SCIENCE CHINA
Life Sciences

Application of Deep Learning in Genomics

Jianxiao Liu, Jiying Li, Hai Wang and Jianbing Yan

Citation: SCIENCE CHINA Life Sciences; doi: 10.1007/s11427-020-1804-5

View online: http://engine.scichina.com/doi/10.1007/s11427-020-1804-5

Published by the Science China Press

REVIEW

Application of deep learning in genomics

Jianxiao Liu[1,2], Jiying Li[3], Hai Wang[4,*] & Jianbing Yan[1,*]

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, 430070;
[2]College of Informatics, Huazhong Agricultural University, Wuhan, 430070;
[3]Microsoft Corporation, 1 Microsoft Way, Redmond, WA, 98052, USA;
[4]National Maize Improvement Center, Key Laboratory of Crop Heterosis and Utilization, Joint Laboratory for International Cooperation in Crop Molecular Breeding, China Agricultural University, Beijing 100193, China

Corresponding authors (Jianbing Yan, email: yjianbing@mail.hzau.edu.cn; Hai Wang, email: wanghai@cau.edu.cn)

Received June 13, 2020; accepted August 15, 2020

Abstract: In recent years, deep learning has been widely used in diverse fields of research, such as speech recognition, image classification, autonomous driving and natural language processing. Deep learning has showcased dramatically improved performance in complex classification and regression problems, where the intricate structure in the high-dimensional data is difficult to discover using conventional machine learning algorithms. In biology, applications of deep learning are gaining increasing popularity in predicting the structure and function of genomic elements, such as promoters, enhancers, or gene expression levels. In this review paper, we describe the basic concepts in machine learning and artificial neural network, followed by elaboration on the workflow of using convolutional neural network in genomics. Then we provide a concise introduction of deep learning applications in genomics and synthetic biology at the levels of DNA, RNA and protein. Finally, we discuss the current challenges and future perspectives of deep learning in genomics.

Keywords: Deep learning; genomics; convolutional neural network

1. Introduction

Artificial intelligence (AI) powers many aspects of modern society, from traditional industries (agriculture, industry, transportation, *etc*) to modern industries (education, culture, catering, tourism, *etc*), and it continues to transform more and more sectors. As the core technology in artificial intelligence, machine learning studies the algorithms that computer systems utilize to perform tasks by learning from data instead of following explicit instructions. Despite their extensive applications, conventional machine learning techniques are limited in their capability to process natural data in their raw forms and learn intricate patterns in complex dataset. Compared to conventional machine learning algorithms, deep neural network stands out with the ability of automatic feature extraction and greater data representation capability in dealing with high-dimensional datasets. This method has gained dramatically improved performance compared to the state-of-the-art in dealing with the complex classification and regression tasks, such as speech recognition, image classification, autonomous driving and natural language processing (Hinton *et al*., 2006; LeCun *et al*., 2015). Meanwhile, deep learning has also been used in the field of genomics such as functional annotation of biological sequences (Ritchie et al. 2015; Libbrecht et al., 2015; Camacho et al., 2018). There have been many beautiful reviews summarizing recent progresses in this area,

including Park et al., 2015; Angermueller et al., 2016; Mamoshina et al., 2016; Min *et al*., 2017; Ching *et al*., 2018; Wainberg *et al*., 2018; Webb, 2018; Yue *et al*., 2018; Zou *et al*., 2019; Wang *et al*., 2020, *etc*. Among them, Angermueller et al. mainly discuss applications of deep learning in regulatory genomics and cellular imaging (Angermueller et al., 2016). Min *et al*. present the current research work of using deep learning in omics, biomedical imaging and biomedical signal processing (Min *et al*., 2017). Ching *et al*. mainly discuss the applications of deep learning in predicting enhancer, promoter, interactions among genomic elements, splicing of transcripts, *de novo* drug design, as well as text mining in healthcare and electronic health records (Ching *et al*., 2018). Zou *et al*. mainly concentrate on the application of deep learning in the fields of regulatory genomics, variant calling and pathogenicity scores, and it also provides a practical guide to tools and resources in deep learning (Zou *et al*., 2019). Wang et al. mainly describe the flow of information from genomic DNA sequences to molecular phenotypes and how to prioritize functional variants in natural populations using deep learning models (Wang *et al*., 2020).

Besides the fast development and applications of deep learning algorithms, the size of biological datasets has grown exponentially. Advanced sequencing technologies enable faster genome sequencing and assembly at reduced costs, and the assembled genomes can be automatically annotated with high-throughput techniques. With unprecedentedly large amount of biological data, modeling the functions of genomic elements becomes increasingly crucial. First, experimentally unravelling important genomic elements for every sequenced genome is unfeasible. Instead, it is more economical to build deep learning models that predict functional genomic elements in well-studied genomes, and apply these models in less well-studied genomes. Second, rather than experimentally determining the phenotypic effects of natural variants, deep learning models can be used to predict variants with desirable functions for downstream crop improvement. Last, interpretation of the above-mentioned models provides novel insights for the biological processes being studied.

Although existing reviews summarize recent progress of deep learning in genomics, an in-depth analysis of deep learning in plant and animal breeding is still lacking. In addition, application of generative models in synthetic biology is rarely mentioned in previous reviews. Here, we firstly discuss concepts and processes in machine learning and the popular deep learning methods. Then we describe common steps in sequence analysis by convolutional neural networks. We then focus on the applications of different deep learning methods in the research related to DNA (enhancer, promoter, non-coding DNA, TSS, methylation states, replication domains, cis-regulatory region, lab-of-origin of DNA, interaction), RNA (alternative splicing, lncRNA, MicroRNA, messenger RNA, expression), and protein (transcription factor, DNA binding proteins, RNA binding proteins). We also describe the applications of deep generative models to generate functional elements (DNA sequence, promoter sequence, protein sequence, single-cell RNA-seq data, Hi-C data). Finally, we discuss the caveats and future perspectives of exploiting deep learning in genomic research as well as plant and animal breeding. Overall, the goal of this article is to summarize recent progress in this field, organize useful recourses in different categories, provide valuable insights to facilitate the application of deep learning in genomic studies, and hopefully point out promising directions of further research in this area.

2. Machine Learning and Deep Learning

2.1 Machine learning

Machine learning algorithms are usually categorized as supervised learning, unsupervised learning and semi-supervised learning. The most common form of machine learning is supervised learning, where each example in the data set is labeled. The machine is expected to learn the mapping from the input to output during the training process and be able to produce sensible prediction on new data. For example, an image classification machine

learning system should be able to classify an unseen image to its category after being trained over hundreds of millions of labeled images, as shown in Figure 1 (Krizhevsky *et al*., 2009). In genomics, we can use supervised learning to predict gene expression levels, population structure and so on (Krogel *et al*., 2004). In contrast, in unsupervised learning, examples in the data set are without pre-existing labels. The learning algorithm is supposed to properly group data examples by learning the function that minimizes the intra-group gap and maximizes inter-group gap. Two of the main methods used in unsupervised learning are principle component analysis (PCA) and clustering analysis (Bowden *et al*., 1997), both of which are widely used in transcriptomic analysis over RNA-Seq datasets (Kiselev *et al*., 2019). Semi-supervised learning falls between supervised and unsupervised learning, as it generates appropriate functions by learning from both labeled and unlabeled data.

2.2 A typical workflow of a machine learning system

A typical workflow in a machine learning system generally includes six steps: data collection, data preprocessing, model training, model evaluation, model usage and model interpretation. The typical process of machine learning systems is shown in Figure 2.

2.3 Deep learning

Deep learning is a machine learning technique that has recently made major breakthroughs in solving problems that have resisted the best efforts in the artificial intelligence community for many years (LeCun *et al*., 2015). Deep learning essentially refers to deep neural network architecture, which consists of an input layer, many hidden layers and an output layer. The multilayer architecture of deep neural network mimics the structure in visual neuroscience and is able to transform the data representation in an increasingly abstract form via non-linear modules. It turns out to be surprisingly successful in learning the non-linear input-output mapping with both increased selectivity and the invariance of the representation. The automatic feature extracting ability with high selectivity and invariance is the key advantage of deep learning.

At present, the following neural network methods have been widely used in genomics: Boltzmann machine (BM), autoencoder (AE), deep belief network (DBN), recurrent neural network (RNN), long short-term memory (LSTM), convolutional neural network (CNN), *etc*. The architecture of these methods is described in the Supplementary materials (Supplementary materials.docx). Until now, the convolutional neural network (CNN) is the most commonly used deep learning model in genomics. The detailed execution process of CNN applied in genome research is shown in Figure 3. One-hot encoding, in which the four nucleotides (A, C, G, and T) are encoded as their corresponding vectors ([1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1]), is commonly used to convert DNA sequences to matrices which serve as inputs for deep learning models.

3. Deep Learning for Genomics

Genomics mainly studies the structure, function, evolution and editing of genomes at the systems level, while at the molecular level, we follow the central dogma of molecular biology to study and characterize the functions of individual molecule in a fine-grained manner. The central dogma of molecular biology refers to the process that genetic information is transferred from DNA to RNA, and then from RNA to protein, that is, to complete the transcription and translation of genetic information (Crick, 1970). Thanks to tremendous advance in high throughput technologies, omics data at all levels of central dogma become available. With the unprecedentedly large amount of omics data, we are probably in the best era to apply machine learning and deep learning at all levels of biological systems (Figure 4).

3.1 Deep learning at the DNA level

3.1.1 Promoter

Promoter is a segment of DNA sequence typically located upstream of the transcription start sites of genes (Busby *et al*., 1994). RNA polymerase and accessory factors recognize and bind to promoters to start transcription. Importantly, conservative sequences within promoters play critical roles in specific binding and transcription initiation by RNA polymerase, therefore accurate prediction of promoters is crucial for interpreting gene expression patterns and understanding genetic regulatory networks. Kh *et al*. applied CNN to construct prediction models to analyze sequence characteristics of promoters in several prokaryotic and eukaryotic organisms, including human, mouse, plant (*Arabidopsis*) and bacteria (*Escherichia coli* and *Bacillus subtilis*). Experiment results demonstrate that deep learning method can predict complex promoter sequence and have significantly higher accuracy compared to previous promoter prediction methods (Kh *et al*., 2017). In addition, Basset is a framework of CNN that learns the functional activity of DNA sequences from genomics data. It applies SGD to learn all model parameters, and computes loss and gain scores for every nucleotide. Basset learns the relevant sequence motifs and the regulatory logic to collectively determine cell-specific DNA accessibility. As claimed by the authors, researchers could benefit from using this framework to understand chromatin accessibility code and annotate every mutation in the genome with its influence on present or potential accessibility (Kelley *et al*., 2016).

