Deep learning models in genomics; are we there yet?

Author: Lefteris Koumakis

2023-12-1

Introduction

01

The advances in biotechnology

02

The rapid increase in computer speed

03

The spread of information technology to many different areas

The new era of Big Data

Introduction

- Previously, the most influential methodologies in bioinformatics were Machine Learning (ML) methodologies.

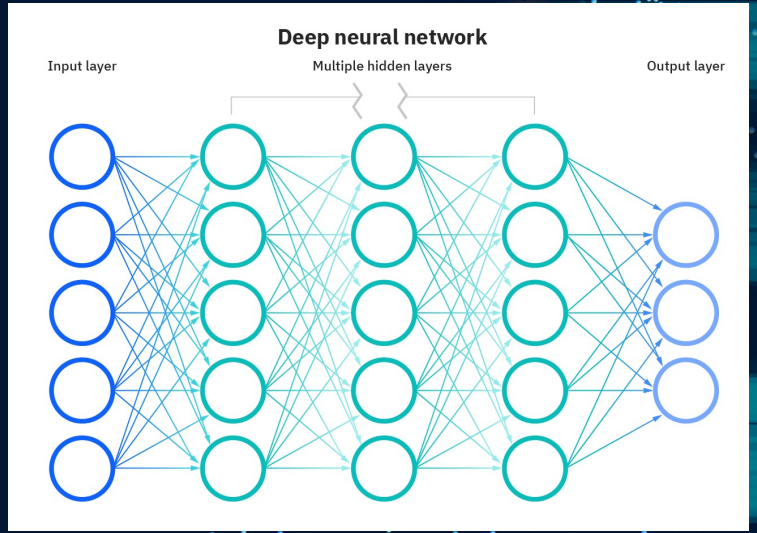
- With the massive generation of data, Deep Learning (DL) approaches proved to be more efficient than ML.

Introduction

- The main advantages of DL over ML:
  - ✓ DL can handle natural data in their raw form efficiently.
  - ✓ DL provides models with higher accuracy at discovering patterns in high dimensional data.
- However, in case of DL, the amount of training data is more demanding, and drastically affects the predicting value of the trained model.

The start of DL models

- ✓ first theorized in the 1980s based on the notion of neurons
- ✓ The term “deep” in DL refers to the number of layers through which the data is transformed.
- ✓ The hard requirements of DL models for large amounts of training data and substantial computing power, placed them unrealistic or limited until the introduction of special hardware such as the high-performance GPUs with parallel architecture.
- ✓ Nowadays deep learning architectures, as the state-of-the-art predictive models, have been applied to many fields.



DL architectures

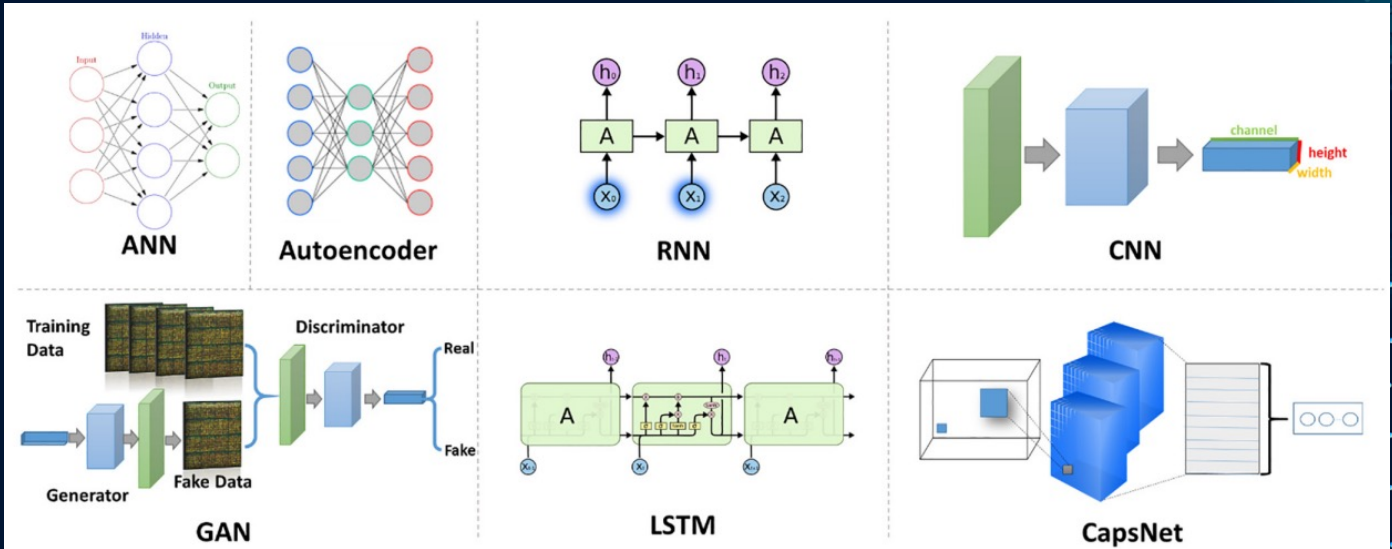
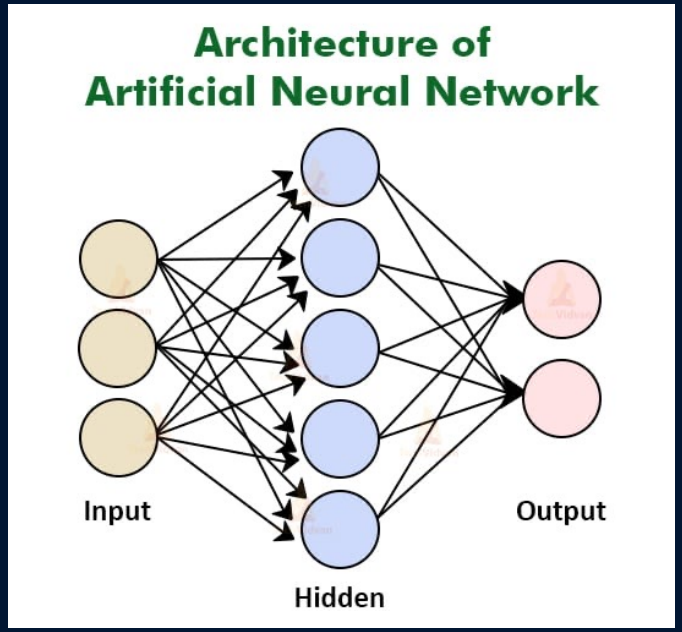


