Acta Scientific Veterinary Sciences (ISSN: 2582-3183)

Volume 5 Issue 7 July 2023

Review Article

Deep Learning in Neural Networks and their Application in Genomics

Kanaka KK[1], Nidhi Sukhija[1], Jayakumar Sivalingam[2]*, Rangasai Chandra Goli[1], Pallavi Rathi[1], Komal Jaglan[3] and Chethan Raj[4]

[1]ICAR-National Dairy Research Institute, Karnal, Haryana, India

[2]ICAR-Directorate of Poultry Research, Hyderabad, India

[3]Lala Lajpat Rai University of Veterinary and Animal Sciences, Hisar, Haryana, India

[4]ICAR-Indian Veterinary Research Institute, Izatnagar, UP

***Corresponding Author:** Jayakumar Sivalingam, ICAR- Directorate of Poultry Research, Hyderabad, India.

DOI: 10.31080/ASVS.2023.05.0683

Received: May 23, 2023

Published: June 07, 2023

Abstract

Deep learning has emerged as a powerful tool in genomics, utilizing neural networks to uncover complex patterns in large datasets. This review explores the application of deep learning in genomics, focusing on supervised and unsupervised learning tasks. The process involves training models with appropriate evaluation metrics and curated datasets to optimize performance. Balancing training data and model flexibility is crucial to avoid underfitting or overfitting. Deep learning models, with their high capacity and flexibility, outperform traditional techniques like logistic regression and support vector machines in genomics. Various applications of deep learning in genomics are includes predicting protein sequence specificity, determining cis-regulatory elements, analyzing splicing regulation and gene expression, and predicting genomic variants. Deep learning proves particularly effective in studying functional genomics and regulatory elements, leveraging techniques from computer vision and natural language processing. Overall, deep learning shows promise in advancing genomics research and understanding complex biological processes.

Keywords: Deep Learning; Supervised Model; Unsupervised Model; Genomics

Introduction

Neural networks, which are part of machine learning techniques have been extensively used in biology research [1]. A class of machine learning techniques known as “deep learning” is capable of finding extremely complex patterns in large datasets. Recent developments in applications ranging from computer vision to natural language processing have been impressive. A typical neural network consists of input layer to which we have to feed the data, the output of input layer is then utilized as input for hidden layers through networks and finally the analysis being transferred to output layer (Figure 1). The ultimate goal in many machine learning tasks is to optimize model performance not on the available data (training performance) but instead on independent datasets (generalization performance). With this goal, data are randomly split into at least three subsets: training (used for learning the model parameters), validation (used to select the best model) and test sets (to estimate the generalization performance) (Figure 3).

Machine learning tasks: supervised and unsupervised learning

Supervised learning

Getting a model that accepts features as input and outputs a prediction for a so-called target variable is the aim of supervised learning. Predicting whether or not an intron is spliced out (the target) given characteristics of the RNA, such as the presence or ab-

sence of the canonical splice site sequence, the location of the splicing branch point, or intron length, is an example of a supervised learning problem. When a machine learning model is trained, its parameters are learned. This process typically entails minimizing a loss function on training data in order to make precise predictions about unobserved data [2] (Figure 1).

