Artificial intelligence for Next generation sequencing data analysis

Martina Elena Tarozzi, indipendent researcher. Florence, Tuscany, Italy

https://doi.org/10.57098/SciRevs.Biology.3.1.2
Received March 22, 2024. Revised April 09, 2024. Accepted April 10, 2024.

Abstract: In the rapidly evolving field of genomics, our capacity to decipher genetic data encoded in DNA has been transformed by Next Generation Sequencing (NGS) technologies. These advanced technologies produce an enormous volume of data, posing substantial challenges in extracting meaningful biological insights. Artificial intelligence (AI) algorithms offer distinctive possibilities to unravel the biological information embedded in such extensive and intricate datasets. This review offers a synopsis of AI classifications and algorithms, elucidating how these techniques can be employed on sequencing data. Subsequently, a selection of the most typical, promising, or illustrative applications of AI on NGS data to tackle unresolved technical or biological issues are showcased.

Introduction

Artificial intelligence (AI) algorithms and sequencing technologies represent two groundbreaking innovations that witnessed outstanding improvements in the last few decades. In both AI and sequencing technologies, the first ilestones date back to the early '50s, and the rapid and simultaneous advancements in the two fields resulted in the new hybrid research branch of computational biology[1]. The large and complex datasets produced by sequencing experiments contain the information needed to understand many unanswered biological and medical questions, but that information is often difficult to extract. As the biomedical sector is becoming more data-intensive and AI algorithms more able to handle biological complexity, the interconnection between these two research fields is bound to strengthen.

Overview of artificial intelligence, machine learning and deep learning

Artificial Intelligence (AI) is a broad term that covers a plethora of computational approaches and algorithms able to mimic cognitive abilities (Figure 1). The term Machine Learning (ML) refers to several algorithms able to perform different tasks, such as pattern recognition, classification and prediction tasks based on models derived from existing data. The key feature of these methods is the absence of coded algorithms given by the developer to describe the steps towards which input data are transformed in output results. The ML method therefore learns from the data to create a hierarchy of concepts, each one defined by its relation to simpler concepts, that it uses to perform a task. By building knowledge from previous data and training, this approach avoids the need for the developer to explicitly define all the knowledge that the machine needs. The goal of many ML tasks is to optimize the model performance so that they can be generalized on independent datasets (generalization performance).

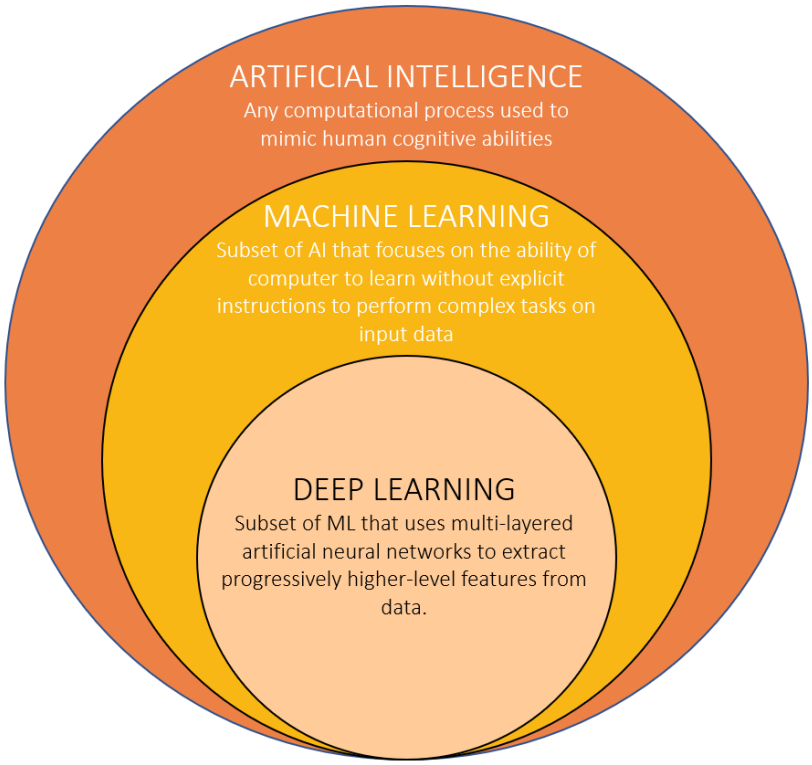

Figure 1: *Schematic representation of differences and shared features between Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL).*