Cis-regulatory elements are distributed mainly in noncoding regions of a genome and are involved in the regulation of gene expression. Li *et al*. introduced a supervised deep learning approach to identify active *cis*-regulatory regions (CRRs) across the human genome, and delineated locations of 300,000 candidate enhancers and 26,000 candidate promoters genome-wide (Li *et al*., 2018). In the light of the fact that determining the origin of DNA sequence is difficult and time-consuming, Nielsen *et al*. used CNN to predict the lab-of-origin of a DNA sequence. It turns out that this approach can be extended to unravel sequences of malicious intent (Nielsen *et al*., 2018).

3.1.2 Enhancer

Enhancers are small DNA segments remotely located upstream or downstream of coding regions but could greatly enhance gene expression level via binding to gene transcription machinery (Khoury *et al*., 1983). DEEP is the first enhancer prediction framework using neural network. The method firstly trains support vector machines (SVM) models using different subsets of the original data, then aggregates decisions and uses artificial neural network (ANN) to derive the final prediction (Kleftogiannis *et al*., 2015). Min *et al*. proposed a computational framework of CNN named DeepEnhancer to distinguish enhancers from genomic sequences. Experimental results show that DeepEnhancer has superior efficiency and effectiveness compared to traditional sequence-based classifiers (Min *et al*., 2016). Liu *et al*. developed a deep learning based algorithmic framework (PEDLA) to predict enhancers from massively heterogeneous datasets. PEDLA can learn from massively heterogeneous data to fully capture universal patterns of enhancers. It also generalize enhancer predictions in ways that are mostly consistent across various cell types/tissues (Liu *et al*., 2016). BiRen is another method to predict enhancers using a deep learning-based hybrid structure that is trained with limited experimentally validated noncoding elements. The hybrid model integrates CNN with bidirectional recurrent neural network (BRNN). It makes full use of the power of CNN in sequence encoding and representation, as well as the superior capacity of gated recurrent unit-based bidirectional recurrent neural network (GRU-BRNN) for handling the long-term dependencies in long DNA sequences (Yang *et al*., 2017).

3.1.3 Non-coding region

Noncoding DNA refers to the sequence that does not encode proteins but plays important roles in regulating

various biological processes, such as gene expression, translation, DNA replication and others (Andolfatto, 2005). DanQ is a hybrid framework combining CNN and bi-directional long short-term memory recurrent neural network (BLSTM) to predict non-coding function *de novo* from sequence. DanQ learns a regulatory grammar to improve predictions, and provides novel insights into non-coding genomic regions (Quang *et al*., 2016). In another research, Zhou *et al*. developed a deep learning-based framework (DeepSEA) to predict the noncoding-variant effects *de novo* from sequence. It calculates functional significance scores based on chromatin effect predictions and the evolutionary information-derived scores. DeepSEA directly learns the regulatory sequence code from large-scale chromatin-profiling data, and predicts chromatin effects of sequence alterations with single-nucleotide sensitivity. DeepSEA is the first approach for prioritization of functional variants using *de novo* regulatory sequence information (Zhou *et al*., 2015). Later, Zhou *et al*. used deep-learning-based framework (ASDbrowser) to predict the specific regulatory effects and the deleterious impact of genetic variants, and detect contribution of noncoding mutations to disease. ASDbrowser uses the interactions between DNA binding proteins or RNA binding proteins and their targets as the training dataset. This work demonstrates for the first time the important role of proband-specific signal in regulatory noncoding region (Zhou *et al*., 2019).

3.1.4 Interactions between genomic elements

Predicting enhancer-promoter interactions helps us understand how the genome regulates complex cellular functions in a living organism. SPEID is a deep learning model that predicts enhancer-promoter interactions solely based on sequence features, such as locations of putative enhancers and promoters in a particular cell type. Experiment results show that SPEID more accurately predicts the enhancer-promoter interactions compared to state-of-the-art methods that use non-sequence features extracted from functional genomic signals. It is the first report that uses sequence-based features alone to predict genome-wide enhancer-promoter interactions (Singh *et al*., 2016). Yuan *et al*. developed CNNC method to mine gene-gene relationship by learning from single-cell expression data. CNNC can improve upon prior methods in tasks ranging from predicting transcription factor targets to identifying disease related genes (Yuan *et al*., 2019). In addition, Huang *et al*. proposed an end-to-end prediction model called GCLMI to predict lncRNA-miRNA interactions by combining graph convolution and auto-encoder (Huang *et al*., 2019).

3.1.5 Other domains

Based on the gene annotations in one species, CNN can predict the annotations in a different species if the mechanisms of interpreting the genomes are conserved in the two species. As an example, Khodabandelou *et al*. used CNN to predict the transcription start sites (TSS) across genomes (DeepTSS). The ratio between positive and negative examples was optimized to obtain the highest prediction scores to identify TSS (Khodabandelou *et al*., 2018). Besides, Eser *et al*. introduced an open source data-agnostic flexible integrative deep learning framework (FIDDLE), which learns an unified representation from multiple data types to infer other data types. This framework demonstrates that one data type could be inferred from other sources of data types without manually specifying the relevant features or dataset preprocessing. As a case study, the authors used multiple *Saccharomyces cerevisiae* genomic datasets to predict TSS through the simulation of TSS-seq data (Eser *et al*., 2016).

DNA methylation has important impact on chromatin structure, cell differentiation, cancer progression, DNA stability, DNA conformation and interactions between DNA and proteins, and gene expression. Angermueller *et al*. used deep neural network to predict single-cell methylation states and model the sources of DNA methylation variability (DeepCpG). DeepCpG uncovers both previously known and *de novo* sequence motifs that are associated with methylation changes and methylation variability between cells (Angermueller *et al*., 2017). In another research, Wang *et al*. applied stacked denoising autoencoder deep learning algorithm to predict DNA methylation status of CpG sites. This algorithm uses two stages to train the model: an unsupervised pre-training stage using unlabeled

training data and a supervised fine-tuning stage using labeled data (Wang *et al*., 2016).

DNA replication refers to the process of synthesizing offspring DNA using parent DNA as templates. Liu *et al*. developed a novel hybrid architecture (DNN-HMM) combining deep neural network and hidden Markov model for *de novo* identification of replication domains(Liu *et al*., 2016). DNN-HMM uses the posterior probabilities of states as the output of DNN, and experiment results demonstrate that DNN-HMM significantly outperforms existing methods.

3.2 Deep learning at the RNA level

3.2.1 Splicing

Alternative splicing (AS) refers to the process of producing different splicing isomers of mRNA through executing different splicing modes (selecting different splicing site combinations) on a mRNA precursor. There are emerging research work using neural network to predict AS patterns. One initial work by Leung *et al*. developed a deep neural network to predict splicing patterns in individual tissues and the differences across tissues. Experiment results show that the deep architecture surpasses the performance of the Bayesian method for predicting AS patterns (Leung *et al*., 2014). In order to more accurately predict AS regulatory factors, research work has been done to improve the neural network model. For example, Anupama *et al*. developed a new target function using Bayesian neural network (BNN) and deep neural network (DNN) for AS prediction (Anupama *et al*., 2017).

A splice junction refers to the boundary between a pair of adjacent exon and intron. Identifying splice junctions of a gene is important for deciphering its primary structure and function. In order to realize the precise identification of spice junction, Lee *et al*. exploited deep RNN to model DNA sequences and predict splice junctions thereon. This approach significantly outperforms conventional machine learning methods as well as a recent deep belief network-based technique (Lee *et al*., 2015).

3.2.2 Non-coding RNA

Non-coding RNA (ncRNA) refers to RNA that does not encode proteins, which can be roughly classified as miRNAs (micro-RNAs), snRNAs (small-nuclear RNAs), siRNAs (short-interfering RNAs), shRNAs (short-hairpin RNAs), circRNAs (circular-RNAs) and lncRNAs (long-non-coding RNAs) (Hüttenhofer *et al*., 2005). Recent studies show that ncRNAs play important roles in RNA modification, RNA splicing, regulation of transcription and translation, RNA interference, *etc* (Wang *et al*., 2018). In the past few years, several deep learning approaches have been proposed to predict ncRNAs utilizing sequence statistics.

lncRNA

The long non-coding RNAs (lncRNAs) play significant roles in various cellular functions, such as immune response, genetic regulations, and embryonic pluripotency (Fatica *et al*., 2014; Deng *et al*., 2018; Deng *et al*., 2019). Research has been done using neural network to identify lncRNAs. Tripathi *et al*. proposed lincRNA prediction method (DeepLNC) using deep neural network. In their approach, k-mer information is generated based on Shannon entropy function to improve the classification accuracy. Two datasets, LNCipedia and RefSeq, are used as experiment benchmark, and the method successfully identified known lncRNAs with 99 % accuracy (Tripathi *et al*., 2016). Yu *et al.* adopted the autoencoder deep learning algorithm to detect lincRNA. This algorithm captures useful features and the information correlation along genome sequences for lincRNA detection. The experimental results show that the autoencoder algorithm has better performance compared with SVM and traditional neural network (Yu *et al*., 2017).

MicroRNA

MicroRNAs (miRNAs) are endogenous non-coding RNAs with regulatory functions in eukaryotic organisms. MicroRNAs play a crucial role in post-transcriptional gene regulation by attaching itself to the 3' untranslated region of the target mRNA (Xu *et al*., 2018). Park *et al*. proposed a novel learning approach (deepMiRGene) that identifies precursor miRNAs using RNNs, specifically LSTM network. Applying learning algorithm in microRNA prediction is difficult due to the palindromic structure of a precursor miRNA. To this end, deepMiRGene divides the input sequence into the forward and backward streams and each structure stream is learned in different sequential directions (Park *et al*., 2016). Lee *et al*. proposed an end-to-end miRNA target prediction framework (deepTarget) using the RNN-based auto-encoding. By combining unsupervised and supervised learning approaches, deepTarget not only achieves an unprecedented high level of accuracy, but also makes manual feature extraction unnecessary. DeepTarget successfully discovers the inherent sequence representations due to the fact that it processes miRNA and RNA sequences with RNN-based autoencoders without alignment (Lee *et al*., 2016).

3.2.3 Messenger RNA

Messenger RNA (mRNA) is transcribed from DNA and conveys the genetic information by guiding protein synthesis. Hill *et al*. used deep RNN to discover complex biological rules and decipher RNA protein-coding potential. Their method trains a gated RNN on human mRNA and lncRNA sequences firstly, and then uses it to predict protein-coding potential. It surpasses the state-of-the-art methods despite being trained with less data and no prior concept of what features define mRNA (Hill *et al*., 2018). Sample *et al*. used CNN to predict the effect of human 5' UTR variants on ribosome loading. They combined polysome profiling of 280,000 randomized 5' untranslated regions with deep learning to build a model that predicts translation efficiency from human 5' UTR sequences. In addition, they also used the genetic algorithm to design new 5' UTR sequences, which accurately direct specified levels of ribosome loading (Sample *et al*., 2019).