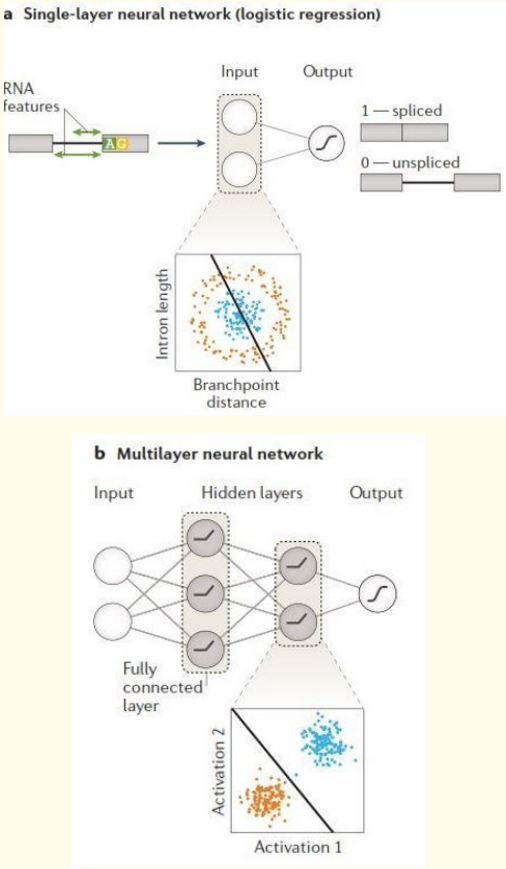

Figure 1: Supervised learning; neural networks with hidden layers used to model presence or absence of splice sites (Figure adopted from Park and Kellis, 2015 [6]).

a) Shown is an example of splice site classification based on two RNA features. It predicts the probability of the output being class 1 using a weighted sum of the input that is mapped to the (0, 1) interval with a sigmoid function.

b) Neural networks with intermediate layers, also called hidden layers, transform the inputs using intermediate nonlinear transformations into a space where the classes become linearly separable.

Unsupervised learning

Getting a model that accepts features as input and outputs a prediction for a so-called target variable is the aim of supervised learning. Predicting whether or not an intron is spliced out (the target) given characteristics of the RNA, such as the presence or absence of the canonical splice site sequence, the location of the splicing branch point, or intron length, is an example of a supervised learning problem. When a machine learning model is trained, its parameters are learned. This process typically entails minimising a loss function on training data in order to make precise predictions about unobserved data [2] (Figure 1).

The encoder and the decoder are the two components of an autoencoder. In the so-called bottleneck layer, the encoder reduces the dimensions of the input data. The compressed data in the bottleneck layer is used by the decoder to attempt to reconstruct the original input. The loss function between the original data and the reconstructed data measures the accuracy of the reconstruction. Although pseudo time estimation is not a feature of autoencoders, the reconstruction's denoising effect can help the data's underlying structure become more apparent [4] (Figure 2).

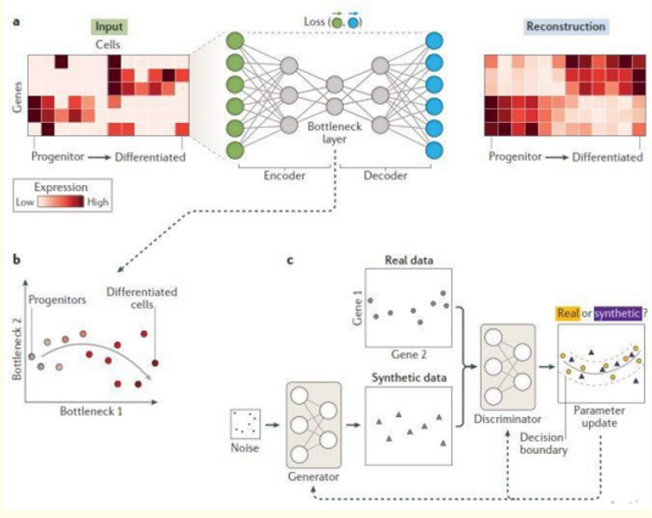

Figure 2: Unsupervised learning in a gene expression study (Figure adopted from Park and Kellis, 2015 [6]).

a) An autoencoder and its two parts: an encoder and a decoder.

b) The bottleneck layer is a low-dimensional representation of the original input revealing the cell differentiation process.

c) Generative adversarial networks consist of generator and discriminator neural networks that are trained jointly. The discriminator classifies whether a given data point was drawn from the real data (circles) or whether it was synthetically generated (triangles). The generator aims to generate realistic samples and thereby tries to deceive the discriminator into mistakenly classifying synthetic samples as real.

Points to consider while developing valid models

The amount of training data and model flexibility need to be balanced properly in machine learning. A model that is too simple won't fit the data well and won't be able to reveal data patterns. Unable to generalise, an overly flexible model will overfit fictitious patterns in the training data. A class of machine learning algorithms known as large neural networks, which constitute one of the main types of deep learning, are capable of dimensionality reduction and prediction. Deep learning models have a higher capacity and are much more flexible than traditional machine learning techniques used in genomics, such as logistic regression and support vector machines. Millions of trainable parameters are used in typical deep learning models. Deep learning can automatically learn features and patterns with less skilled handcrafting when given properly curated training data. Additionally, training in and interpreting the underlying biology calls for more caution [5].