Fig. 1. Architecture of the main deep learning models.

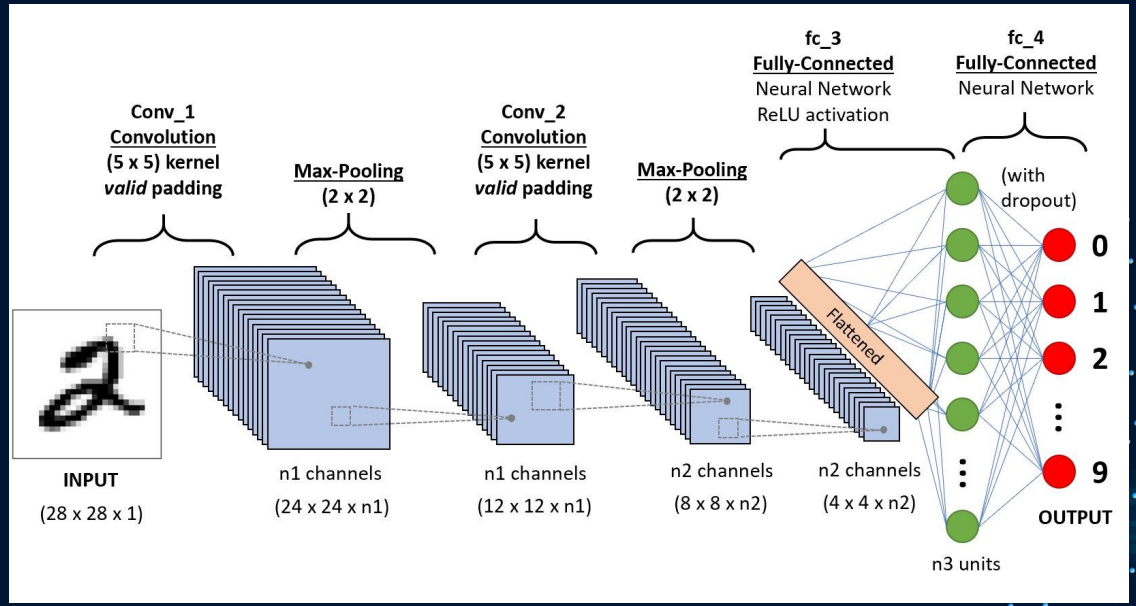
DL architectures

Artificial Neural Networks (ANN)