ML methods are usually defined either as supervised or unsupervised. In unsupervised learning, no predefined labels are provided for the objects under study. Here, the goal is to explore the data and discover similarities between objects (e.g., samples) based solely on the input data. Clustering and most dimensionality reduction techniques represent examples of unsupervised algorithms. Unsupervised methods are commonly used in exploratory data analysis and for quality control tasks to identify potential issues such as outliers and batch effects, as well as to discover recurrent patterns and unknown sources of variation in highly dimensional datasets [2,3]. Examples of unsupervised dimensionality reduction techniques are principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE)[4] and Uniform Manifold Approximation and Projection (UMAP) [5]. PCA is a fast, linear transformation, which maps the data over a space whose coordinates are linear combinations of input features that capture most of the variance of the data. This algorithm is usually preferred when the aim is to separate the data points as far as possible. On the other hand, methods such as UMAP or t-SNE rely on more complex mathematical assumptions and steps, to "guess" the manifold on which data are located. Such methods preserve only local similarities and hence produce a higher clusterization of the data in the embedding space, with data subpopulations more separated among themselves compared to PCA embeddings [6,7]. Clustering methods are used to identify in an unsupervised way groups of similar data points based on the measure of similarity of choice. Examples of commonly used clustering algorithms in sequencing data analysis are hierarchical clustering with dendrograms, K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [8].

On the other hand, supervised learning involves using labeled data, where each input has an associated output. This allows the algorithm to learn a set of rules to predict the correct output for new input data based on its features, attributes, or labels. If the output is qualitative, then the process is called classification [9,10]. Otherwise, in the case of quantitative values, it is called regression. Well-trained ML models can learn rules about the underlying patterns and relationships in the data that can then be applied to make accurate predictions on

new, unseen data. These rules can display new insights into the relevant features used to correctly identify the studied classes[11]. Examples of classifiers are decision trees, random forests, support vector machines and neural networks. Examples of regression methods are linear or generalized linear models, with the possible addition of random effects (mixed-effect models), penalization terms (regularized models), non-linearity (e.g., kernel regression), or basis functions (e.g., splines). Nevertheless, it is worth underlying that supervised and unsupervised learning are not formally defined terms, and the boundaries between them are often blurred since many ML methods can be used to perform both types of tasks[12].

Neural networks are ML algorithms consisting of interconnected artificial neurons, which represent the building blocks of neural networks and deep learning algorithms[13]. Artificial neurons have as input a vector of values and compute weighted sums of these values followed by a non-linear transformation. The activation function defines whether the input values reach the threshold needed to activate the neurons and consequently determines the output of the node given the set of inputs received. The weights are parameters that are adjusted during the training step, as learning proceeds[11]. The term Deep Learning (DL) refers to multi-layered artificial neural networks with a high number of hidden layers used to extract progressively higher-level features from data. The first layer of neurons, also referred to as the input layer, is the one that receives the experimental input data, followed by layer two, made of neurons that receive as input the outputs of layer one, and so forth for deeper hidden layers[13]. The goal of these networks is to model a function f. In a classifier, for example, the function $y = f(x)$ maps an input x to a category y. A feed-forward network defines a mapping $y = f(x; \theta)$ and learns from the training set the value of the parameters θ to model the function. Training a neural network involves optimizing its parameters (weights) to minimize a certain error metric, which is typically defined by a loss function. Such loss function measures the difference between the predicted output of the neural network and the actual output. Optimal weight updates are enabled by backpropagation, which is a well-established algorithm in neural network training, as it enables the computation of the loss function's gradient with respect to the network's weights.

Neural networks can be categorized into three main types based on the type of architecture in which neurons are organized: feed-forward, recurrent, and convolutional (Figure 2). In feed-forward networks, the connections between nodes do not form a cycle, and the information proceeds exclusively forward from the input layer to the hidden layers and lastly to the output layer. On the contrary, recurrent neural networks have connections that form cycles, allowing the output of a node to affect subsequent input to the same nodes. Convolutional neural networks are made of convolution kernels processing input data, followed by pooling layers simplifying the information to its most meaningful concepts, and ultimately followed by hidden fully connected layers for further data processing, like for example a prediction task. The ultimate output of the neural network represents its prediction or classification of the input data, which is built based on its experience of recurrent patterns learned from the data.