Following knowledge is necessary in building an effective deep learning model:

- Will be able to choose an appropriate evaluation metric and curate an appropriate training dataset.
- Confounding biases that might artificially inflate performance should be avoided when building the training set. For instance, known neutral genetic variants may be more evenly distributed throughout the genome, whereas known pathogenic genetic variants may cluster in specific regions of the genome, such as exons or promoters.
- In reality, a neural network trained on these unbalanced data would probably learn to identify regions of the genome enriched in pathogenic variants without being able to distinguish between neutral and pathogenic variants within these crucial genomic regions. This neural network would appear to perform well.
- Thus, it is important to design training datasets that are appropriately balanced for confounders that would detrimentally affect performance when applied to real-world-use cases [6].
- As with all other machine learning techniques, deep learning requires domain expertise for successful application. For instance, domain knowledge is used to build features from data in logistic regression and support-vector-machine classification, and expertise is incorporated into the prior distributions in Bayesian models.
- Domain knowledge is incorporated into the network architecture design in deep learning. Understanding the underlying premise and constraints of various architectures is essential for network performance [7].

Deep learning workflow in genomics:

- **Step (a):** Divide the dataset into training, validation, and test sets at random. To ensure that the predictor learns salient features rather than confounders, the positive and negative examples should be balanced for potential confounders (such as sequence content and location) (a in Figure 3).
- **Step (b):** Using domain expertise, the appropriate architecture is chosen and trained. For instance, RNNs capture more flexible spatial interactions, while CNNs capture translation invariance (b in Figure 3).
- **Step (c):** We evaluate the rates of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Precision and recall are frequently taken into consideration when there are more negative than positive examples (c in Figure 3).
- **Step (d):** By calculating how each nucleotide in the input affects the prediction, the learned model is understood (d in Figure 3).

Applications in genomics

Deep learning methods and tools for studying the genome are being presented in an increasing number of publications. Genomics offers state of art techniques to identify, measure, interpret diversity which is needed for abiotic stress management in keeping view of climatic change [14]. Deep learning's leading application area is functional genomics [9]. Here are a few examples in brief: 1) Predicting the sequence specificity of proteins that bind to DNA and RNA predicting the regulators and enhancers. 2) Determining cis-regulatory elements and regions measuring the level of methylation. 3) Examining splicing regulation and gene expression predicting the start sites for transcription. 4) Genomics variants (SNPs, CNVs) prediction. These tools are based on information obtained from DNase I sequencing, the ATAC-seq assay for transposase-accessible chromatin, chromatin immunoprecipitation (ChIP) sequencing, ChIP-on-chip, RNA immunoprecipitation sequencing, transcription factor datasets, and chromatin state [10].