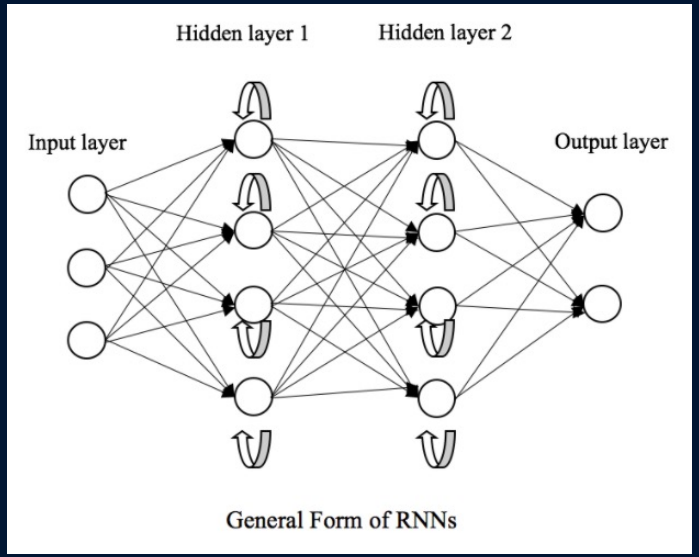
DL architectures

Convolutional Neural Network (CNN)



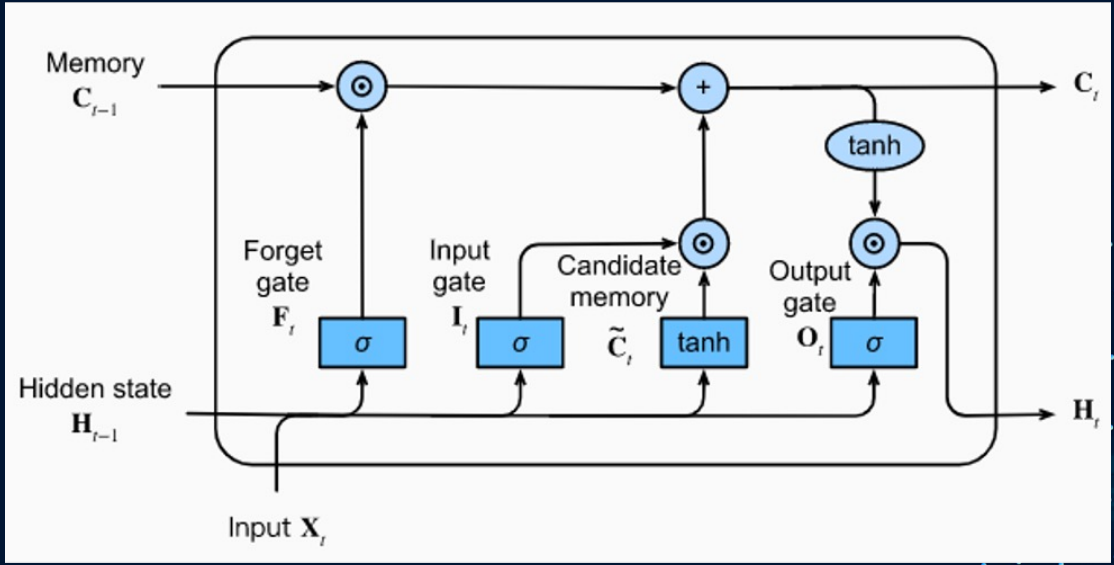
DL architectures

Recurrent Neural Network (RNN)



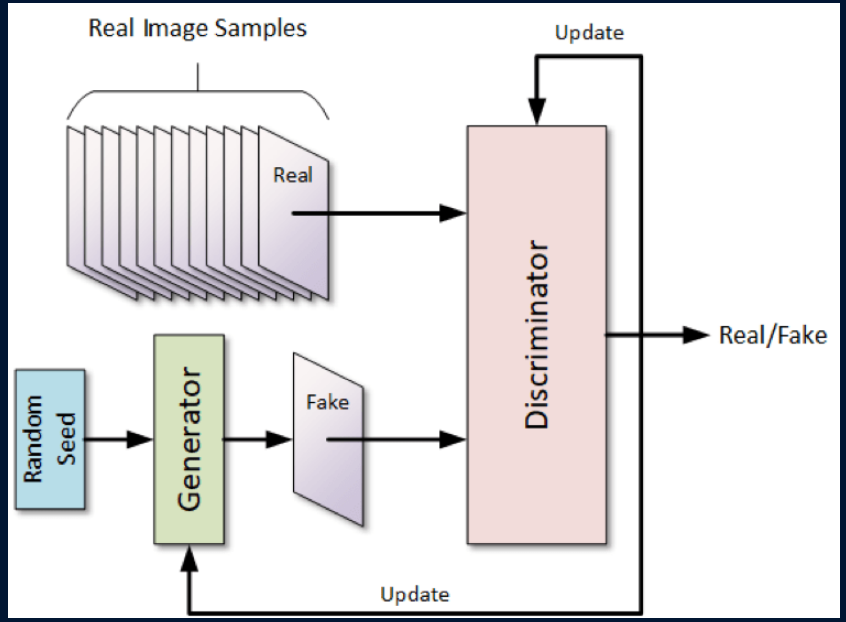
DL architectures

Long short-term memory (LSTM)



