactome knowledgebase demonstrated that deep learning methods yielded higher numbers of significant

biological pathways than conventional clustering algorithms. In particular, DCEC resulted in the highest number of enriched pathways, suggesting the strength of its unified framework based on visual similarities.

The **human microbiome** plays a number of critical roles, impacting almost every aspect of human health and well-being. Conditions in the microbiome have been linked to a number of significant diseases. Additionally, revolutions in sequencing technology have led to a rapid increase in publicly-available sequencing data. Consequently, there have been growing efforts to predict disease status from metagenomic sequencing data, with a proliferation of new approaches in the last few years. Many of these efforts have explored utilizing deep learning. [Lapierre *et. al.*] reviews some of these methods and the algorithms that they are based on, with a particular focus on deep learning methods. A deeper analysis of Type 2 Diabetes and obesity datasets has eluded improved results, using a variety of machine learning and feature extraction methods. The article offers perspectives on study design considerations that may impact results and future directions the field can take to improve results.

Chromosomal higher-order folding is an important feature of genome organization, which plays a critical role in genome functioning including transcriptional regulation. [Bkhetan *et. al.*] presents deep learning models of the human genome three-dimensional structure that combine one dimensional (linear) sequence specificity, epigenomic information and transcription factor binding profiles, with the polymer-based biophysical simulations in order to explain the extensive long-range chromatin looping observed in ChIA-PET experiments. Furthermore, the authors designed three-dimensional models of chromatin contact domains (CCDs) using real (ChIA-PET) and predicted looping interactions. Initial results show a similarity between both types of 3D computational models (constructed from experimental or predicted interactions). This observation confirms the association between genome sequence, epigenomic and transcription factor profiles, and three-dimensional interactions.

The identification of **therapeutic biomarkers** predictive of drug response is crucial in personalized medicine. A number of computational models to predict response of anti-cancer drugs have been developed as the establishment of several pharmacogenomics screening databases. In [Su *et. al.*], a deep cascaded forest model, Deep-Resp-Forest, was proposed to classify the anti-cancer drug response as “sensitive” or “resistant”. With this model, diverse molecular data could be effectively integrated to provide more information than a single type of data for the classification. Two structures based on the multi-grained scanning to transform the raw features into high-dimensional feature vectors and integrate the diverse data were proposed. The original deep and time-consuming architecture of cascade forest was improved by a feature optimization operation, which emphasized the most discriminative features across layers. On the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) datasets, Deep-Resp-Forest demonstrated the promising use of deep learning and deep forest approach on the drug response prediction tasks.

Digital breast tomosynthesis (DBT) is a newly developed three-dimensional tomographic imaging modality in the field of breast cancer screening designed to alleviate the limitations of conventional digital mammography-based breast screening methods. A computer-aided detection (CAD) system was designed for masses in DBT using a faster region based convolutional neural network (faster-RCNN). An efficient detection architecture of convolution neural network with a region proposal network (RPN) was used for each slice to generate region proposals (i.e., bounding boxes) with a mass likelihood score. In each DBT volume, a slice fusion procedure was used to merge the detection results on consecutive 2D slices into one 3D DBT volume. The performance of the CAD system was evaluated using free-response receiver operating characteristic (FROC) curves. The results suggest that the faster R-CNN has the potential to augment the prescreening and FP reduction in the CAD system for masses.

Last but not least, the **biomedical literature** provides a rich source of knowledge such as protein-protein interactions (PPIs), drug-drug interactions (DDIs) and chemical-protein interactions (CPIs). Bio-medical relation extraction aims to automatically extract biomedical relations from biomedical text for various biomedical research. State-of-the-art methods for biomedical relation extraction are primarily based on supervised machine learning and therefore depend on (sufficient) labeled data. However, creating large sets of training data is

prohibitively expensive and labor-intensive, especially so in biomedicine as domain knowledge is required. In contrast, there is a large amount of unlabeled biomedical text available in PubMed. Hence, computational methods capable of employing unlabeled data to reduce the burden of manual annotation are of particular interest in biomedical relation extraction. [Zhang *et. al.*] presents a novel semi-supervised approach based on variational autoencoder (VAE) for biomedical relation extraction. The model consists of the following three parts, a classifier, an encoder and a decoder. The classifier is implemented using multi-layer CNNs, and the encoder and decoder are implemented using both bidirectional long short-term memory networks (Bi-LSTMs) and CNNs, respectively. The semi-supervised mechanism allows the model to learn features from both the labeled and unlabeled data. Experimental results show that this method effectively exploits the unlabeled data to improve the performance and reduce the dependence on labeled data. This is the first semi-supervised VAE-based method for (biomedical) relation extraction. The results suggest that exploiting such unlabeled data can be greatly beneficial to improved performance in various biomedical relation extraction, especially when only limited labeled data (e.g. 2000 samples or less) is available in such tasks.

We are confident that the collection of articles in this special issue will serve as an inspiring compendium for future deep learning advancement and deployment in biological and biomedical fields.