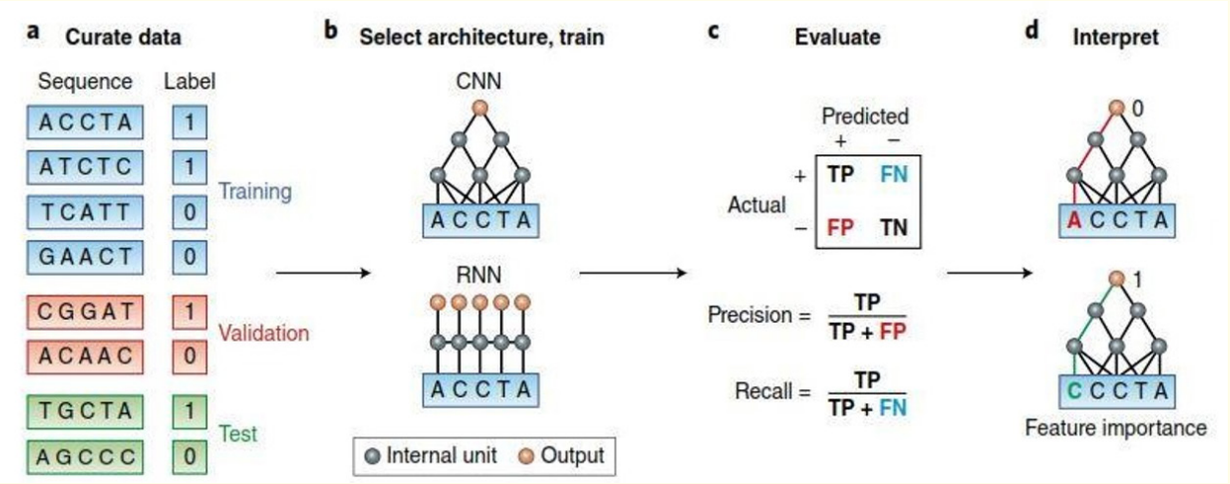

Figure 3: Deep learning workflow in genomics (Figure adopted from Zou., et al. 2018 [8]).

DNA sequence patterns and three-dimensional genome topology (based on Hi-C) have been used to infer the methylation state of DNA, which also affects gene expression. Some of the real world examples have been discussed as follows: Kanaka et al. (2021) [15].

It will also help to choose candidate promoters for their incorporation in transgenic cassettes [16]. In the analysis of sequence data to identify SNPs and CNVs in Indian cattle and Buffaloes many tools and software have been used and they were built using ANN and CNN [17,18] and also for finding significant association of SNPs with phenotypes using GWAS [19,20]. Also these tools are helpful for finding epitopes [21] and to design vaccines against food allergies and pathogens [22]. At 1-kb resolution, Hi-C contacts are predicted by nucleotide sequence and DNase I assay signals [10].

Using architectures directly adapted from contemporary computer vision and natural language processing applications, deep learning has been particularly effective when applied to regulatory genomics. Although modifications to the deep learning architecture tailored to the needs of genomics can be helpful, the majority of methods use CNN or RNN, which are excellent for the tasks of modeling regulatory elements [11]. The forward- and reverse-complement versions of the same DNA sequence can yield different predictions using conventional deep learning models that do not explicitly model this property. With the aim of calculating how much RNA is produced from a DNA template in a specific cell or condition, various tools are capable of extracting transcriptome patterns from large datasets of gene expression data [12]. Deep learning can be used to study the splicing-code model, identify long non-coding RNAs, and create predictive models of gene expression from genotype data. As a final application, deep learning has been used to interpret regulatory control in single cells, such as the detection of DNA methylation in single cells3 and the identification of cell subgroups by enhancing the representation of single-cell RNA-seq data [13]. Table 1 lists resources pertinent to deep learning in general.

Resource type	Name	URL	Comment
Cloudplatform	Amazon EC2	https://aws.amazon.com/ec2/	Most popular cloud platform
	Microsoft Azure	https://azure.microsoft.com/	Second-largest cloud platform
Plug-and-play cloud	FloydHub	https://www.floydhub.com/	All startups in the GPU service space; pay-by-the-hour model on top of basic monthly subscriptions
GPU services	Google CloudML	https://cloud.google.com/ml-engine/	Can run your own models on Google's hardware, including tensor processor units Google
	Colaboratory	https://colab.research.google.com/	Notebook environment with free GPUs (during 12 h)

Design services fordeep learning models	Fabrik	https://github.com/Cloud-CV/Fabrik/	Model export to Keras code; no training
	IBM Data Cloud	https://datascience.ibm.com/	Model export to Keras, PyTorch, TensorFlow or Caffe
	Deep Cognition	http://deepcognition.ai/	Training and evaluation included
Prebuilt images with CUDA support	Docker Hub	https://hub.docker.com/r/nvidia/cuda/	Docker images from NVIDIA with CUDA/cuDNN GPU support
	AmazonDeep LearningAMIs	https://aws.amazon.com/machine-learning/amis/	Amazon Machine Images (AMIs)with GPU support
Softwarelibraries (general)	Keras	https://keras.io/	More high-level than TensorFlow (below) but can be integrated withit in many ways
	TensorFlow	https://www.tensorflow.org/	Developed by Google; most popular deep learning framework
	PyTorch	http://pytorch.org/	Developed by Facebook
Softwarelibraries	DragoNN	https://kundajelab.github.io/dragonn/	specific for genomics
	Kipoi	http://kipoi.org/	Model zoo for deep learning ingenomics
Educationalresources	fast.ai	http://www.fast.ai/	E.g., Deep Learning for Coders 1 and 2
	Coursera	https://www.coursera.org/specializations/deep-learning/	Deep-learning-specialization coursepackage
	Textbook	http://neuralnetworksanddeeplearning.com/	Free online textbook with examplecode

Table 1: Resources relevant for deep learning in general.