3.2.4 Expression

Gene expression refers to the process of synthesizing functional RNA with genetic information from genes. Gene expression is affected by many factors at various levels, including genetic variants at the DNA level. Recently, more and more research work, including the neural network method, concentrates on the gene expression prediction based on genomic sequence. Chen *et al*. proposed a deep learning method (D-GEX) to infer the expression of target genes from the expression of landmark genes (Chen *et al*., 2016). Besides, DeepChrome is a CNN trained on histone modification data to predict gene expression. DeepChrome extracts complex interactions among important features automatically. Specifically, it uses a novel optimization-based technique to generate feature pattern maps, and visualize the combinatorial interactions among histone modifications (Singh *et al*., 2016). In addition, Xie *et al* applied a new deep learning model named multilayer perceptron with stacked denoising autoencoder (MLPSAE) to predict gene expression profiles from genotypes. Experiment results show that it outperforms the methods of MLP-SAE without dropout, Lasso and random forests (Xie *et al*., 2017). Cuperus *et al*. used a model to predict the protein expression of the 5' UTR of mRNAs. The trained CNN with random library performs well at predicting the protein expression of both the random and native 5' UTRs. Their method can also capture the effect of sequence variation adjacent to the coding region in several biological processes including transcription, translation and protein stability (Cuperus *et al*., 2017). ExPecto is a modeling framework for *ab initio* prediction of tissue-specific gene expression levels. This framework integrates CNN with spatial feature transformation and L2-regularized linear models to predict tissue-specific expression (Zhou *et al*., 2018). Finally, Washburn *et al*. developed two CNN architectures to predict mRNA expression levels from DNA promoter and/or terminator regions. Their first work is to predict whether a given gene is expressed or unexpressed by constraining training and testing sets to include different gene families. The second work is to predict which of the two compared gene orthologs has higher mRNA abundance. In the second work, evolutionarily informed comparisons between orthologous genes is used to both

control and leverage evolutionary divergence(Washburn *et al*., 2019).

Gupta *et al*. applied the deep architectures to learn intricate structure in gene expression data for gene clustering. This method uses denoising autoencoder deep architectures to pre-train data in an unsupervised manner and learn the properties of gene expression profiles. The generated features by the model are useful for gene clustering and would facilitate understanding the interactions and regulation among genes (Gupta *et al*., 2015).

3.3 Deep learning at the protein level

3.3.1 Transcription factor

Transcription factors (TFs) are DNA binding proteins that bind to gene promoter and enhancer regions, and thus play an important role in gene expression regulation. Predicting TF binding sites has attracted more and more researchers in recent years. Quang and Xie developed a convolutional-recurrent neural network model (FactorNet) to interpret binding patterns and reveal insights into regulatory grammar. They also introduced several novel strategies to reduce the computation overhead of deep neural network (Quang *et al*., 2019). In another research, Chen *et al*. used a hybrid approach between kernel methods and deep neural network, convolutional kernel network (CKN), to improve the prediction of TF binding sites (Chen *et al*., 2017). Gapped k-mers frequency vectors (gkm-fvs) is an effective sequence-based prediction (e.g., TF binding site prediction) method (Ghandi *et al*., 2014). However, it is computationally expensive, especially for a large kernel matrix and large amount of data. To solve this problem, Cao *et al*. proposed a flexible and scalable framework (gkm-DNN) to achieve efficient feature representation and accurate prediction using deep neural networks (DNN). Experiment results show that gkm-DNN not only overcomes the drawbacks of high dimensionality, colinearity and sparsity of gkm-fvs, but also produces better accuracy compared with gkm-SVM in much shorter training time (Cao *et al*., 2017). Lanchantin *et al*. proposed the Deep Motif Dashboard (DeMoDashboard) to explore three different DNN architectures for TF binding site prediction (Lanchantin *et al*., 2016).

DNA binding proteins play significant roles in transcription, translation, DNA repair, alternative splicing and replication machinery. Predicting the sequence specificities of a protein can help interpret a genomic sequence to detect potential binding sites. Alipanahi *et al*. adapted CNN to predict binding sequence specificities and patterns (DeepBind). DeepBind can discover new patterns even when the locations of patterns within sequences are unknown. In addition, DeepBind can predict deleterious SNVs in promoters and identify deleterious genomic variants (Alipanahi *et al*., 2015). DeeperBind is another novel doubly-deep model for the prediction of protein binding specificities with respect to DNA probes. DeeperBind makes full use of the complementary modeling capabilities of LSTM and CNN. Compared to DeepBind, DeeperBind removes the positional dimension of the intermediate features and it is capable of dealing with varying-length sequences by exploiting LSTM layers (Hassanzadeh *et al*., 2016). In a separate research, Zeng *et al*. applied a CNN architecture to predict DNA sequence binding using a large compendium of transcription factor datasets. Experimental results show that deploying more convolutional kernels is always important for motif-based tasks. In addition, the proposed method has improved performance compared to DeepBind through a systematic exploration of CNN architectures (Zeng *et al*., 2016).

To study the binding of TFs to DNA sequence in a cell line without corresponding ChIP-seq data, the prerequisite question is determining the presence of binding motif in the DNA sequence. However, even if the motif is present, it only contains sequence information but cannot reflect the cell type-specificity of TF binding. To this end, Qin *et al*. combined deep neural network with a multi-task learning setting to share information across transcription factors and cell lines. The developed TFImpute achieves cell type-specific binding prediction for TF-cell line combinations without ChIP-seq data (Qin *et al*., 2017).

3.3.2 RNA-specific binding proteins

RNA-binding proteins (RBPs) play important roles in multiple cellular processes, such as RNA editing, translational regulation, alternative splicing and mRNA localization, *etc*. Zhang *et al*. proposed a deep learning framework (deepnet-rbp) to predict structural binding preferences and binding sites of RBPs. The proposed deepnet-rbp is the first study of integrating additional RNA tertiary structural features to improve the model performance (Zhang *et al*., 2016). Besides, iDeep is a hybrid framework of CNN and DBN to integrate multiple heterogeneous datasets to predict RBP interaction sites on RNAs. The DBN learns high-level features that are determined by hidden variables for different inputs. The CNN component of iDeep captures low-level regulatory motifs with biological functions, which are recurring patterns in RNA sequences (Pan *et al*., 2017).

3.4 Deep generative models in biology

In the field of deep learning, two most commonly used and efficient generative models are Variational auto-encoder (VAE) and Generative Adversarial Networks (GAN). At present, the above two methods start to be used in genomics studies.

3.4.1 Variational autoencoder

Variational autoencoder is a kind of neural network that maps the input to the same-sized output via encoder and decoder. Encoder extracts and compresses the high-dimensional input data to a bottleneck distribution presentation, and decoder subsequently re-constructs an output based on the bottleneck distribution. VAE is commonly used to generate new data or to denoise data (Kingma *et al*., 2013; Rezende *et al*., 2014). In genomics, VAE has been used by several groups to generate new data, such as microbial genomes (Nissen *et al*., 2018). Grønbech *et al*. used VAE to learn biologically plausible groupings of scRNA-seq data with higher quality. The network predicts gene expression counts using appropriate discrete probability distribution as likelihood functions (Grønbech *et al*., 2018). In another study, researchers used VAE to generate protein sequences. Sinai *et al*. presented an embedding of natural protein sequences using VAE to predict how mutations affect protein function (Sinai *et al*., 2017). Davidsen *et al*. used VAE to generate T cell receptor protein sequences, which can perform accurate cohort frequency estimation. They also demonstrated that VAE-like models can distinguish between real sequences and generated sequences according to a recombination-selection model (Davidsen *et al*., 2019; Isacchini *et al*., 2019).

3.4.2 Generative adversarial networks

Generative adversarial networks (GAN) is deep generative model that generates new synthetic data via an adversarial process. GAN is composed of a generative model and a discriminative model, where the generative model generates new data point based on the captured data distribution, and discriminative model can estimate the probability of a sample coming from training data rather than from the generative model. The main purpose of GAN is that by training both models, the generator is able to synthesize new instances of data that the discriminator is unable to distinguish from the real data (Goodfellow *et al*., 2014). Until now, generative adversarial network has also been used in genomics, such as for inference of target gene expression profiles (Wang *et al*. 2018), reproducing high-resolution Hi-C data (Hong *et al*., 2019; Liu *et al*., 2019), *etc*. Same as VAE, GAN has also been used in the generation and augmentation of single-cell RNA-seq data, such as cscGAN (Marouf *et al*., 2020), scPhere (Ding *et al*., 2019), scRNAseq-WGAN-GP (Ghahramani *et al*., 2018), scRNA-seq data imputation (Gunady *et al*., 2019), *etc*.

In addition, GAN also has important applications of generating sequences, including protein sequence, DNA sequence, promoter, *etc*. Anand *et al*. applied GAN to generate protein structures for fast *de novo* protein design (Anand *et al*., 2018). Repecka *et al*. developed the ProteinGAN to learn natural protein sequence diversity and

generate functional protein sequences (Repecka *et al*., 2019). Chhibbar *et al*. used GAN to generate protein sequences from antibiotic resistance genes. Experiment result shows that the generated sequences can be used to study and expand functionality associated with the antibiotic resistance determinants (Chhibbar *et al*., 2019). In 2017, Killoran *et al*. firstly used GAN to generate and design DNA sequence. It opens the door for applying deep generative models to advance genomics research (Killoran *et al*., 2017). Later, Gupta *et al*. applied GAN to generate synthetic DNA sequences encoding proteins of variable length. They proposed a novel feedback-loop architecture named FBGAN to optimize the synthetic gene sequences for desired properties (Gupta *et al*., 2018). Instead of optimizing the input seed of a pre-trained GAN by Killoran *et al*., 2017, Linder *et al*. optimized the weights of the generator to maximize both sequence fitness and diversity. They developed the deep exploration networks (DENs) to obtain generators capable of sampling hundreds of thousands of high-fitness DNA sequences (Linder *et al*., 2019). Not only that, Yelmen *et al*. have trained GANs and restricted Boltzmann machines (RBMs) to learn the high dimensional distributions of real genomic datasets and created high quality artificial genomes (Yelmen *et al*., 2019). Different from generating DNA and protein sequences, Wang *et al*. applied GAN in *de novo* promoter sequence design to generate entirely new promoter sequences in *Escherichia coli* (Wang *et al*., 2019). This work indicates the potential of deep generative models in designing genetic elements in the future.

3.5 A summary of deep learning in genomics

The above-mentioned applications of deep learning models in genomics are summarized in Table 1. In the table, ANN refers to artificial neural network, BRNN refers to bidirectional recurrent neural network. MLP-SAE refers to multi-layer perceptron and stacked denoising auto-encoder. DA refers to denoising autoencoder. CKN refers to convolutional kernel network. FNN denotes the feedforward neural network. BLSTM refers to bi-directional long short-term memory recurrent neural network. BNN refers to Bayesian neural network, and SD-AE refers to stacked denoising autoencoder. Concrete information can be found in the supplementary materials.

