財團法人醫藥品查驗中心 Center for Drug Evaluation, Taiwan 當代醫藥法規月刊 2023-11 157
回 CDE 官網首頁
訂閱/取消電子報

[名家專欄]

回顧次世代定序的卷積神經網路運用

曾嶔元[1]

前言

隨著大規模平行高通量定序技術的出現，對人類基因體進行定序以識別基因變異，已成為基因體學的必要步驟。大數據管理和解讀需要龐大的伺服器和熟練的生物資訊分析。從次世代定序的海量資訊中確定並獲得的基因變異型，其資訊和類別(包括良性、可能良性、意義未明、可能致病和致病變異型) 對臨床意義至關重要，但卻相當困難完成。人工智能的深度學習衍生出許多的深度神經網絡，其中使用最廣的就是卷積神經網絡(convolutional neural network, CNN)。從 CNN 產生的模型可有效地處理次世代定序所產生的大規模和複雜的資料，包括變異型辨認(variant calling)和變異型註解(variant annotation)、疾病變異型(disease variants)、基因表達和調控(gene expression and regulation)、表觀遺傳學(epigenomics)、以及藥物基因體學(pharmacogenomics)。這些都是近幾年發生的事情，值得我們回顧其發展。

次世代定序

近年來，生命科學領域產生的數據量急劇增加，主要是由高通量定序(high-throughput sequencing, HTS)技術的快速進步所推動。次世代定序(Next Generation Sequencing, NGS)可從取自於血液樣本、腫瘤樣本、細胞系、福爾馬林固定石蠟包埋塊、和液體切片樣本的 DNA 和 RNA，來生成生物體的整個數據庫。這些技術包括全基因體定序(whole genome sequencing, WGS)、全外顯子定序(whole exome sequencing, WES)、目標定序 (target sequencing)、轉錄體(transcriptomic)、和蛋白質體(proteomic)分析[1-5]。NGS 也是各種疾病相關的表觀遺傳學研究的基石[4]。它們在

[1] 元鼎診所院長/醫師

理解嚴重急性呼吸綜合症冠狀病毒 2 (SARS-CoV2) 的起源和流行病學方面，發揮很大的作用[5]。

多體學(multi-omics)研究產生了大量的數據，NGS 產生的數據量通常以吉字節(十億位元組, gigabyte)、太字節(兆位元組, terabyte)、或拍字節(又稱千兆位元組, petabyte) 呈現。NGS 使用文件格式，例如 FASTQ (用於對齊參考序列)、BAM (序列對齊/映射的二進製版本)、和 VCF (變異型辨認格式)，來生成大型數據庫。文件大小取決於覆蓋範圍和讀取長度[6]。面對這些大型數據庫的挑戰是，主要來自執行臨床相關結果的分析和解釋。這些成千上萬個基因表達或非編碼轉錄，在判讀和臨床運用上造成非常大的困擾。

近年來，人工智能(artificial intelligence, AI)的卓越性能在解讀次世代定序數據上充分得到發揮[7]。這些實例包括找到基因型-表型相關性、生物標誌物的鑒定和基因功能預測，以及映射生物醫學活躍的基因體區域，如轉錄強化子(enhancer)[7-11]。

人工智能

人工神經網路(artificial neural network, ANN)在人工智能中有悠久的歷史，最早可以追溯到 1958 年 Rosenblatt 所提出的受生物啟發的感知器模型(biologically inspired Perceptron model)[12]。

機器學習(machine learning, ML)演算法，旨在以機器的學習能力來解決問題和做出決策[13-15]。機器學習在各個領域和應用中被廣泛地使用，例如圖像和語音識別、自然語言處理、金融預測、醫學診斷、自動駕駛、以及許多其他需要數據驅動決策的領域。機器學習的主要目標是允許機器自動從數據和模式中學習，適應新信息，並在沒有人類干預的情況下進行預測或決策。傳統的機器學習使用較簡單的神經網絡架構，例如前饋神經網絡(feedforward neural networks, FNN)，也稱為多層感知器(multilayer perceptrons)。這些網絡由輸入層、一個或多個隱藏層、和一個輸出層組成。它們非常適合用於分類和回歸等任務。

機器學習的分類任務(如預測疾病的存在/不存在)，最常應用於基因體學研究領域，以及用於比較不同模型的表現。例如，曲線下面積(area under the curve, AUC)是評估模型表現最廣泛使用的指標之一。它測量真陽性率(TPR)或靈敏度、真陰性率(TNR)或特