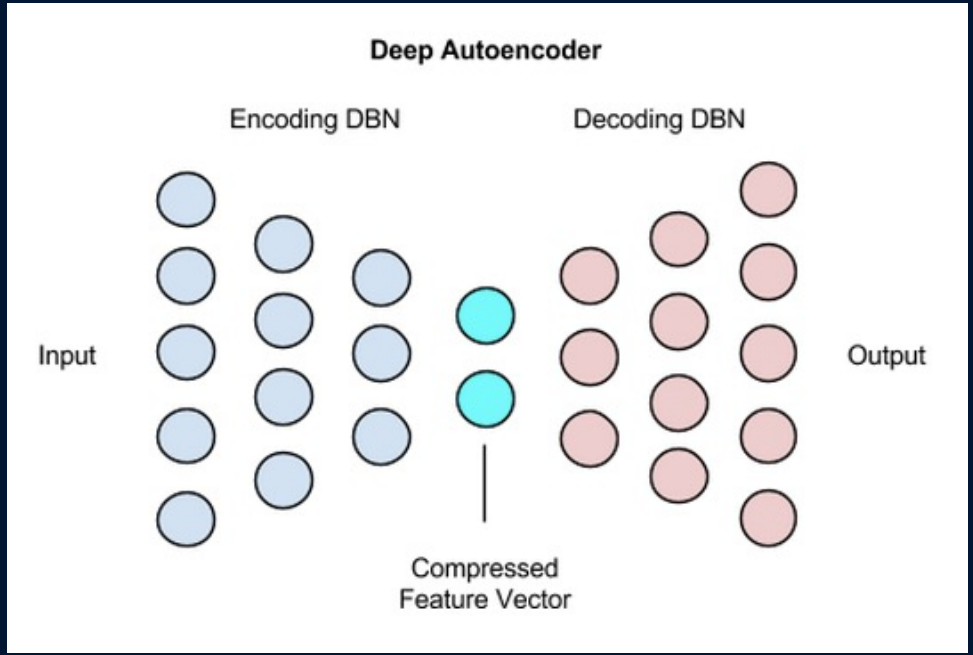
DL architectures

Generative Adversarial Networks (GANs)



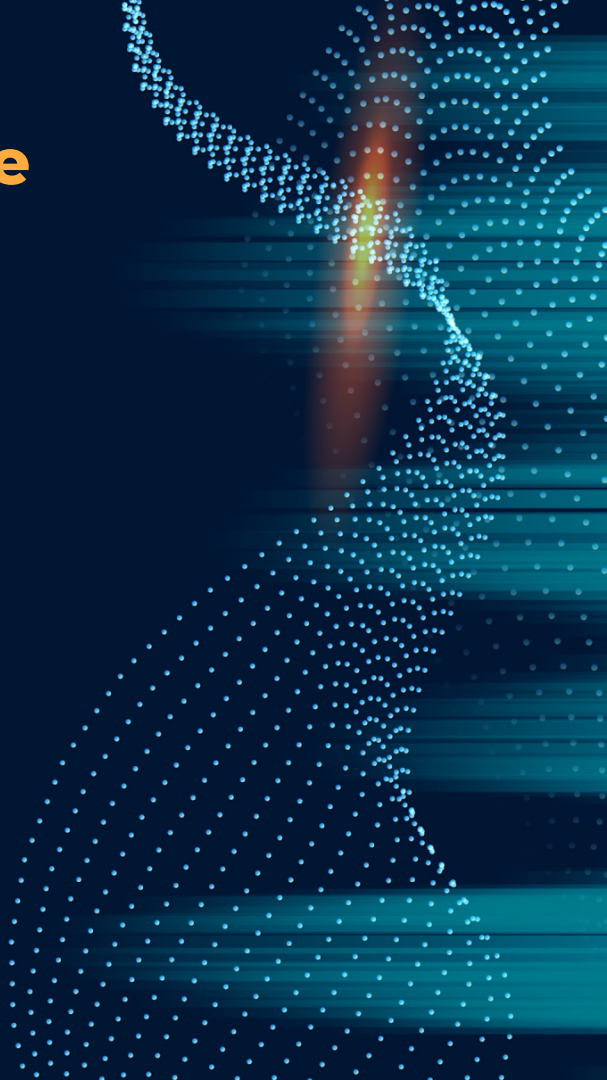
DL architectures

Autoencoders (AE)