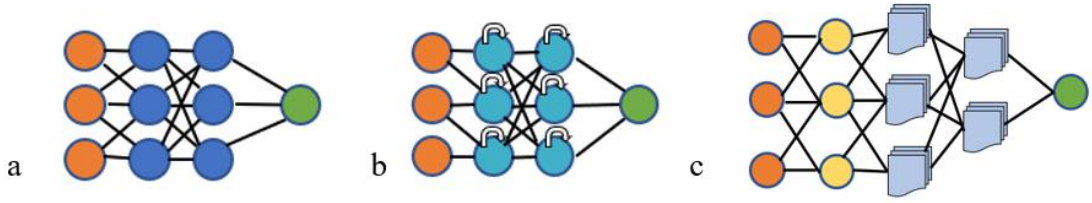

Figure 2: *Schematic representation of the three main types of neural networks: feed-forward neural network (a), recurrent neural networks (b) and convolutional neural networks (c). In all three representations, orange circles stand for the input neurons and green circles for the output neurons. In panel a, blue circles represent the hidden layers. In panel b, light blue circles indicate recurrent hidden layers, while in panel c, yellow circles indicate the kernels and the gray 3D squares represent convolutional layers.*

Applications of AI in genomics and transcriptomics

The complexity of data generated by high-throughput sequencing technologies can make traditional analysis methods insufficient for identifying patterns and extracting insights.ML and DL methods have been applied to sequencing data with a vast number of scopes. Here we provide a selection of some of the most relevant fields of application. This section aims at providing common, promising or exemplifying applications of AI methods on NGS data in biology and bioinformatics, while it should not be considered a complete overview of all its possible applications in biology.

Liquid biopsies and personalized medicine

Liquid biopsies are minimally invasive diagnostic methods that analyze bodily fluids, such as blood, urine, or cerebrospinal fluid, to detect and monitor diseases, and are especially relevant in early diagnosis of cancer and neurodegenerative diseases [14,15]. These samples contain cell-free Nucleic Acids (cfNA), circulating tumor DNA (ctDNA), circulating tumor cells (CTCs), exosomes, and other biomarkers that allow the extraction of genomic, transcriptomic and epigenomic information, which can be used for early detection, monitoring of progression and support personalized therapeutic decisions to target the disease [15]. These types of data are extremely complex, subject to many confounders and for most features characterized by a high signal-to-noise ratio. AI algorithms have significantly advanced data analysis and interpretation of this data and consequently the whole field in several aspects, such as in risk assessment and early diagnosis [16], disease subtype classification [17], treatment response prediction [18] and in monitoring minimal residual disease [19]. For example, SVM were effectively used to predict the probability of reoccurrence based on gene expression data or specific gene signature in different types of cancers[20,21], improving the monitoring of the molecular profile of the patient's tumor and the prediction of personalized treatments at different times. Furthermore, ctDNA methylation patterns have been extensively studied with several ML classification or regression methods as well as with neural networks to achieve effective early detection both in cancer research[22] and in the context of neurodegenerative diseases[23]. In this context, AI reaches some of the most notable results in terms of tangible impacts in molecular biology and medicine, and it is expected that its role in personalized medicine will increase in the near future.