異度和假陽性率(FPR)。此外，F1 得分用於測試高度不平衡數據集中模型的準確性，是精確度和召回率之間的調和平均數(harmonic mean)。對於 AUC 和 F1 得分，越大的值反映出更好的模型表現。同樣，混淆矩陣(confusion matrix)通過測量模型的準確性來描述完整的模型表現，計算出真陽性值加上真陰性值，並將總和除以樣本的總數[16,17]。

就次世代定序的數據而言，未使用深度神經網絡(deep neural networks, DNN)的機器學習演算法的例子如下：SIFT (Sorting Intolerant From Tolerant)、PolyPhen2 (Polymorphism Phenotyping v2)、PROVEAN (Protein Variation Effect Analyzer)、MutationTaster、和 AlignGVGD。它們基於不同類型的演算法，包括序列比對、演化保守性、和其他統計方法，來預測遺傳變異的功能影響。MaxEntScan 和 NNSPLICE 可以預測剪接位點。MaxEntSca 使用最大熵模型，NNSPLIC 則使用前饋神經網絡來識別 DNA 序列中的剪接位點。至於 GeneSplicer 和 Human Splicing Finder，它們分別關注剪接位點預測和剪接位點分析。

機器學習通常使用人工神經網路作為從數據中學習，並進行預測的強大工具。人工神經網絡由連接的節點或神經元組成，並組織成層。每個神經元都接收輸入，使用激活函數進行處理，並產生輸出。神經元之間的連接具有相應的權重，這些權重在訓練期間進行調整，以使網絡能夠從數據中學習並進行準確的預測。機器學習具有廣泛的技術應用；然而，標準的機器學習方法不適於複雜的、自然的、高度維度的原始數據，例如基因組學數據。

相反地，深度學習(deep learning, DL)是機器學習的一個子領域，它專注於構建和訓練具有多個隱藏層的深度神經網路，改進了人工智慧的準確性[18,19]。本質上，深度學習是機器學習的一個子集，而人工神經網路是深度學習中使用的基礎架構之一[20-23]。

在深度學習中，先驗(prior)指的是關於數據生成過程，或輸入輸出之間潛在關係的先驗知識 (prior knowledge)或假設。先驗知識是許多機器學習算法(包括深度學習模型) 的基本。它提供了一種將現有知識或信念納入學習過程的方式，引導模型的行為並改善其性能。先驗可以針對特定的任務或問題採取不同的形式。例如，在圖像識別中，先驗可能是某些模式或形狀在數據中更有可能出現的假設。在自然語言處理中，先驗可能是某些單詞，更有可能在句子中跟隨其他單詞。通過納入這樣的先驗知識，深度學習模型可以做出更明智的預測，並拓展到新的、未見過的數據。深度學習模型通常對數據

生成過程編碼了一個額外的先驗：即輸入到輸出的關係，可以通過多個較簡單的函數或構建塊的階層組合來表達。這個明顯較為溫和的假設，提供了比簡單的局部恆定性先驗更多的結構。此外，對先驗的進一步假設可以通過深度神經網絡的拓撲結構來編碼。

沃爾珀特的無免費午餐定理(Wolpert's no free lunch theorem)[24]指出，在所有類型的數據和任務中，沒有單一的機器學習或優化算法是普遍優於所有其他算法的。該定理意味著任何算法的性能，都取決於其應用問題的具體特徵。也就是說，對於所有數據生成分佈的平均值，每種可能的機器學習算法，在以前未觀測到的數據點上具有相同的錯誤率。因此，如果對於數據生成過程無法進行額外的假設，則在以前未觀測到的數據上，沒有任何機器學習方法可以被認為優於或劣於其他方法。這個定理提醒我們，當為特定的任務選擇算法時，開發人員和研究人員應考慮數據和問題本身的特性，以選擇最適合的方法。

然而，在現實面，這是不切實際的，因為有關問題領域的先驗知識通常會使得某些數據生成分佈比其他分佈更可能。這些假設，可以被制定成對數據生成過程的先驗。其中一個常見的先驗是平滑或局部恆定性先驗[25]，亦即相似的輸入數據點應該具有相似的輸出值。簡單來說，如果平滑性是我們推斷值的唯一依據，則我們須要接近每個我們想要進行推斷的數據點的訓練範例。因此，要成功地構建機器學習模型，既需要對數據生成分佈的預期特性有所了解，又需要在模型中將這種知識編碼進去。