Table 1. The applications of deep learning in genomics

Level	Type	Authors	Abbr.	Methods	Website
DNA	Enhancer	Kleftogiannis *et al*., 2015	DEEP	ANN	http://cbrc.kaust.edu.sa/deep/
		Liu *et al*., 2016	PEDLA	DNN	https://github.com/wenjiegroup/PEDLA
		Min *et al*., 2016	DeepEnhancer	CNN	-
		Yang *et al*., 2017	BiRen	CNN, BRNN	https://github.com/wenjiegroup/BiRen
	Promoter	Kelley *et al*., 2016	Basset	CNN	http://www.github.com/davek44/Basset
		Kh *et al*., 2017	CNNProm	CNN	http://www.softberry.com
	Non-coding DNA	Zhou *et al*., 2015	DeepSEA	CNN	http://deepsea.princeton.edu/
		Quang *et* al., 2016	DanQ	CNN, BLSTM	http://github.com/uci-cbcl/DanQ
		Zhou *et al*., 2019	ASDbrowser	CNN	https://hb.flatironinstitute.org/asdbrowser/help
	TSS	Eser *et al*., 2016	FIDDLE	CNN	-
		Khodabandelou *et al*., 2018	DeepTSS	CNN	https://github.com/StudyTSS/DeepTSS/
	Methylation states	Wang *et al*., 2016	DeepMethyl	SD-AE	http://dna.cs.usm.edu/deepmethyl/
		Angermueller *et al*., 2017	DeepCpG	CNN	https://github.com/PMBio/deepcpg
	Replication	Liu *et al*., 2016	DNN-HMM	DNN	https://github.com/wenjiegroup/DNN-HMM
	cis-regulatory	Li *et al*., 2018	DECRES	FNN	https://github.com/yifeng-li/DECRES
	lab-of-origin	Nielsen *et al*., 2018	-	CNN	https://github.com/VoigtLab/predict-lab-origin.
	Interaction	Singh *et al*., 2016	SPEID	LSTM	-
		Yuan *et al*., 2019	CNNC	CNN	https://github.com/xiaoyeye/CNNC
		Huang *et al*., 2019	GCLMI	AE	-
RNA	Alternative	Leung *et al*., 2014	-	DNN	-

	splicing	Lee *et al*., 2015	-	RNN	-
		Anupama *et al*., 2017	-	BNN, DNN	https://majiq.biociphers.org/jha_et_al_2017/
	lncRNA	Tripathi *et al*., 2016	DeepLNC	DNN	http://bioserver.iiita.ac.in/deeplnc
		Yu *et al*., 2017	-	AE	https://github.com/ningyu12/lincRNA_predict/
	MicroRNA	Park *et al*., 2016	deepMiRGene	LSTM	-
		Lee *et al*., 2016	deepTarget	RNN-based AE	http://data.snu.ac.kr/pub/deepTarget
	Messenger RNA	Hill *et al*., 2018	mRNN	RNN	http://github.com/hendrixlab/mRNN
		Sample *et al*., 2019	Optimus 5-Prime	CNN	https://github.com/pjsample/human_5utr_modeling
	Expression	Gupta *et al*., 2015	-	DA	-
		Chen *et al*., 2016	D-GEX	DNN	https://github.com/uci-cbcl/D-GEX
		Singh *et al*., 2016	DeepChrome	CNN	https://github.com/QData/DeepChrome
		Xie *et al*., 2017	MLP-SAE	MLP-SAE	https://github.com/shilab/MLP-SAE/
		Cuperus *et al*., 2017	Deep-learning-yeast-UTRs	CNN	https://github.com/Seeliglab/2017---Deep-learning-yeast-UTRs
		Zhou *et al*., 2018	ExPecto	CNN	https://github.com/FunctionLab/ExPecto
		Washburn *et al*., 2019	-	CNN	https://bitbucket.org/bucklerlab/p_strength_prediction/
Protein	**Transcription factor**	Lanchantin *et al*., 2016	DeMoDashboard	CNN, RNN	-
		Chen *et al*., 2017	CKN-Seq	CKN	https://gitlab.inria.fr/dchen/CKN-seq
		Cao *et al*., 2017	gkm-DNN	DNN	http://page.amss.ac.cn/shihua.zhang/software.html
		Qin *et al*., 2017	TFImpute	CNN	https://bitbucket.org/feeldead/tfimpute
		Quang *et al*., 2019	FactorNet	CRNN	https://github.com/uci-cbcl/FactorNet
	DNA binding proteins	Alipanahi *et al*., 2015	DeepBind	CNN	http://tools.genes.toronto.edu/deepbind/
		Hassanzadeh *et al*., 2016	DeeperBind	LSTM, CNN	-
		Zeng *et al*., 2016	-	CNN	http://cnn.csail.mit.edu
	RNA binding proteins	Zhang *et al*., 2016	deepnet-rbp	RBM	https://github.com/thucombio/deepnet-rbp
		Pan *et al*., 2017	iDeep	CNN, DBN	https://github.com/xypan1232/iDeep
Generative models	**Protein sequence**	Sinai *et al*., 2017	-	VAE	https://github.com/samsinai/VAE_protein_function
		Davidsen *et al*., 2019	-	VAE	https://github.com/matsengrp/vampire/
		Anand *et al*., 2018	-	GAN	-
		Repecka *et al*., 2019	ProteinGAN	GAN	https://github.com/biomatterdesigns/ProteinGAN
		Chhibbar *et al*., 2019	W-GAN	GAN	-
	DNA sequence	Killoran *et al*., 2017	-	GAN	-
		Gupta *et al*., 2018	FBGAN	GAN	-
		Yelmen *et al*., 2019	-	GAN, RBM	-
	Promoter	Wang *et al*., 2019	WGAN-GP	GAN	-

Based on our review of deep learning in genomics, we concluded that CNN is the most widely used method at present. The popularity of CNN is due to the merit of local connection via convolution kernel, sharing kernel weights, automated feature extraction, simple yet efficient learning procedures, high selectivity and high invariance. Application of GANs is also merging in genomics due to their roles in unsupervised learning and advantages of producing clearer and realistic samples, saving cost, and so on. For the experiment dataset, most of the current research work use the Human ENCODE dataset (de Souza, 2012). Dataset of mouse, *Saccharomyces cerevisiae*, yeast, maize and sorghum are also analyzed later. Figure 4 summarizes the published deep learning models along the central dogma of molecular biology.

PyTorch and TensorFlow are the two most commonly used frameworks for deep learning. PyTorch was released by Facebook's AI Research lab in 2017. It primarily includes APIs in Python to be more declarative and thus fits smoothly into the Python machine learning ecosystem. TensorFlow, on the other hand, was created at Google Brain at 2015. It has APIs in multiple programming languages. However, it is the high-level Keras APIs for TensorFlow that has proven very successful within the deep learning community. PyTorch is preferred by deep-learning

researchers, while TensorFlow is widely used in production environment. The reason for the divide is two-folds. PyTorch's intuitive APIs combined with eager execution mode make it easy for quick testing on simple solutions and smaller-scale models. But in terms of production environment deployment, TensorFlow makes it easy to maintain and update the trained models on the server-side and allows compression of trained model so that it can be used on mobile devices. We have summarized and listed the deep learning framework used in various genomics studies in the supplementary table. However, no definitive answer exists regarding which one is better. As a rule of thumb, PyTorch is a general recommendation for deep learning researchers, while TensorFlow might be a better choice for deploying model in production environment. In either case, understanding the concepts and principles of deep neural networks regardless of framework is the key to build robust and efficient models.

4. Caveats of Deep Learning Algorithm

4.1 Model architecture

Different neural network architectures have their own advantages and disadvantages. Appropriately selecting neural network or combining neural networks for specific biological problems requires deep understanding of the network as well as the biological context. For example, BiRen uses the hybrid model that integrates CNN and BRNN to predict enhancers (Yang *et al*., 2017). DanQ combines CNN and BLSTM to predict non-coding function *de novo* from sequence (Quang *et al*., 2016). Both example indicates specific reasons and probably quite a lot of trials on model selecting and testing. To let the biologists focus on the biological problem and be worry-free when using deep learning tools, automatic model selection could hopefully provide friendly usage of various deep learning models. For example, AutoGenome is a tool that enables researchers to perform end-to-end learning with the most cutting edge neural network architectures easily (Liu *et al*., 2019). In addition, optimizing existing deep neural networks and combining machine learning methods are promising research directions. For example, DEEP combines SVM and ANN to realize enhancer prediction (Kleftogiannis *et al*., 2015). Sample *et al*. used CNN and genetic algorithm to better predict the effect of human 5' UTR variants on ribosome loading (Sample *et al*., 2019). Liu *et al*. used deep learning and hidden Markov model in *de novo* identification of replication domains using replication timing profiles (Liu *et al*., 2016).

4.2 Hyperparameter optimization

Hyperparameters refer to model parameters that are set before training. By contrast, the values of other parameters are adjustable during model training stage. Hyperparameters are related to model selection and learning process. Better hyperparameters are conducive to the rapid convergence of the model, and could improve process of model construction. At present, it is common to start with multiple sets of parameters, and then select the parameters with the best learning effect to train the model. As we known, hyperparameter configurations are data and application dependent, tuning hypermeters are often necessary due to limited pre-knowledge about the data. In deep learning, based on empirical knowledge, some hyperparameters, such as number of hidden layers, length of convolutional filters, and learning rate, can be recommended to users. For example, setting the number of hidden layers to 3 has been suggested in a large number of genomic research (Kelley *et al*., 2016; Khodabandelou *et al*., 2018; Washburn *et al*., 2019; Zhou *et al*., 2015). Besides, the number of units in neural network is mainly related to specific predicted objects. For example, it is usually set to 0~500 for the region prediction problem, a range of 16, 32, 64, 128 for prediction of transcription factor, binding proteins and expression prediction, and 0~1000 for the function annotation, splicing, methylation states and interaction prediction.

4.3 Training set/test set splitting

The training set is used for training the model, and the test set is used to evaluate model performance after

training is completed. The partition of training set/test set should keep the consistency of data distribution as much as possible, and avoid the influence of extra deviation on the final result. The commonly used training set/test set splitting methods include hold-out, cross validation, bootstrap, *etc*. The legend representation of cross validation about training/test dataset splitting is shown in Figure 5(A). Importantly, it is necessary to split appropriate training and testing sets according to the data characteristics of specific problems, such as the specific biological relevance. For example, Washburn *et al*. used the gene-family guided splitting method to solve the problem of closely related genes appear in both training set and testing set. They used gene-family relationships to ensure that genes within the same family do not appear both in the training and testing sets (Washburn *et al*., 2019).