Conclusion

Deep learning techniques have revolutionized the field of genomics by enabling the analysis of large and complex datasets. The application of deep learning in genomics has led to advancements in functional genomics, including protein-DNA and protein-RNA interaction prediction, cis-regulatory element identification, splicing regulation analysis, and genomics variants prediction. With the adaptation of computer vision and natural language processing architectures, deep learning has proven particularly effective in regulatory genomics. Overall, deep learning holds great promise for uncovering insights into the complex biological mechanisms underlying the genome.

Bibliography

1. Telenti A., *et al*. "Deep learning of genomic variation and regulatory network data". *Human Molecular Genetics* 27 (2018): R63-R71.

2. Libbrecht M W and Noble W S. "Machine learning applications in genetics and genomics". *Nature Reviews Genetics* 16 (2015): 321-332.

3. Vincent P., *et al*. "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion". *Journal of Machine Learning Research* 11 (2010): 3371-3408.

4. Eraslan G., *et al*. "Single-cell RNA-seq denoising using a deep count autoencoder". *Nature Communication* 10 (2019): 390.

5. Khodabandelou G., *et al*. "Genome Functional Annotation using Deep Convolutional Neural Networks". *bioRxiv* 330308 (2020).

6. Park Y and Kellis M. "Deep learning for regulatory genomics". *Nature Biotechnology* 33 (2015): 825-826.

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

8. Zou J., *et al*. "A primer on deep learning in genomics". *Nature Genetics* 51.1 (2019): 12-18.

9. Yue T and Wang H. "Deep learning for genomics: A concise overview". *arXiv preprint arXiv:1802.00810* (2018).

10. Kelley DR., *et al*. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks". *Genome Research* 26 (2016): 990-999.

11. Zeng W., *et al*. "Prediction of enhancer-promoter interactions via natural language processing". *BMC Genomics* 19 (2018): 84.

12. Shaham U., *et al*. "Removal of batch effects using distribution-matching residual networks". *Bioinformatics* 33 (2017): 2539-2546.

13. Angermueller C., *et al*. "Deep CpG: accurate prediction of single-cell DNA methylation states using deep learning". *Genome Biology* 18 (2017): 67.

14. Kanaka, K. K., et. al. On the concepts and measures of diversity in the genomics era. Current Plant Biology, 33 (2023):100278.

15. Kanaka, K. K., et. al. Cloning, characterisation and expression of the SERPINB14 gene, and association of promoter polymorphisms with egg quality traits in layer chicken. British Poultry Science, 62(6) (2021): 783-794.

16. Kanaka, K. K., et. al. Development of protocol for production of primary antibody against ovalbumin protein in chicken for detection of the protein through western blotting. Journal of Animal Research, 9(6) (2019): 849-853.

17. Murugesan, K. D., et. al. Profiling and integrated analysis of whole-transcriptome changes in uterine caruncles of pregnant and non-pregnant buffaloes. Genomics, 113(4)(2021): 2338-2349.

18. Jaglan, K., et. al. Impact of non-genetic factors on clinical mastitis incidence in Murrah buffaloes (2022).

19. Jaglan, K., et. al. Genomic clues of association between clinical mastitis and SNPs identified by ddRAD sequencing in Murrah buffaloes. Animal Biotechnology (2023):1-9.

20. GEORGE, L., et. al. Genetic improvement of economic traits in Murrah buffalo using significant SNPs from genome wide association study (2023).

21. Kanaka KK.,et al. "Network analysis of beta casomorphin-7 revealed genes and pathways associated with human diseases". Acta Scientific Veterinary Sciences 5.3 (2023): 31-46.

22. Nidhi Sukhija., et al. "Epitope tailored vaccine construct for goat milk allergy". Acta Scientific Veterinary Sciences 5.3 (2023): 23-30.