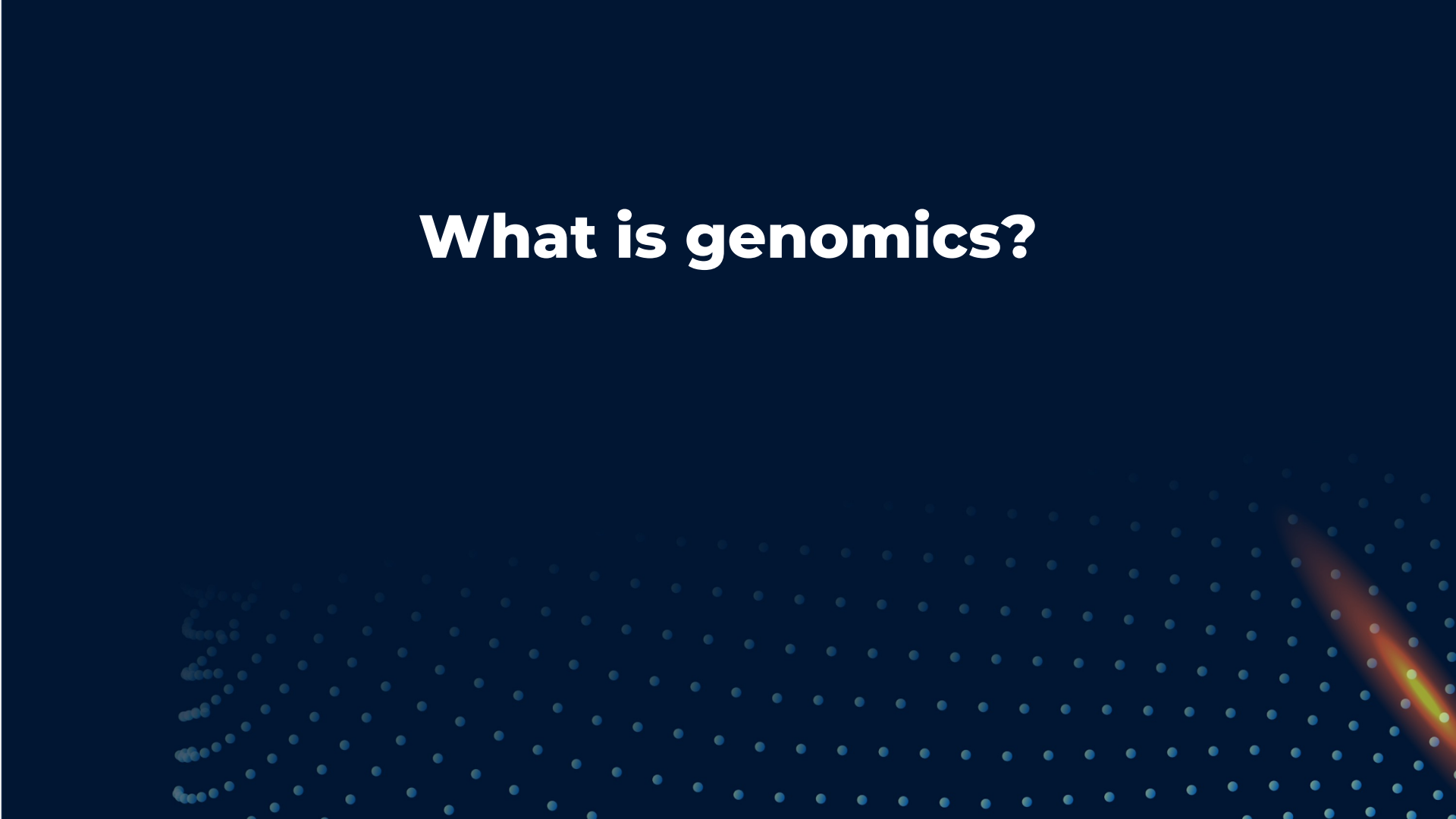


Other methodologies to improve accuracy

- ✓ Combining ML and DL: multi model fusion
- ✓ Transfer learning

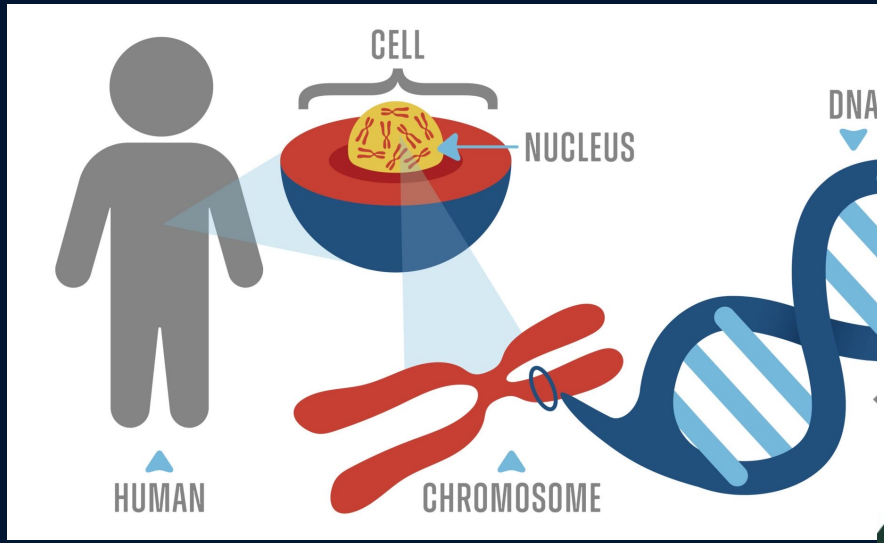


What is genomics?