4.4 Ensemble learning

As we know, putting heads together could come up with good ideas and two heads are usually better than one. In the same way, it is likely to achieve better performance by combining the classification results of several classifiers instead of relying on a single classifier. Ensemble learning denotes the generation of multiple learners through certain rules and the integration of all learners as the final comprehensive output. It could effectively solve a common problem in deep learning that results of neural network method sometimes differ greatly and hard to reproduce (Liu *et al*., 1999). In the field of deep learning, ensemble learning mainly includes the following three modes: varying training data, varying models and varying combination.

Firstly, from the aspect of varying training data, the commonly used methods include *k*-fold cross validation and resampling. (1) In the *k*-fold cross validation, all training data sets are divided into *k*-sub training sets, then each sub training set is used to train the model separately, and finally results of *k* models are integrated as the final result. (2) In the resampling method, the composition of each training set can be different, and there may be duplicate data in different training sets. Resampling method allows the trained model to have slightly different expectations for sample density and different generalization errors. Secondly, for varying models, the following three kinds of methods are mainly used. (1) Different parameters are randomly used to initialize the models with the same configuration. (2) Change the configuration parameters of the model, including the hidden state vectors of different dimensions, hidden layers, learning rates, learning strategies, regularization methods, *etc*. (3) When a single model may need a long training time, it saves the best model periodically in the training process of other models and then integrates the saved models. Lastly, for varying combination, the simple way is to average the prediction results of all models. The improved method called model blending is to average the prediction results of all models by weighting, in which the weight is set using the validation set. In addition, we can design a new model to dynamically learn the weight of each model, which is generally called model stacking or stacked generalization.

4.5 Complexity within the black box

Neural networks are often considered as black boxes because they are difficult to interpret. It is usually tricky to discover the key features that affect the decision-making in the neural networks. To this end, developed methods include fitting a simple model in the local area of input (Ribeiro *et al*., 2016; Turner, 2016), or observing the change of output by providing a local perturbation to the input (Shrikumar *et al*., 2017; Sundararajan *et al*., 2017; Zeiler *et al*., 2014). However, both methods rely on some fixed deep neural network frameworks, and the results are usually not stable and vulnerable to noise. In order to solve these problems, the idea of knockoffs is introduced into the neural network (Lu *et al*., 2018). By constructing the knockoff features of the original features, the processes happen inside the black box of neural network are somehow revealed. The legend representation of using knockoff features to open the black box of deep learning is shown in Figure 5(B). Besides, we can use gradient and perturbation methods to identify the importance of sequence regions that a neural network uses to make decisions. For example, Washburn *et al*. applied two gradient-based methods (Saliency and DeepLIFT) and one perturbation-based method (Occlusion) to identify motifs/putative *cis* elements. Their pseudo-gene model indicates

that promoter is more important than the terminator to determine the on/off of gene expression (Washburn *et al*., 2019).

5 Future Perspectives

5.1 Deep generative models

Deep generative models are powerful methods to effectively learn complex data distribution using unsupervised learning and generate new data points that are indistinguishable from the training set. In only few years, it has already achieved great success in many fields, such as creating new image content, and still remains as one of the hottest research areas. As described in section 3.4, VAE and GAN have been widely used in synthetic biology, such as for generation of DNA sequence (Linder *et al*., 2019; Yelmen *et al*., 2019), promoter sequence (Wang *et al*., 2019), protein sequence (Sinai *et al*., 2017; Repecka *et al*., 2019), single-cell RNA-seq data (Marouf *et al*., 2020; Grønbech *et al*., 2018) and high-resolution Hi-C data (Hong *et al*., 2019; Liu *et al*., 2019), *etc*. The generated new DNA elements, as described above, would possibly save the sequencing cost on a large number of samples, and more importantly, these functional elements could constitute the building blocks for synthetic biology. The representation of using GANs to generate sequences is shown in Figure 5(C).

5.2 Interaction prediction

The gene-gene interactions are important research directions in functional genomics. Constructing the gene regulatory network mainly uses the gene expression data, which is called as reverse engineering (Margolin *et al*., 2006). The commonly used gene regulatory network construction methods include weighting matrix, Boolean network, linear function, mutual information, Bayesian network, *etc*. Until now, there are few researches concentrating on gene interaction prediction based on genome sequence. In addition, chromatin loops play important roles in transcriptional regulation by bringing together remote regulatory elements and their target genes. Such long-range interactions contribute to variations in gene expression, metabolism, and terminal traits (Peng *et al*., 2019). Recently, using deep learning to detect these interactions has attracted researchers' attention. Singh *et al*. used a deep learning model (SPEID) to predict enhancer-promoter interactions using only the sequence information (Singh *et al*., 2016). This is the first work that uses sequence-based features alone to predict genome-wide enhancer-promoter interactions. Later, GCLMI is developed to predict lncRNA-miRNA interactions by combining graph convolution and auto-encoder (Huang *et al*., 2019). With continuous development of 3D genome technology and increasing amount of interactome data, it will be increasingly convenient for researchers to detect the sequence interactions. We believe that there will be more and more research work using deep learning to predict interactions.

5.3 Transfer learning

Transfer learning aims to use the knowledge learned from one environment to facilitate the learning tasks in another environment. The parameters in the pre-trained model are re-used in the new model as feature extractor. The parameters in the new model will be trained on a relevant small dataset. In such a way, transfer learning alleviates the demand on large data size and still enables us to produce an accurate model. The representation of transfer learning is shown in Figure 5(D).

In scenarios that the data sets in two tasks are closely related, the pre-trained model can be shared to the new model through transfer learning. It is unrealistic to train the large-scale neural networks with tens of millions of parameters from scratch when the number of data samples is small. Training a large model with inadequate data would easily lead to the overfitting problem. Using the pretrained model obtained by learning similar problems with available large training dataset, transfer learning effectively helps to solve the problem with insufficient samples. For example, we can share the model parameters among different species. The model parameters of

training rice can be used in the model that studies the function of maize sequence. Similarly, model parameters can be re-used among different sequence data types. The model parameters of sequence specific DNA binding protein identification can be used in the model for sequence specific RNA binding protein.

5.4 Applications in plant and animal breeding

How can deep learning models be used to guide the genetic improvement of livestock and crops? There are at least three approaches. (1) Deep learning models are extremely powerful at predicting the effects of natural genomic variants on molecular phenotypes, irrespective of the frequencies of these variants in natural populations, or the magnitude of their effects. Thus deep learning models, combined with models linking molecular phenotypes to terminal traits (such as association mapping and genomic selection), will prove helpful to guide breeding programs. As shown in Figure 6(A), we can predict the variation loci for specific phenotypes using deep learning except the prediction of the type of sequence and expression level of specific gene sequence. (2) We can use deep learning to detect DNA sequence interactions (such as gene-gene interactions, interactions between variant regulatory elements and target genes, lncRNA-miRNA interactions), which tremendously help us to draw the genetic variation topology network of gene expression as well as phenotype variation (Peng *et al*., 2019). The detected sequence interactions can help the functional genomic research, thus enhancing studies about the genetic architecture associated with complex traits. As shown in Figure 6(A), deep learning can be used to predict whether there is interaction or not between two sequences, and thus serve the functional genomic research of plants and animals. (3) Generated DNA elements could constitute the building blocks for synthetic biology. We can use deep generative models (eg. GAN, VAE) to generate novel genomic elements, so as to achieve desirable molecular phenotypes or terminal traits. As shown in Figure 6(B), functional sequences for specific DNA elements (such as enhancer, high expression) generated by deep generative models would be combined or further integrated to bio-engineered system to produce desirable phenotypes. Moreover, sequences with multiple functions may be generated via joint training of several deep learning models that are targeting at maximize individual phenotype separately. A versatile sequence that can interact with several sequences can also be generated, and it will be helpful for the subsequent research of biological synthesis. More than that, deep generative models can produce sequences favorable for multiple better phenotypes, thus to simplify the study on synthetic biology. Taken together, we propose that deep learning will be a key technique in future livestock and crop breeding.

Conclusion

Deep learning has transformed many aspects of genomics studies, the usage of CNN in sequence analysis and application of GANs in generating new dataset have especially gained a great deal of success. We expect to witness more successes in the near future because of deep learning's merits of automated feature extraction, little requirement for handcrafted engineering, high selectivity and high invariance. These properties allow researchers to easily take advantages of huge amount of data. New learning algorithms and architectures that are currently being developed, as well as continuous application of deep learning in genomics will innovate and accelerate research in sequence analysis, function prediction, expression prediction, interaction identification and breeding of plants and animals.

Compliance and ethics

The author(s) declare that they have no conflict of interest.

Acknowledgements This work was supported by the National Key Research and Development Program of China (2016YFD0100303), the National Natural Science Foundation of China (31525017), the Fundamental Research Funds for the Central Universities (2662018JC030).

References

Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015, 33(8): 831-839.

Anand N, Huang P. Generative modeling for protein structures. Advances in Neural Information Processing Systems. 2018: 7494-7505.

Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila. Nature, 2005, 437(7062): 1149-1152.

Angermueller C, Lee H J, Reik W, et al. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biology, 2017, 18(1): 67.

Angermueller C, Parnamaa T, Parts L, et al. Deep learning for computational biology. Molecular Systems Biology, 2016, 12(7):878.

Anupama J, Gazzara M R, Yoseph B. Integrative deep models for alternative splicing. Bioinformatics, 2017, 33(14): i274-i282.

Bowden R, Mitchell T A, Sarhadi M. Cluster based nonlinear principle component analysis. Electronics letters, 1997, 33(22): 1858-1859.

Busby S, Ebright R H. Promoter structure, promoter recognition, and transcription activation in prokaryotes. Cell, 1994, 79(5): 743-746.

Camacho D M, Collins K M, Powers R K, et al. Next-generation machine learning for biological networks. Cell, 2018, 173(7): 1581-1592.

Cao Z, Zhang S H. gkm-DNN: efficient prediction using gapped k-mer features and deep neural networks. bioRxiv, 2017.

Chen D X, Jacob L, Mairal J. Predicting transcription factor binding sites with convolutional kernel networks. bioRxiv, 2017.

Chen Y, Li Y, Narayan R, et al. Gene expression inference with deep learning. Bioinformatics, 2016, 32(12):1832-1839

Chhibbar P, Joshi A. Generating protein sequences from antibiotic resistance genes data using Generative Adversarial Networks. arXiv preprint arXiv:1904.13240, 2019.

Ching T, Himmelstein D S, Beaulieu-Jones B K, et al. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface, 2018, 15(141): 20170387.

Crick F. Central dogma of molecular biology. Nature, 1970, 227(5258): 561-563.

Cuperus J T, Groves B, Kuchina A, et al. Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. Genome Research, 2017, 27(12): 2015-2024.

Davidsen K, Olson B J, DeWitt III W S, et al. Deep generative models for T cell receptor protein sequences. Elife, 2019, 8.

de Souza N. The ENCODE project. Nature Methods, 2012, 9: 1046.