Regulatory genomics

Regulatory genomics is the field of genomics that studies gene expression regulation trying to identify regulatory regions (such as enhancers, promoters, transcription start sites (TSS), and genome accessibility) and the regulatory hierarchy between these regions and other genes. In this context, deep learning and more specifically Convolutional Neural Networks have been applied with the best results. One of the commonly used architectures involves treating the input DNA sequence as categorical variables. Each position in the sequence is one-hot encoded, resulting in a vector where only one channel corresponds to the A-C-G-T nucleotides (with a value of 1) provided to the input layer. These kernels are followed by convolutional layers, which simplify the information to extract the most relevant concepts. Convolutional filters are initially trained on specific regions of interest with known regulatory properties. The knowledge gained by the convolutional neural network (CNN) during training can then be applied to new regions for accurate predictions. This architecture has been successfully applied to various types of sequencing data, particularly in the context of epigenomic studies. This overall architecture has been used on different types of sequencing data and has provided particularly interesting results in terms of epigenomic studies. For example, this type of architecture has been applied to DNAase-seq data to predict cell-type specific regions of accessible chromatin [24], to identify promoters and distal regulatory regions along mammalian genomes [25], to predict cell-type specific gene expression from DNA sequencing data and alterations of it associated to variant alleles [26], and to identify genomic regions responsible for the three-dimensional chromatin folding in the nucleus [27] from genomic and Hi-C data. Considering that both the experimental and computational technologies used in these studies are relatively young, this is arguably one of the most promising research fields for the next decades, with the potential to answer many of the open questions in functional genomics.

Improvement of genome editing specificity

During the past decade, technological innovations in molecular biology have made genome editing easier, allowing the modification of the DNA sequence at a single nucleotide resolution[28]. The most successful technique to perform genome editing is CRISPR/cas9, where the identification of the target genomic region is mediated by a guide RNA (gRNA). The gRNA is a chimeric RNA consisting of a ~20nt guide sequence that identifies through base complementarity the target site in the genome and precisely directs the Cas9 protein to it. Some mismatches in the guide sequence can be tolerated and do not affect the ability to align and cut DNA, resulting in off-target cleavages [29,30]. Accurate gRNA design maximizes on-target efficacy (sensitivity) and minimizes off-target effects (specificity). ML and DL models have been used in this context to predict gRNA sequencing with high sensitivity and specificity, and several specific tools have been released in the last few years, for example, DeepSpCas9 [31], DeepCRISPR [32] , DeepCpf1 [33], CRISPRscan [34], among many other. These tools differ in terms of models and network architectures, nevertheless, the fundamental overall architecture described in the previous section still applies also in this context. Of course, other more complex networks exist, like the one used by DeepCRISPR[32], where a hybrid deep neural network is used combining unsupervised and supervised representation learning to model single gRNAs using a set of genome-wide sgRNAs. Despite the outstanding results already achieved, this type of application of ML and DL on sequencing data still has some relevant limitations [35]. In this context, the large amount of data needed to train prediction models is not always available, and quite often the available data show some challenges caused by the heterogeneity of sequencing platforms and cell types.

Future Perspectives

Both computational and molecular biology are continuing to grow at an impressive pace, as they did in the last decades. As shown in this review, integrating AI, ML and DL techniques with NGS data holds immense promise, but it also presents challenges. For example, the quality and quantity of NGS data pose hurdles— e.g. noise, biases, and artifacts- that can impact analyses and computational resources and strain existing infrastructure due to DL model demands. In addition, these models require a huge amount of data which is not always available, therefore one of the currently most relevant challenges in the field is addressing small sample sizes and class imbalance affects model robustness. Another limitation is that interpretability remains an issue: black-box models lack transparency and the understanding of the molecular processes underneath a given prediction or classification based on nucleotides pattern often remains under understood. Bridging the gap between AI expertise and biological domain knowledge is critical and could improve transferability and generalization across diverse biological contexts as well. Last but not least, ethical concerns arise from handling sensitive genomic data. Finding a balance between the extraordinary potential of AI in the medical field and the need for a careful safeguard of individual privacy are really hot topic involving both law and tech experts which currently remains an open issue. In summary, while AI, ML, and DL offer exciting prospects for genomics research, overcoming these challenges is crucial for realizing their full potential. Taken together these innovations are a new frontier of research and have the potential to strongly affect our possibilities to understand and interact with genetic information, which will also require ethical considerations.