隨著機器學習的興起，各式各類的神經網絡架構，在各個領域變得普遍。目前在基因資料庫上使用的人工神經網路，包括前饋神經網路、卷積神經網絡(convolutional neural network, CNN)、遞歸神經網絡(recurrent neural networks, RNNs)、長短期記憶網絡(long short-term memory, LSTM)、和自編碼器(autoencoder, AE)等等。其中，卷積神經網絡(CNN)是次世代定序數據中最常用的演算法。

卷積神經網絡

卷積神經網絡(CNN)又稱 ConvNet，是一種特定類型的深度神經網絡(DNN)架構，特別適用於圖像和空間數據的處理。CNNs 最常見的應用是在圖像處理方面，最初被認為是用於分類手寫字體的全自動圖像網路解析器[26]。

CNN 使用卷積層、池化層、和全連接層來自動學習和提取輸入數據中的相關特徵[27, 28]。在 CNN 中，至少在一些層中，矩陣乘法被替換為使用(通常是相當小的)過濾器核矩陣(filter kernel matrix)進行卷積，其中核矩陣的元素在訓練期間進行學習。鑑於卷積允許通過空間鄰域的加權平均值來替換值，它能夠編碼有關空間結構或模式的資訊。卷積通常跟隨著一個稱為池化(pooling)的操作，它將幾個卷積結果合併成單一的輸出。例如，最大池化(max-pooling)用最大值替換一組卷積結果。這允許某種程度的不變性；例如，在電腦視覺中，將相同的過濾器核對幾個相鄰像素進行最大池化卷積，可以產生平移不變性。鑑於在每個輸入值處計算相同的卷積通常是有用的(意味著將相同的空間模式與每個輸入點進行比較)，CNN 通常使用參數共享，其中幾個神經元被限制為使用相同的過濾器核。

CNN 層中的卷積單元可以從前一層的單元獲取輸入數據，這些單元共同生成一個預測。這種深度構造的關鍵原理是大量的處理和連接特徵，用於推斷輸入和輸出之間的非線性關聯性[29, 30]。

在分析 NGS 數據的背景下，通常會將 CNN 應用於多個序列讀取對齊堆疊圖像中，以檢測局部模式，從而識別突變或序列模式。基本概念是，將以感興趣的位置 (候選變異型) 為中心的對齊讀取轉換成圖像，並應用在解決電腦視覺任務中已知的 CNN。這是因為輸入可以被視為對讀取堆疊的圖像，而變異與相鄰的對齊讀取之間的複雜依賴性可以由卷積核(convolutional kernels)來建模。

以卷積神經網路解讀次世代定序的數據

由於卷積神經網絡(CNN)是次世代定序數據最常用的演算法。本文就以此演算法討論次世代定序數據的解讀。讀者如果對以其他演算法解讀基因數據有興趣的話，請參閱拙作《以人工智能解碼人類基因》(刊載於 2023 年 9 月出刊的《生物醫學雜誌》第 16 卷第三期雜誌)。

變異型辨認 (variant calling) 和變異型註解 (variant annotation)

變異型辨認是指藉由對比參考基因體和檢測樣本基因體的差異，來鑒定和標記出檢測樣本中的變異型。通過變異型辨認，可以得到檢測樣本中各種變異型的基因型資訊，為後續變異型註解和分析奠定基礎。變異型註解是指對已標記的變異型進行註釋，包括

該變異型在基因組中的位置、對應的基因和轉錄本、變異型的功能類型等資訊。通過變異型註解，可以對變異型進行進一步的分析，為疾病的分子機制研究提供重要的線索。

變異型辨認和變異型註解是基因體學研究中的關鍵步驟，它們通常涉及大量數據的處理和複雜的分析。這些步驟的準確性和可靠性，對於確定基因體變異的意義和疾病相關性至關重要，並且為基因體學和臨床基因測試提供了基礎。隨著技術的不斷發展，變異型辨認和變異型註解的方法和工具也在不斷進化，使得基因體學研究變得更加深入和準確。