Deng P, Liu S, Nie X, et al. Conservation analysis of long non-coding RNAs in plants. Science China Life Sciences, 2018, 61(2): 190-198.

Deng P, Wu L. LncRNAs are cool regulators in cold exposure in plants. Science China Life Sciences, 2019, 62(7): 978-981.

Ding J, Regev A. Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces. BioRxiv, 2019: 853457.

Eser U, Churchman L S. FIDDLE: An integrative deep learning framework for functional genomic data inference. bioRxiv, 2016.

Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nature Reviews Genetics, 2014, 15(1): 7-21.

Ghandi M, Lee D, Mohammad-Noori M, et al. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Computational Biology, 2014, 10(7):e1003711.

Ghahramani A, Watt F M, Luscombe N M. Generative adversarial networks uncover epidermal regulators and predict single cell perturbations. bioRxiv, 2018: 262501.

Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in neural information processing systems. 2014: 2672-2680.

Grønbech C H, Vording M F, Timshel P N, et al. scVAE: Variational auto-encoders for single-cell gene expression data. bioRxiv, 2018: 318295.

Gunady M K, Kancherla J, Bravo H C, et al. scGAIN: Single Cell RNA-seq Data Imputation using Generative Adversarial Networks. bioRxiv, 2019: 837302.

Gupta A, Wang H, Ganapathiraju M. Learning structure in gene expression data using deep architectures, with an application to gene clustering. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2015, 1328-1335.

Gupta A, Zou J. Feedback GAN (FBGAN) for DNA: A novel feedback-loop architecture for optimizing protein functions. arXiv preprint arXiv:1804.01694, 2018.

Hassanzadeh H R, Wang M D. DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins. IEEE International Conference on Bioinformatics and Biomedicine(BIBM), 2016, 178-183.

Hill S T, Rachael K, Amy T, et al. A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. Nucleic Acids Research, 2018, 46(16): 8105-8113

Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 313, 2006, 504-507.

Hong H, Jiang S, Li H, et al. DeepHiC: A Generative Adversarial Network for Enhancing Hi-C Data Resolution. BioRxiv, 2019: 718148.

Huang Z A, Huang Y, You Z H, et al. Predicting lncRNA-miRNA interaction via graph convolution auto-encoder. Frontiers in genetics, 2019, 10: 758.

Hüttenhofer A, Schattner P, Polacek N. Non-coding RNAs: hope or hype?. Trends in Genetics, 2005, 21(5):289-297.

Isacchini G, Sethna Z, Elhanati Y, et al. On generative models of T-cell receptor sequences. arXiv preprint arXiv:1911.12279, 2019.

Jian Z, Theesfeld C L, Kevin Y, et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics, 2018, 50, 1171-1179.

Kelley D R, Snoek J, Rinn J L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 2016, 26(7): 990-999.

Kh. U R, Solovyev V V, Rogozin I B. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLOS ONE, 2017, 12(2): e0171410.

Khodabandelou G, Mozziconacci J, Routhier E. Genome functional annotation using deep convolutional neural network. bioRxiv, 2018: 330308.

Khoury G, Gruss P. Enhancer elements. Cell, 1983, 33(2): 313-314.

Killoran N, Lee L J, Delong A, et al. Generating and designing DNA with deep generative models. arXiv preprint arXiv:1712.06148, 2017.

Kingma D P, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Kiselev V Y, Andrews T S, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature

Reviews Genetics, 2019, 20(5): 273-282.
Kleftogiannis D, Kalnis P, Bajic V B. DEEP: a general computational framework for predicting enhancers. Nucleic Acids Research, 2015, 43(1):e6.
Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Handbook of Systemic Autoimmune Diseases, 2009, 1(4).
Krogel M A, Scheffer T. Multi-relational learning, text mining, and semi-supervised learning for functional genomics. Machine Learning, 2004, 57(1-2): 61-81.
Lanchantin J, Singh R, Wang B, et al. Deep Motif Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2016, 22:254-265.
LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444.
Lee B, Baek J, Park S, et al. deepTarget: End-to-end Learning Framework for microRNA Target Prediction using Deep Recurrent Neural Networks. The 7th ACM International Conference, 2016, 434-442.
Lee B, Lee T, Na B, et al. DNA-level splice junction prediction using deep recurrent neural networks. arXiv, 2015.
Leung M K K, Xiong H Y, Lee L J, et al. Deep learning of the tissue-regulated splicing code. Bioinformatics, 2014, 30(12): i121.
Li Y F, Shi W Q, Wasserman W W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics, 2018, 19(1):202.
Libbrecht M W, Noble W S. Machine learning applications in genetics and genomics. Nature Reviews Genetics, 2015, 16(6): 321-332.
Linder J, Bogard N, Rosenberg A B, et al. Deep exploration networks for rapid engineering of functional DNA sequences. bioRxiv, 2019: 864363.
Liu D, Xu C, He W, et al. AutoGenome: An AutoML Tool for Genomic Research. bioRxiv, 2019: 842526.
Liu F, Li H, Ren C, et al. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. Scientific Reports, 2016, 6(1): 28517.
Liu F, Ren C, Li H, et al. De novo Identification of replication-timing domains in the human genome by deep learning. Bioinformatics, 2016, 32(5):641-649.
Liu Q, Lv H, Jiang R. hicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics, 2019, 35(14): i99-i107.
Liu Y, Yao X. Ensemble learning via negative correlation. Neural networks, 1999, 12(10): 1399-1404.
Lu Y Y, Fan Y, Lv J, et al. DeepPINK: reproducible feature selection in deep neural networks. The 32nd Conference on Neural Information Processing Systems, Montréal, Canada, 2018, 1-11.
Mamoshina P, Vieira A, Putin E, et al. Applications of deep learning in biomedicine. Molecular pharmaceutics, 2016, 13(5): 1445-1454.
Margolin A A, Wang K, Lim W K, et al. Reverse engineering cellular networks. Nature protocols, 2006, 1(2): 662.
Min S, Lee B, Yoon S. Deep learning in bioinformatics. Briefings in Bioinformatics, 2017, 18(5): 851-869.
Marouf M, Machart P, Bansal V, et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. Nature Communications, 2020, 11(1): 1-12.
Min X, Chen N, Chen T, et al. DeepEnhancer: Predicting enhancers by convolutional neural networks. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2016.
Mohamed A, Dahl G E, Hinton G. Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20(1): 14-22.
Nielsen A A K, Voigt C A. Deep learning to predict the lab-of-origin of engineered DNA. Nature communications, 2018, 9(1): 3135.

Nissen J N, Sonderby C K, Armenteros J J A, et al. Binning microbial genomes using deep learning. BioRxiv, 2018: 490078.

Pan X, Shen H B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC Bioinformatics, 2017, 18(1):136.

Park S, Min S, Choi H, et al. deepMiRGene: Deep Neural Network based Precursor microRNA Prediction. bioRxiv, 2016.

Park Y, Kellis M. Deep learning for regulatory genomics. Nature Biotechnology, 2015, 33(8): 825.

Peng Y, Xiong D, Zhao L, et al. Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. Nature communications, 2019, 10(1): 1-11.

Qin Q, Feng J X. Imputation for transcription factor binding predictions based on deep learning. PLOS Computational Biology, 2017, 13(2):e1005403.

Quang D, Xie X H. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research, 2016, 44(11): e107.

Quang D, Xie X H. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods, 2019, 166:40-47

Repecka D, Jauniskis V, Karpus L, et al. Expanding functional protein sequence space using generative adversarial networks. bioRxiv, 2019: 789719.

Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082, 2014.

Ribeiro M T, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016, 1135-1144.

Ritchie M D, Holzinger E R, Li R, et al. Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics, 2015, 16(2): 85-97.

Sample P J, Wang B, Reid D W, et al. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. Nature Biotechnology, 2019, 37(7): 803.

Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Proceedings of the 34th International Conference on Machine Learning, 2017, 70, 3145-3153.

Sinai S, Kelsic E, Church G M, et al. Variational auto-encoding of protein sequences. arXiv preprint arXiv:1712.03346, 2017.

Singh R, Lanchantin J, Robins G, et al. DeepChrome: Deep-learning for predicting gene expression from histone modifications. Bioinformatics, 2016, 32(17): i639-i648.

Singh S, Yang Y, Poczos B, et al. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. bioRxiv, 2016: 085241.

Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 2017, 70, 3319-3328.

Tripathi R, Patel S, Kumari V, et al. DeepLNC, a long non-coding RNA prediction tool using deep neural network. Network Modeling Analysis in Health Informatics and Bioinformatics, 2016, 5(1):21.

Turner R. A model explanation system, IEEE International Workshop on Machine Learning for Signal Processing. 2016, 1-6.

Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine. Nature biotechnology, 2018, 36(9): 829.

Wang H, Cimen E, Singh N, et al. Deep learning for plant genomics and crop improvement. Current Opinion in Plant Biology, 2020, 54: 34-41.

Wang J, Qi Y. Plant non-coding RNAs and epigenetics. Science China Life Sciences, 2018, 61(2): 135-137.

Wang X, Ghasedi Dizaji K, Huang H. Conditional generative adversarial network for gene expression inference. Bioinformatics, 2018, 34(17): i603-i611.

Wang Y, Liu T, Xu D, et al. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Scientific Reports, 2016, 6:19598.

Wang Y, Wang H, Liu L, et al. Synthetic Promoter Design in Escherichia coli based on Generative Adversarial Network. BioRxiv, 2019: 563775.

Washburn J D, Mejia-Guerra M K, Ramstein G, et al. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proceedings of the National Academy of Sciences, 2019, 116(12), 5542-5549.

Webb S. Deep learning for biology. Nature, 2018, 554(7693).

Xie R, Wen J, Quitadamo A, Cheng J L, Shi X H. A deep auto-encoder model for gene expression prediction. BMC Genomics, 2017, 18(S9): 845.

Xu L, Hu Y, Cao Y, et al. An expression atlas of miRNAs in *Arabidopsis thaliana*. Science China Life Sciences, 2018, 61(2): 178-189

Yang B, Liu F, Ren C, et al. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics, 2017, 33(13): 1930-1936.

Yelmen B, Decelle A, Ongaro L, et al. Creating Artificial Human Genomes Using Generative Models. bioRxiv, 2019: 769091.

Yu N, Yu Z, Pan Y. A deep learning method for lincRNA identification using auto-encoder algorithm. IEEE International Conference on Computational Advances in Bio & Medical Sciences. IEEE, 2017.

Yue T, Wang H. Deep learning for genomics: A concise overview. arXiv, 2018.

Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. bioRxiv, 2019: 365007.

Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. European conference on computer vision. Springer, Cham, 2014: 818-833.

Zeng H, Edwards M D, Liu G, et al. Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics, 2016, 32(12):i121-i127.

Zhang S, Zhou J, Hu H, et al. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Research, 2015:gkv1025.

Zou J, Huss M, Abid A, et al. A primer on deep learning in genomics. Nature Genetics, 2019, 51(1): 12-18.