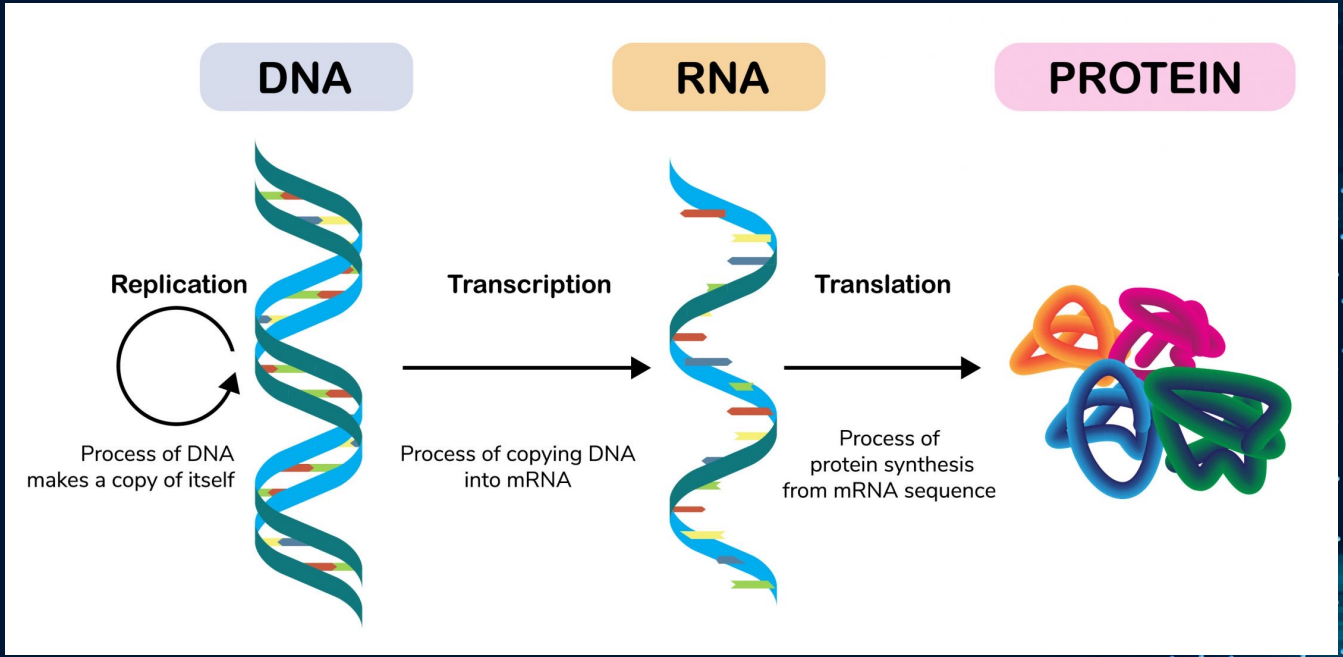
Some basic biological concepts

What is a gene?



Some basic biological concepts

The central dogma

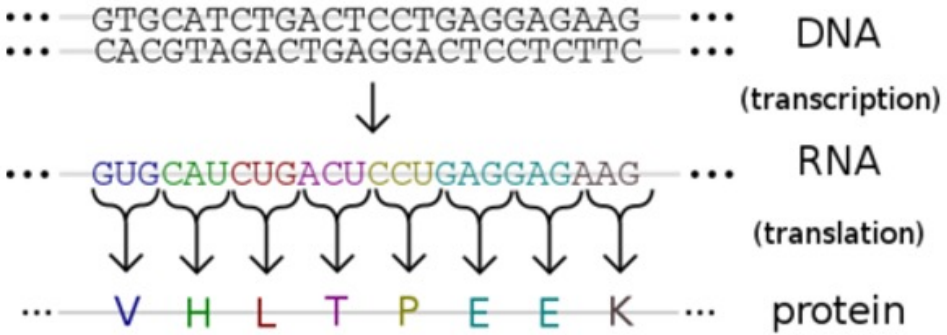


Some basic biological concepts

The process of translation

Second base in codon

First base in codon	U	C	A	G	Last base in codon
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA STOP	UGA STOP	A
	UUG Leu	UCG Ser	UAG STOP	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met (start)	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G



Now imagine how much data we can have!