Google 開發的 DeepVariant[31]可辨認單核苷酸變異(SNV)和小插入/刪除(indels)。DeepVariant 依賴於輸入圖像的圖形差異，來執行 NGS 短序列變異辨認的分類。它將映射的定序數據集視為圖像，並將變異辨認轉換為圖像分類任務[31]。Kumaran 等人將 DeepVariant 與傳統的變異型辨認程序 (例如 SAMtools 和 GATK)相結合，證明可以提高單核苷酸變異和 Indel 檢測的準確性[32]。網絡輸出候選位置的基因型概率，亦即同型合子參數(homozygous reference)、異型合子或同型合子替代然而，此模型不提供有關變異的詳細資訊。因此，DeepVariant 被歸類為不完整的變異辨認模型(incomplete variant caller model)[33]。

Clairvoyante 模型可以預測變異類型(SNV 或 Indel)、雜合度、等位基因和 Indel 長度。因此，它克服了 DeepVariant 模型缺乏完整變異資訊的缺點，包括精確的替代等位基因和變異類型。Clairvoyante 模型是專門設計用於利用從 SMS 技術(例如 PacBio 和 ONT)生成的長讀序列數據，但也通常適用於短讀數據集[33]。Clairvoyante[34] 使用多任務 CNN 進行單分子定序數據處理，與 DeepVariant 相比，所需的參數數量減少了大約一個數量級，並且在長讀取技術方面取得了良好的性能。

Neusomatic[35]是一種計算方法，用於從次世代定序數據中的腫瘤樣本進行體細胞變異型辨認。它利用深度學習技術，尤其是 CNN 架構，來識別癌症基因組中體細胞遺傳變異。該方法不使用讀取堆疊(read pile-ups)作為影像，而是使用從每個對齊列中提取的特徵作為神經網絡的輸入。這導致了一個簡化的 CNN 架構，具有較低的計算複雜性，既適用於訓練，也優於傳統的體細胞突變檢測方法。Neusomatic 的主要重點是檢測可能涉及癌症發展的 SNV 和小 indels。通過應用 Neusomatic，研究人員可以獲得

有關腫瘤細胞中發生的基因變化的見解，有助於理解癌症進展的潛在機制，和開發個人化的治療。

DeepSV[36]可用於檢測長缺失，因為 DeepSV 能從序列讀取圖像中提取長的基因刪除(>50 bp)，而不是其他類型的結構變異 (如長插入或反轉)。它輸入 BAM 格式或 VCF 文件，並以 VCF 形式輸出結果。關於性能方面，DeepSV 曾經與另外八個刪除辨認工具和一個名為 Concod 的機器學習工具進行了比較[37]。結果顯示，儘管 Concod 在較少的訓練樣本情況下具有較短的訓練時間，但 DeepSV 在使用相同數據集時顯示出更高的準確性和更少的訓練損失[38]。

疾病變異型 (Disease variants)

疾病變異型指的是某種基因變異與某種疾病的發病率有關。這些變異可以是單一的基因突變，也可以是多個基因的組合。研究疾病變異型的目的，是要確定哪些基因突變與特定疾病有關，以便更好地理解疾病的成因和發展。

CNNScoreVariants 是一個基於 CNN 的計算模型，用於預測基因變異的功能影響[39]。它旨在評估基因突變對蛋白質功能的潛在影響，特別是在遺傳疾病的情況下，CNNScoreVariants 將蛋白質序列資訊作為輸入，並輸出一個預測分數，該分數顯示變異型的可能性。該模型已被證明優於其他傳統基於機器學習的變異型分析算法，並已集成到 Genome Analysis Toolkit (GATK)軟件中，用於變異型辨認。

當幾個較不清楚的候選變異型對應於特定的表現型時，我們無法用 in silico 預測模型得到答案。對於這樣的問題，目前可用深度神經網絡架構進行變異型的優先排序，例如一個依賴 CNN 算法的變異型註解器 Basset[40]，可以使用 DNase I 敏感性定序數據作為輸入，來預測致病性的單核苷酸多型性(SNP)。

NGS 技術問世後，WGS 解釋了整個人類基因體中的編碼和非編碼片段的基因體變異。目前已有幾種基於機器學習的方法提供優先處理非編碼變異型的方法；然而，辨識那些與複雜特徵(例如癌症)相關的變異型，是一項具有挑戰性的任務。此外，為了預測一般和精確的新相關性，需要大量與特定表現型(phenotype)相關的陽性變異型。現在也有幾種深度學習方法來克服這些挑戰。例如，DeepWAS 模型依賴於 CNN 演算法，使得它能夠對每個變異型，在眾多細胞類型特定的染色質特徵上，進行調控影響預測[41]。