Zhou J, Park C Y, Theesfeld C L, et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nature Genetics, 2019, 51(6): 973.

Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning-based sequence model. Nature Methods, 2015, 12(10): 931-934.

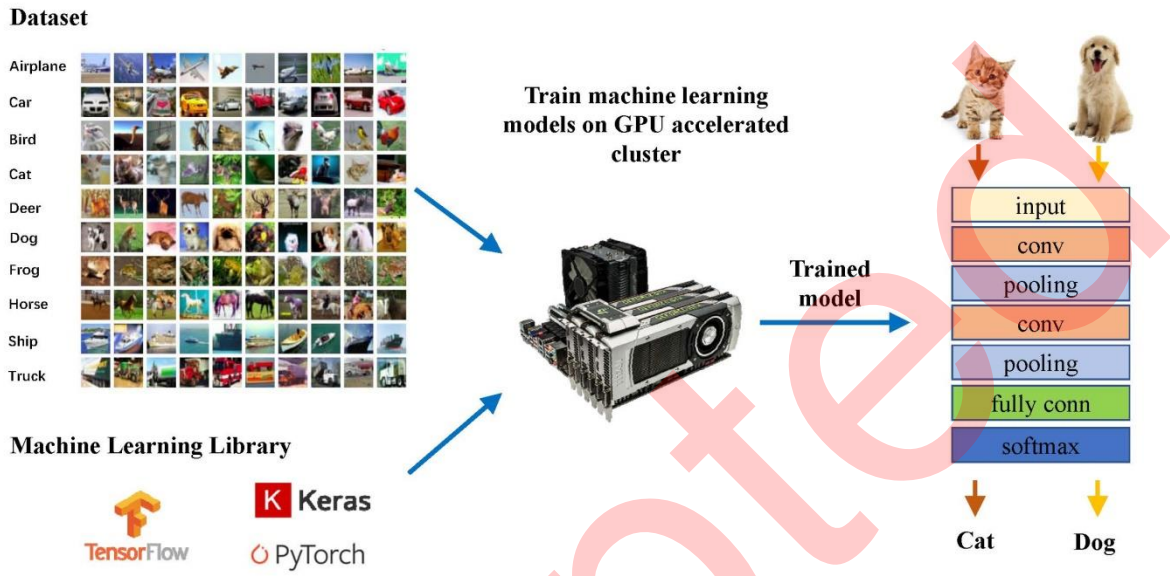

Figure 1 Overview of training machine learning models. Training efficient and robust machine learning models require large dataset with labels, Graphics Processing Unit (GPU) accelerated clusters and machine learning models provided in model libraries. After training a CNN with millions of images, the model can predict the category of unseen images accurately.

http://engine.scichina.com/doi/10.1007/s11427-020-1804-5

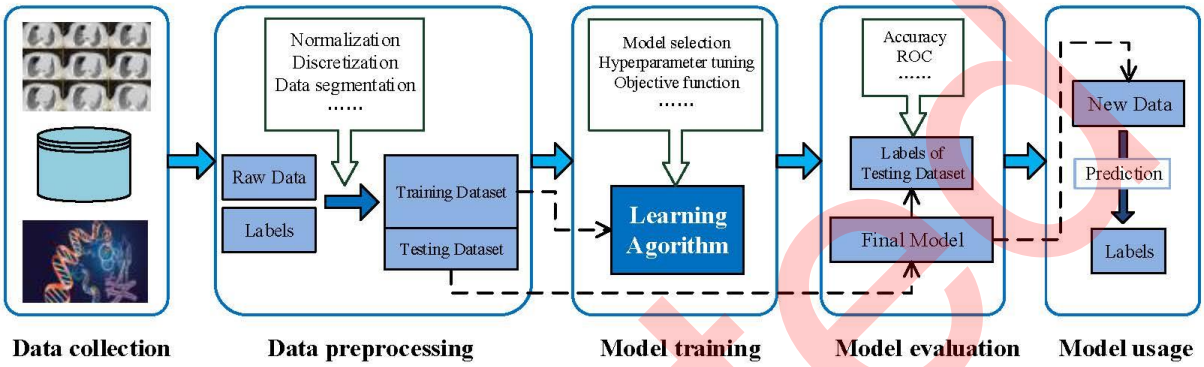

Figure 2 Specific process of machine learning. Initially, datasets are collected from different sources in various forms such as structured, unstructured or semi-structured data. Data preprocessing includes normalization, discretization, missing value filling, removing collinearity, training set and test set segmentation, data wrangling, etc. Model training is the core stage of machine learning, which includes model selection, objective function optimization, training stop condition setting, cross validation, hyperparameter tuning, etc. In model evaluation stage, test data set is used to evaluate model performance by measuring accuracy and drawing receiver operating characteristic (ROC) curve, etc. Then the trained model is employed to make predictions on new datasets. In addition, sometimes we would like to know how a model makes its predictions. In such a case, the importance of individual features, or interaction among features, is needed to explain a model's predictions.

http://engine.scichina.com/doi/10.1007/s11427-020-1804-5

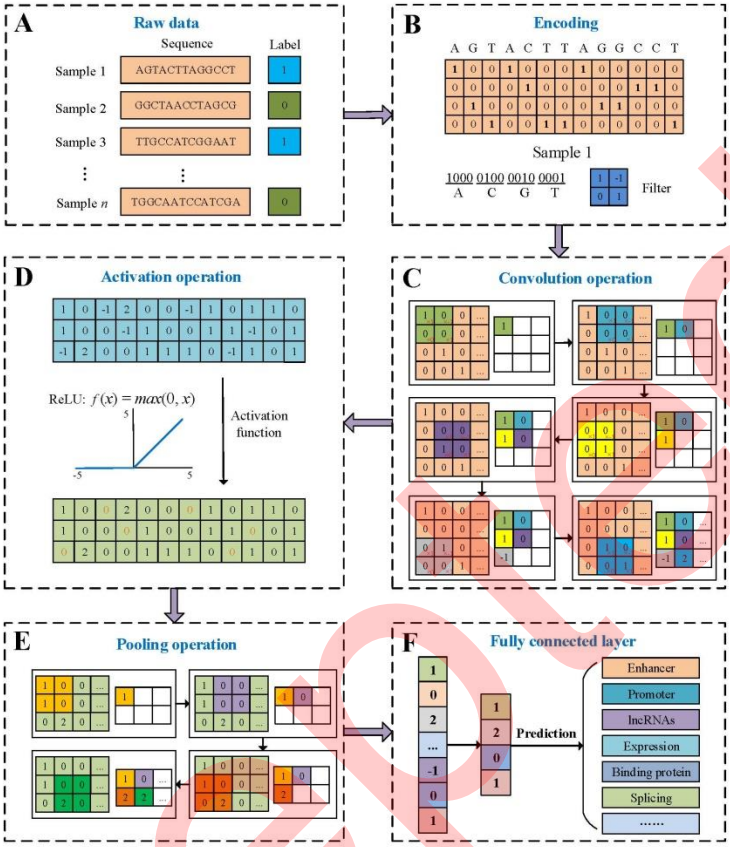

Figure 3 Application of CNN in genomics. (1) One-hot encoding is used to encode DNA sequence to matrix as the input of CNN. All filters (one filter example is shown in Figure 3(B)) are randomly initialized. (2) The encoded DNA is convoluted on the basis of initialized filters. The filter is multiplied by the corresponding input data through the sliding window, and the sum is computed and recorded (Figure 3(C)). The filters are adjustable parameters, often called weights, which are modified in the training process, as to improve the model performance. Sharing filters is one of the key ideas of CNN to reduce the number of connections between each layer and thus reduce the risk of overfitting. (3) The output of the convolution layer is mapped nonlinearly using activation functions. As shown in Figure 3(D), all negative values in the feature map are capped to zero using Rectified Linear Units (ReLU). (4) Based on the feature map obtained by convolution operation, the pooling operation is carried out to further filter the feature map. Generally, average pooling and max pooling are the two major pooling methods, with max pooling (Figure 3(E)) more widely used. The pooling layer is sandwiched in the middle of the convolution layers to reduce the data dimension, the number of parameters and the possibility of overfitting. (5) Multiple layers consisted of convolution and pooling operations are stacked with each layer representing the data in slightly more abstract form than the previous layer. After 10-20 convolutional and pooling layers, a fully connected layer is added as the output layer (Figure 3(F)). (6) Step (1) to (5) illustrate the feedforward pass in the training process. The feedforward pass outputs a prediction of the example. To increase the prediction accuracy, we first calculate the error (distance) between prediction and labeled category. To minimize the prediction error, back propagation is used to calculate the error gradient of all weights in the network. Specifically, stochastic gradient descent (SGD) is commonly used to update all filters to minimize the output error. Step (2)-(6) are repeated for all the input samples until the error stops decreasing. A test data set is then used to evaluate the generalization of the model, indicating whether the model can produce sensible predictions on data never seen before. The trained model could be used for various purposes, such as the predictors for enhancer, promoter, gene expression, interaction, etc.