- The expression level of thousands of genes or non-coding transcripts (e.g., miRNAs)
- Sequences of genomes, produced by next generation sequencing (NGS)
- The associations between genetic markers and disease, obtained by genome-wide association studies (GWAS)
- The state of gene variants (SNPs, Indels) as well as other genome alterations (e.g., copy number variations CNVs) for different populations

Materials and method

- Not a systematic literature review of deep learning methodologies for genomics

Rather:

- **Capturing the current trends in the area**

❑ Excluded studied:

  - Studies focusing on radiogenomics where deep learning architectures used only for the image analysis and then combined with genomics analysis using statistics or traditional machine learning methodologies

  - Studies where deep learning was used only for data augmentation and synthetic data generation paired with genomics or other analysis

Table 1
. List of deep learning methodologies in genomics. From left to right the columns represent the DL model acronym (if any), the respective publication, DL model, omics data used as input, prediction/research question, evaluation metrics and the comparison with other classic ML methods (if any).

Name	Publication	DL model	omics data	Purpose / Prediction	accuracy	performance gap over other methods
DeepTarget	[40]	RNN	miRNA-mRNA pairing	target prediction	0,96	+25% f-measure
DeepMirGene	[41]	LSTM	positive pre-miRNA and non-miRNA	miRNA target	0.89 sensitivity	+4% f-measure
DeepNet	[43]	ANN	RNA-Seq	control-cases	~0.7	same or worst AUC from LASSO
	[44]	AE	time-series gene expression	pre-processing step for clustering		Better than PCA
	[45]	AE	cDNA microarrays	Predict the organization of transcriptomic machinery	–	significant overlap with previous studies
ADAGE	[47]	AE	gene expression	identification/ reconstruction of biological signals	–	significant overlap with post-hoc analysis KEGG
eADAGE	[47]	AE	gene expression	identification of biological patterns	–	significant overlap with post-hoc analysis KEGG
D-GEX	[48]	RNN	expression of landmark genes	Gene expression inference	overall error 0.3204 ± 0.0879	Outperforms Linear Regression(LR) (+15.33%) and KNN-GE in most of the target genes
DeepChrome	[49]	CNN	histone modifications	classify gene expression	Average area under the curve (AUC) = 0.80	(+5%) from support vector machines (SVM), (+21% from random forest (RF)
AttentiveChrome	[50]	LSTM	histone modifications	classify gene expression	Average AUC = 0.81	Marginally better than DeepChrome
Multimodal deep belief network	[54]	DBN	gene expression, DNA methylation and miRNA expression	Identification of Key Genes and miRNAs	average correlations 0.91, 0.73 and 0.69 for the GE, DM and ME	–
DeepVariant	[51]	CNN	whole-genome sequence	variant caller	99,45% F1	produced more accurate results with greater consistency across a variety of quality metrics
	[53]	ANN	cell-line with drug response	predict drug response	0.65 AUC	Outperformed FR 0.54 AUC and elastic nets 0.51 AUC
DeepFIGV	[52]	CNN	whole-genome sequence	predict quantitative epigenetic variation	z-scores DNase rho = 0.0802, P = 5.32e–16	
DeePathology	[55]	Multiple AEs	mRNA and miRNA	predict tissue-of-origin, normal or disease state and cancer type	99.4% accuracy for cancer subtype	95.1% for SVM
DeepCpG	[57]	CNN	Single cell methylation	predicts missing methylation states and detects sequence motifs	89% AUC	86% AUC for Random Forest
CNNC	[56]	CNN	scRNA-seq	predicting transcription factor target	~70% accuracy for multiple experiments	Outperformed GBA (guilt by association) and DNN (fully connected DL) across a variety of experiments
DanQ	[22]	CNN and RNN	DNA-seq	predicting the function of DNA directly from sequence alone	AUC score ~ 70%	Outperformed LR and DeepSEA (CNN DL), with over 10% improvement in AUC
FBGAN	[17]	GANs	DNA-seq	optimize the synthetic gene sequences	Train accuracy 0.94 test accuracy 0.84	Outperformed kmer and Wasserstein GAN trained directly on AMPs

Results and discussion

- Deep learning models are considered the state of the art for classification and clustering when we deal with big data such as the -omics area.
- Nevertheless, we are still far from providing DL models for -omics data that can be used in the precision medicine since the proposed methodologies have not been validated yet in the clinical practice.
- The success of DL depends on finding an architecture to fit the research question and be capable to handle the respective data.

Limitations of DL in genomics

Model interpretation (the black box)

- It is difficult to understand the rational.
- There are problems extracting the causality relationship between the data and the outcome.
- '**White-box**' approaches and **explainable AI techniques** are more preferred.