DeepWAS 模型的主要結果是，直接確定在相關組織中對某種染色質特徵有共同效應的與疾病相關的 SNP。在結合不同資源和組織的表達和甲基化定量特徵座位資料(eQTL 和 meQTL)後，DeepWAS 模型可檢測與疾病相關的轉錄活性基因體位置的能力。DeepWAS 模型用於識別與疾病或特徵相關的 SNP，而 ExPecto 模型則用來預測突變/功能的組織特異性轉錄效應[42]。

基因表達和調控 (Gene expression and regulation)

基因表達指的是基因轉錄成 mRNA，進而轉譯成蛋白質的過程。這個過程涉及到多個步驟，包括激活基因、基因轉錄、RNA 剪切、RNA 運輸、翻譯和後轉譯修飾等。這些步驟都需要多種調節機制的參與，例如轉錄因子、啟動子、強化子、siRNA、miRNA 等。基因表達的調控是一個非常複雜的過程，不同的細胞類型和環境條件都會對其進行調節，以確保基因表達的準確性和時機性。

MPRA-DragoNN 模型[43]，可用於預測和分析非編碼 DNA 序列數據的轉錄調節活性，這些數據是從 MPRAs 數據中測量得來的。SPOT-RNA 模型用於預測 RNA 的次級結構[44]。DeepExpression 模型可透過啟動子序列和強化子-啟動子相互作用，來預測基因表達[45]。Xpresso 模型結合了啟動子序列及其相關 mRNA 穩定性特徵，以預測 mRNA 的基因表達水平。有趣的是，Xpresso 模型可以在幾個任意的細胞類型中簡單訓練，即使它們缺乏 ChIP 和 DNase 等實驗資訊[46]。

2019 年，Jaganathan 等人構建了 SpliceAI，可從 pre-mRNA 序列識別剪接功能[47]。SpliceAI 使用一種深度殘差神經網絡(deep residual neural network)，以前 mRNA 轉錄本序列作為輸入來預測剪接功能。該架構包含一個 32-空洞卷積層(dilated convolutional layer)，用於識別跨越巨大基因體間隙的序列，因為有數萬個核苷酸分隔剪接供體和剪接受體[47]。

表觀基因體學 (epigenomics)

表觀遺傳變化是指不涉及 DNA 序列的遺傳變化，又稱為表觀修飾。這些修飾可以影響基因表達和細胞功能。表觀基因體學研究的主要目的，是要了解表觀遺傳變化如何影響疾病和其他生物過程。近年來，表觀基因體學技術的快速發展已經帶來了許多突破。這些技術包括 DNA 甲基化定序、組蛋白修飾定序、RNA 甲基化定序等。表觀基因體學

的研究對於疾病診斷、治療和預防具有重要意義。表觀遺傳變化已經與多種疾病的發生和發展有關聯，如癌症、心血管疾病、神經系統疾病等。通過深入研究表觀基因體學，我們可以更容易地理解，這些疾病的發病機制和預防方法。

DeepSEA[48]和 DeepBind[49]是第一批成功應用 CNN 模型於大規模染色質分析數據中，以模擬蛋白質結合的序列特異性。這種方法不需手動設計特徵集，能自動學習有信息量的序列特徵。DeepBind 用於預測結合蛋白，並表現出比傳統模型更強大的預測能力[50]。

DeepTACT 模型可用來預測 3D 染色質相互作用[51]，而 Basenji 模型則用於預測大型哺乳動物基因體中特定細胞型的表觀和轉錄體表徵[52]。

Deopen 模型用於從學習的調節 DNA 序列編碼中，預測整個基因體的染色質可及性。DeepHistone[53]也是一種基於 CNN 的算法，用於預測不同的位點特異標記的組蛋白修飾。為了進行精確的預測，該模型結合了 DNA 序列數據和染色質可及性資訊。它有區分功能性單核苷酸多型性(SNP)和其相鄰遺傳變異的能力，因此可用於研究與疾病相關之變異的影響[54]。

藥物基因體學 (pharmacogenomics)

藥物基因體學是研究基因如何影響藥物作用的學科。通過檢查一個人的基因體，藥物基因體學可以讓我們了解一個人對藥物的反應。這種個人化的方法，可以幫助醫師選擇最有效的藥物治療方案，從而提高治療的成功率，減少不良反應的發生率。