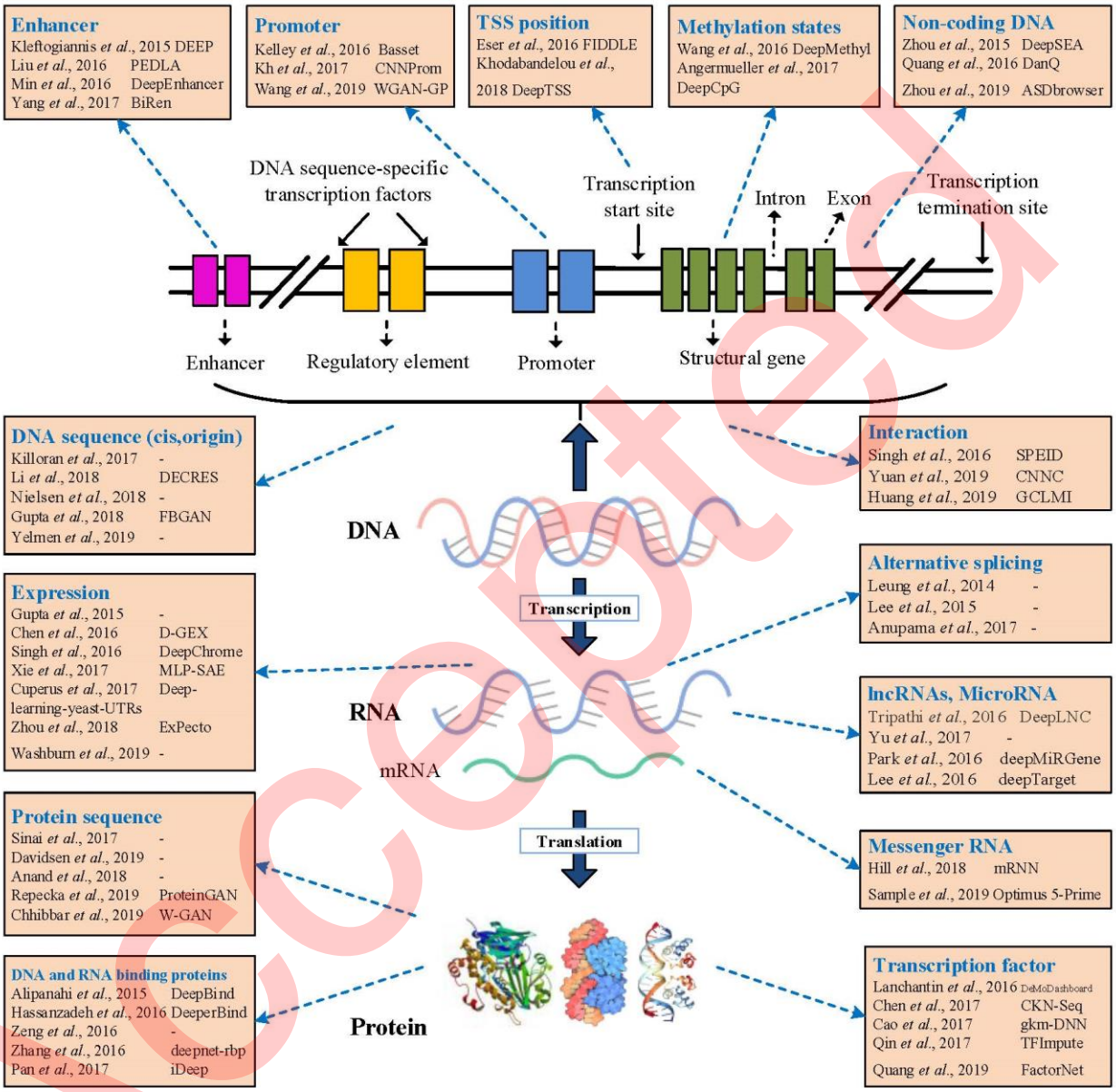

Figure 4. The applications of deep learning in genomics at the levels of DNA, RNA and protein. At the DNA level, deep learning has been applied in research related to enhancer, promoter, non-coding DNA, TSS position, methylation states, *cis*-regulatory, replication, and interaction. At the RNA level, deep learning has been used to study alternative splicing, lncRNA, MicroRNA, messenger RNA and expression. At the protein level, deep learning is used to study transcription factor, DNA binding proteins, RNA binding proteins, and protein sequence generation. GANs have also been applied to solve biological questions at different molecular levels.

http://engine.scichina.com/doi/10.1007/s11427-020-1804-5

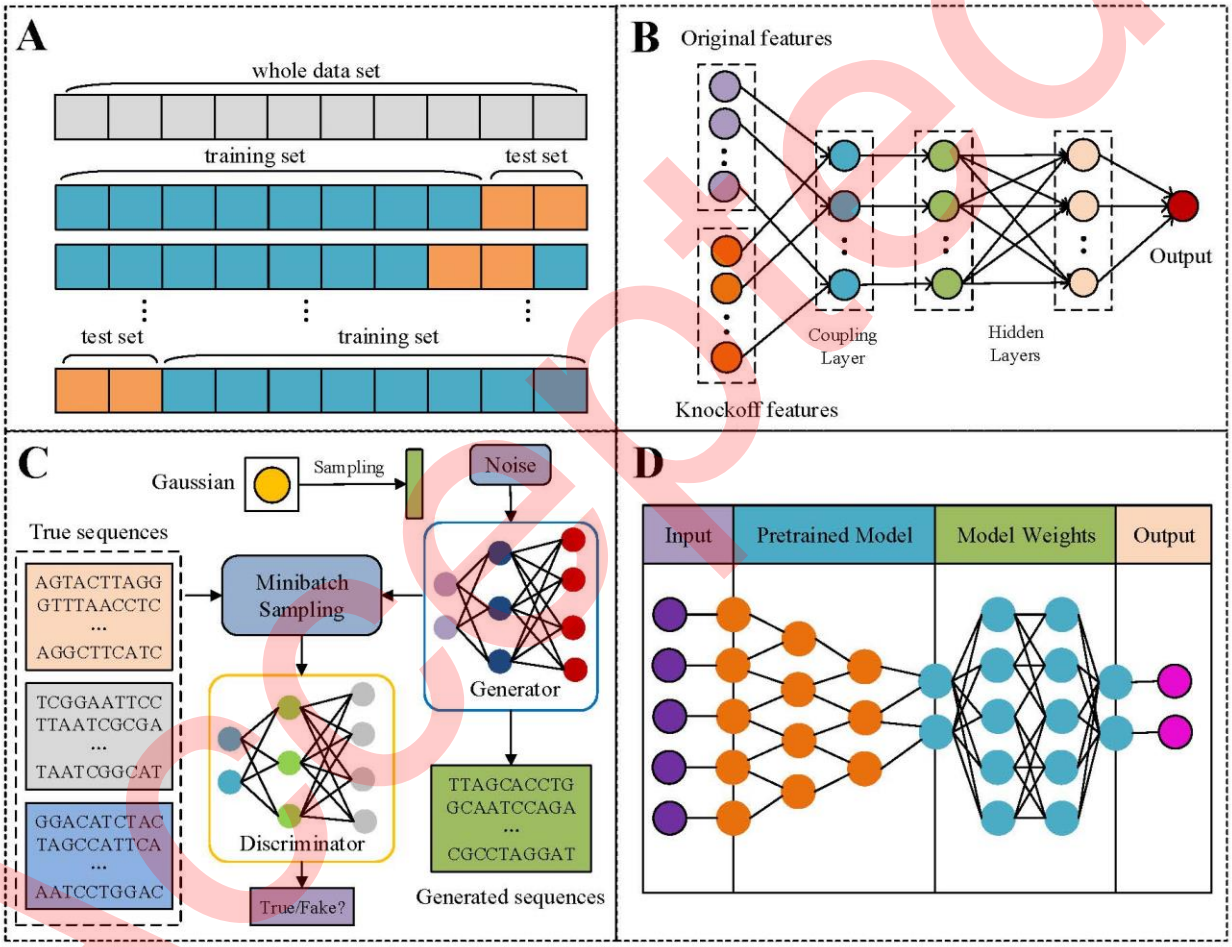

Figure 5. Caveats of using deep learning in genomics. (A) overfitting and underfitting. (B) training/test dataset splitting. (C) ensemble learning. (D) using knockoff features to open the black box of deep learning. (E) using generative adversarial networks to generate sequences. (F) process of transfer learning.

http://engine.scichina.com/doi/10.1007/s11427-020-1804-5

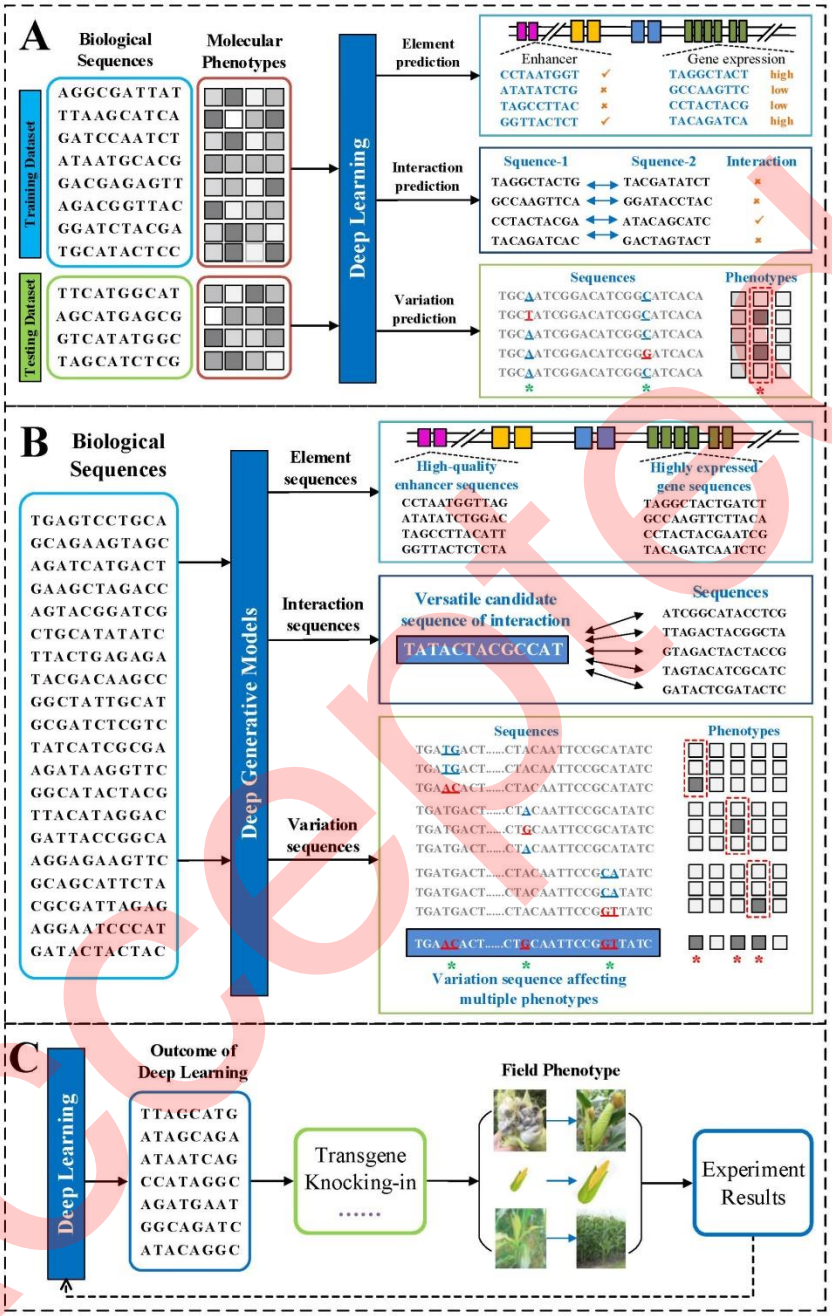

Figure 6. Deep learning in plant and animal breeding. (A) Three applications of deep learning in genomics. A deep learning model, once trained on training set and verified using testing set, can be used in various scenarios, including functional annotation of biological sequences such as prediction of gene-centric properties, prediction of the interactions among sequences, and prediction of phenotypic effects of natural variants. (B) Three applications of generative models in synthetic biology, including the generation of genomic elements with defined functions (such as enhancers or promoters), the generation of interacting sequences, and also generation of biological sequences conferring crops with superior agronomic traits. (C) Deep learning-guided crop genetic improvement. Biological sequences with desirable functions are transferred into crops by transgene or genome editing, in order to improve agronomic traits of crop species more efficiently. By this means, crop improvement becomes a designed process, and is no longer limited by natural variation.