References

1. Muir P, Li S, Lou S, et al. The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biol. 2016; 17:1–9

2. Jia W, Sun M, Lian J, et al. Feature dimensionality reduction: a review. Complex Intell. Syst. 2022 83 2022; 8:2663–2693

3. Meng C, Zeleznik OA, Thallinger GG, et al. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. 2016; 17:628–641

4. Hinton G; L van der M. Visualizing Data using t-SNE. Ann. Oper. Res. 2014; 219:187–202

5. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018;

6. Yang Y, Sun H, Zhang Y, et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Rep. 2021; 36:

7. Li W, Cerise JE, Yang Y, et al. Application of t-SNE to human genetic data. J. Bioinform. Comput. Biol. 2017; 15:1–14

8. Jin J, Wang H, Shu Z, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Cailiao Yanjiu Xuebao/Chinese J. Mater. Res. 2017; 31:219–225

9. Xu C, Jackson SA. Machine learning and complex biological data. Genome Biol. 2019; 20:76

10. Bzdok D, Krzywinski M, Altman N. Points of significance: Machine learning: Supervised methods. Nat. Methods 2018; 15:5–6

11. Murphy KP. Machine Learning A Probabilistic Perspective. MIT Press Cambridge, Massachusetts London, Engl. 2018; 16:

12. Omta WA, Heesbeen RG van, Shen I, et al. Combining Supervised and Unsupervised Machine Learning Methods for Phenotypic Functional Genomics Screening: https://doi.org/10.1177/2472555220919345 2020; 25:655–664

13. Zou J, Huss M, Abid A, et al. A primer on deep learning in genomics. Nat. Genet. 2019; 51:12–18

14. Gong X, Zhang H, Liu X, et al. Is liquid biopsy mature enough for the diagnosis of Alzheimer's disease? Front. Aging Neurosci. 2022; 14:891

15. Lone SN, Nisar S, Masoodi T, et al. Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments. Mol. Cancer 2022 211 2022; 21:1–22

16. Liu L, Chen X, Petinrin OO, et al. Machine learning protocols in early cancer detection based on liquid biopsy: A survey. Life 2021; 11:1–39

17. Peneder P, Stütz AM, Surdez D, et al. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. Nat. Commun. 2021; 12:

18. Gu F, Wang X. Analysis of allele specific expression-a survey. Tsinghua Sci. Technol. 2015; 20:513–529

19. Im YR, Tsui DWY, Diaz LA, et al. Next-Generation Liquid Biopsies: Embracing Data Science in Oncology. Trends in Cancer 2021; 7:283–292

20. Zhou J, Li L, Wang L, et al. Establishment of a SVM classifier to predict recurrence of ovarian cancer. Mol. Med. Rep. 2018; 18:3589–3598

21. Xu G, Zhang M, Zhu H, et al. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. Gene 2017; 604:33–40

22. Constantin N, Sina AAI, Korbie D, et al. Opportunities for Early Cancer Detection: The Rise of ctDNA Methylation-Based Pan-Cancer Screening Technologies. Epigenomes 2022; 6:1–27

23. Bahado-Singh RO, Radhakrishna U, Gordevičius J, et al. Artificial Intelligence and Circulating Cell-Free DNA Methylation Profiling: Mechanism and Detection of Alzheimer's Disease. Cells 2022; 11:1–19

24. Kelley DR, Snoek J, Rinn JL. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016; 26:990–999

25. Onimaru K, Nishimura O, Kuraku S. Predicting gene regulatory regions with a convolutional neural network for processing double-strand genome sequence information. PLoS One 2020; 15:e0235748

26. Kelley DR, Reshef YA, Bileschi M, et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. 2018;

27. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita.

28. Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (80-.). 2012; 337:816–821

29. Doench JG, Hartenian E, Graham DB, et al. Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nat. Biotechnol. 2014 3212 2014; 32:1262–1267

30. Doench JG, Fusi N, Sullender M, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. 2015 342 2016; 34:184–191

31. Kim HK, Kim Y, Lee S, et al. SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance. Sci. Adv. 2019; 5:

32. Chuai G, Ma H, Yan J, et al. DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. Genome Biol. 2018; 19:1–18

33. Kim HK, Min S, Song M, et al. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. Nat. Biotechnol. 2018 363 2018; 36:239–241

34. Moreno-Mateos MA, Vejnar CE, Beaudoin JD, et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat. Methods 2015 1210 2015; 12:982–988

35. Wang J, Zhang X, Cheng L, et al. An overview and metanalysis of machine and deep learning-based CRISPR gRNA design tools. RNA Biol. 2020; 17:13