深度學習方法非常適合預測治療的反應[55, 56]。在藥物基因體學應用中，CNN 被用於創建 DrugCell 模型[57]。DrugCell 可用於預測人類癌細胞對治療的反應。它將模型的中心機制與人類細胞生物結構相結合，並允許預測癌症對藥物的反應，然後規劃治療組合。DrugCell 的 VNN 整合細胞基因型，而人工神經網絡(ANN)則整合藥物設計。VNN 模型的輸入包括人類細胞分子各系統間的階層關聯的文件，其中包含 Gene Ontology 數據庫中的 2086 個生物過程標準。ANN 模型的輸入整合了藥物的摩根指紋(Morgan Fingerprints)文件，該文件包含了藥物的化學結構的矢量符號。這兩個部分的輸出被結合成一層神經元，產生特定基因型對特定治療的反應。每種藥物的預測準確性都顯示，具有顯著準確性的藥物亞群。

DeepBL 的模型採用 Small VGGNet 結構(一種 CNN)，並使用 TensorFlow 庫。這個方法使用蛋白質序列作為輸入，檢測對 β-內酰胺類(lactam)抗生素產生耐藥性的 β-內酰胺酶 (β-lactamases, BLs)及其變異型。該模型基於大規模的 RefSeq 數據集，涵蓋 NCBI 數據庫中的>39K BLs。將該模型與其他傳統的基於機器學習的算法(包括 SVM、RF、NB 和 LR)作比較，在一個包含 10,000 多個序列的獨立測試集上，DeepBL 的性能優於這些方法[58]。

儘管藥物基因體學的應用相對有限，但是它在某些情況下已經被廣泛使用。隨著技術的進步和更多研究的開展，相信藥物基因體學將會在未來發揮越來越重要的作用。

結語

隨著深度學習的興起，複雜的神經網絡架構能夠有效地處理大規模且複雜的問題，並減少人為干預。越來越多的人，對使用人工智能進行複雜計算和評估診斷的興趣，日益增加。當我們對次世代定序產生的大量數據感到惶恐時，各種深度神經網絡模型，尤其是以卷積神經網絡為基礎者，適時地協助我們處理海量的資訊。在人工智能的協助下，讓我們目睹正在發生的精準醫療發展。

參考文獻

1. Auffray C, Imbeaud S, Roux-Rouquié M, Hood L. From functional genomics to systems biology: concepts and practices. C R Biol. 2003;326(10–11):879–92.
2. Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, et al. Medical implications of technical accuracy in genome sequenc- ing. Genome Med. 2016;8(1):24.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
4. Yue T, Wang H. Deep Learning for Genomics: A Concise Overview. 2018
5. Honoré B, Østergaard M, Vorum H. Functional genomics studied by proteomics. BioEssays. 2004;26(8):901–15.
6. K. Y. He, D. Ge, and M. M. He, "Big Data Analytics for Genomic Medicine," Int J Mol Sci, vol. 18, no. 2, Feb 15 2017.

7. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. Brief Bioinform. 2020;2:447.
8. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science (80-). 2016;354(6313):769–73.
9. Kulasingam V, Pavlou MP, Diamandis EP. Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer. Nat Rev Cancer. 2010;10(5):371–8.
10. Nariai N, Kolaczyk ED, Kasif S. Probabilistic protein function prediction from heterogeneous genome-wide data. PLoS One. 2007;2(3):e337.
11. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.
12. Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386
13. Jiang F et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230–43.
14. Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clin Infect Dis 2018;66(1):149–53.
15. Adir O et al. Integrating Artificial Intelligence and Nanotechnology for Precision Cancer Medicine. Adv Mater 2020;32(13):e1901989.
16. Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, et al. Peering into the black box of artificial intelligence: evaluation met- rics of machine learning methods. Am J Roentgenol. 2019;212(1):38–43.
17. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. Am J Roentgenol. 2019;212(3):513–9.
18. Joshi AV. Machine Learning and Artificial Intelligence. Springer Nature Switzerland; 2020.
19. Adir O et al. Integrating Artificial Intelligence and Nanotechnology for Precision Cancer Medicine. Adv Mater 2020;32(13):e1901989.