Limitations of DL in genomics

The curse of dimensionality

- The genomic datasets usually represent a very large number of variables and a small number of samples.
- Not only a problem for DL but also for less demanding (in terms of samples) ML algorithms.
- Fortunately, there are repositories that provide access to public data and one can combine datasets from multiple sources. Nevertheless, in order to collect a representative cohort for DL training, a lot of preprocessing and harmonization is needed.

Limitations of DL in genomics

Imbalanced classes

- Most of the DL and ML models for genomics deal with classification problems e.g. discrimination between disease and healthy samples.
- It is well-known that genomics trials and data gathered from various sources are usually inherently class imbalanced.
- ML/DL models cannot be effective until a sufficient number of instances per class has been fitted.

Transfer learning

Limitations of DL in genomics

Heterogeneity of data

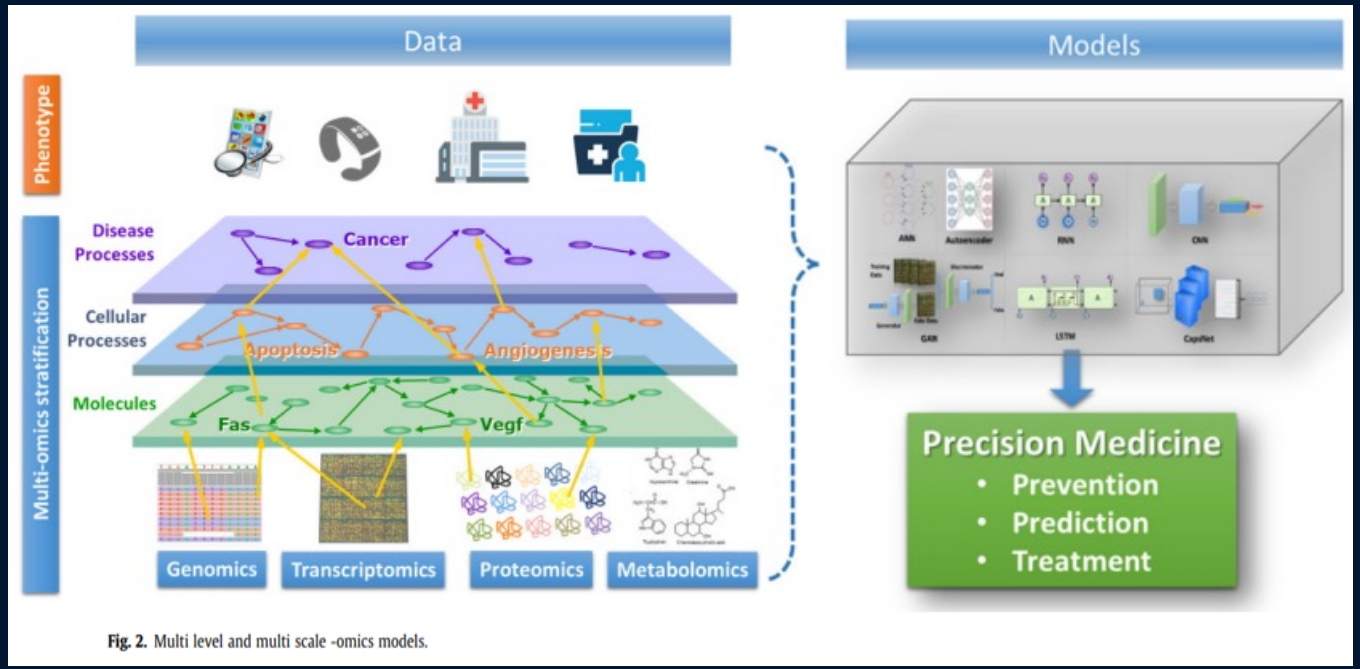


Fig. 2. Multi level and multi scale -omics models.

Limitations of DL in genomics

Parameters and hyper-parameters tuning

- One of the most difficult steps for DL is the tuning of the model.

- The main tuning hyper-parameters for every DL architecture are the learning rate, the batch size, the momentum, and the weight decay. These hyper-parameters act as knobs which can be tweaked during the training of the model. A wrong setting in any of these parameters may result to under-fitting or over-fitting.

Conclusions

- It is evident that DL models can provide higher accuracies in specific tasks of genomics than the state of the art methodologies.
- In addition, deep learning has the ability to deal with multimodal data effectively and genomics offers extremely heterogeneous data making them an excellent candidate for the realization of precision medicine.

But we are NOT there yet!

Future directions:

- The process of translating the knowledge acquired in genomics research into clinically useful tools has been extremely slow.
- More efforts should be made to analyze and combine datasets (private and public) in order to enhance the role of DL genomics in prediction and prognosis.
- Explainable DL models can pave the way for identifying not only novel biomarkers but also regulatory interactions in different pathology conditions such as tissues and disease states.

Thanks for your attention.