20. Koumakis L. Deep learning models in genomics; are we there yet? Comput Struct Biotechnol J. 2020;18:1466–73.
21. Cao C, Liu F, Tan H, Song D, Shu W, Li W, et al. Deep learning and its applications in biomedicine. Genom Proteom Bioinform. 2018;16(1):17–32.
22. Telenti A, Lippert C, Chang PC, DePristo M. Deep learning of genomic variation and regulatory network data. Hum Mol Genet. 2018;27(R1):R63-71.
23. Kopp W, Monti R, Tamburrini A, Ohler U, Akalin A. Deep learning for genomics using Janggu. Nat Commun. 2020;11(1):3488.
24. Wolpert, D.H. (1996) The lack of a priori distinctions between learning algorithms. Neural Comput. 8, 1341–1390
25. Goodfellow, I. et al. (2016) Deep Learning. MIT Press
26. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
27. Chiu Y-C, Chen H-IH, Zhang T, Zhang S, Gorthi A, Wang L-J, et al. Predict- ing drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med Genom. 2019;12(51):18.
28. Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual understanding of convolutional neural network-a deep learning approach. Procedia Comput Sci. 2018;132:679–88.
29. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional. Neural Netw. 2015;5:71143.
30. Rawat W, Wang Z. Deep convolutional neural networks for image classi- fication: a comprehensive review. Neural Comput. 2017;29(9):2352–449.
31. Poplin, R. et al. (2018) A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983–987
32. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal snp and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983.

33. Luo R, Sedlazeck FJ, Lam T, Schatz MC, Kong H, Genome H. Clairvoyante: a multi-task convolutional deep neural network for variant calling in single molecule sequencing. Science. 2018;3:7745.
34. Luo, R. et al. (2019) A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat. Commun. 10, 998
35. Sahraeian, S.M.E. et al. (2019) Deep convolutional neural networks for accurate somatic mutation detection. Nat. Commun. 10, 1041
36. Cai, L. et al. (2019) DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. BMC Bioinf. 20, 665
37. Cai L, Chu C, Zhang X, Wu Y, Gao J. Concod: an effective integration framework of consensus-based calling deletions from next-generation sequencing data. Int J Data Min Bioinform. 2017;17(2):153.
38. Cai L, Wu Y, Gao J. DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. BMC Bioinform. 2019;20(1):665.
39. Friedman, S. et al. (2020) Lean and deep models for more accurate filtering of SNP and INDEL variant calls. Bioinformatics 36, 2060–2067
40. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26(7):990–9.
41. Arloth J, Eraslan G, Andlauer TFM, Martins J, Iurato S, Kühnel B, et al. DeepWAS: multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. PLOS Comput Biol. 2020;16(2):e1007616.
42. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.
43. Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A. Deciphering regulatory DNA sequences and noncoding genetic variants

using neural network models of massively parallel reporter assays. PLoS One. 2019;71:466689.

44. Singh J, Hanson J, Paliwal K, Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun. 2019;10(1):5407.
45. Zeng W, Wang Y, Jiang R. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. Bioinformatics. 2019;6:7110.
46. Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell Rep. 2020;31(7):107663.
47. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-548.e24.
48. Zhou, J. and Troyanskaya, O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods 12, 931–934
49. Alipanahi, B. et al. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. 33, 831–838
50. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.
51. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res. 2019;47(10):e60–e60.
52. Kelley DR, Reshef YA, Bileschi M, Belanger D, Mclean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Science. 2018;71:739–50.
53. Yin, Q. et al. (2019) DeepHistone: a deep learning approach to predicting histone modifications. BMC Genomics 20, 11–23
54. Yin Q, Wu M, Liu Q, Lv H, Jiang R. DeepHistone: a deep learning approach to predicting histone modifications. BMC Genomics. 2019;20(2):193.

55. Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. Bioinformatics. 2018;34(9):1538–46.
56. Chiu Y-C, Chen H-IH, Zhang T, Zhang S, Gorthi A, Wang L-J, et al. Predict- ing drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med Genom. 2019;12(51):18.
57. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predict- ing drug response and synergy using a deep learning model of human cancer cells. Cancer Cell. 2020;38(5):672-684.e6.
58. Wang Y, Li F, Bharathwaj M, Rosas NC, Leier A, Akutsu T, et al. DeepBL: a deep learning-based approach for in silico discovery of beta-lactamases. Brief Bioinform. 2020;7:8859.