生物技术进展 2021 年 第 11 卷 第 4 期 483~488
Current Biotechnology ISSN 2095-2341

基于深度学习的作物基因组学和遗传改良

辛志奇[1]， 赵航[1]， 汪海[2]， 路铁刚[1*]

1.中国农业科学院生物技术研究所，北京 100081；
2.中国农业大学国家玉米改良中心，北京 100193

摘 要： 随着世界人口的不断增长、食物需求量的不断增加，以及气候的不断变化，如何提高农作物产量已成为人类面临的一个巨大挑战。传统设计育种耗时长、效率低，已经不能满足新时代的育种需求。随着基因型和表型数据成本的不断降低，以及各种组学数据的爆炸式增长，人工智能技术作为能够在大数据中高效率挖掘信息的工具，在生物学领域受到了广泛关注。人工智能指导的设计育种将大大加快育种的效率，给育种带来革命性的变化。介绍了人工智能特别是深度学习在作物基因组学和遗传改良中的应用，并进行了总结与展望，以期为智能设计育种提供新的思路。

关键词： 人工智能；设计育种；深度学习；机器学习

DOI：10.19586/j.2095-2341.2021.0084

中图分类号： S892.6, TP181 **文献标识码：** A

Crop Genomics and Genetic Improvement Based on Deep Learning

XIN Zhiqi[1], ZHAO Hang[1], WANG Hai[2], LU Tiegang[1*]

1.*Biotechnology Research Institute, Chinese Academy of Agricutural Sciences, Beijing* 100081, *China*;
2.*National Maize Improvement Center, China Agricultural University, Beijing* 100193, *China*

Abstract: With an ever-increasing world population and demand for food, ensuring food security has becoming more and more challenging, especially when we are facing severe climate change. However, traditional design breeding is time-consuming and low-efficient, which can't meet the needs of this era. With the development of sequencing technology, the cost of genotyping and phenotyping continues to decrease, resulting in the explosive growth of omics data. Artificial intelligence, as a tool that can efficiently mine information in big data, has attracted wide attention in the field of biology. Artificial intelligence directed breeding design will greatly accelerate the efficiency of design breeding and bring revolutionary changes. This review introduced the application of artificial intelligence, especially deep learning, in genomics and crop genetic improvement, summarized the application progress, and put forward the prospect of how artificial intelligence design breeding, which was expected to provide new thought for artificial intelligence designed breeding.

Key words: artificial intelligence; design breeding; deep learning; machine learning

随着全球人口数量不断增加，到2050年，全球对粮食的需求预计将比2005年增加100%~110%[1-2]。为满足人们对农作物产品日益增长的需求，创新育种技术显得尤为重要。在漫长的农业历史中，育种主要经历了三个阶段：通过观察植株表型，选育优良自交系的传统育种；应用统计学、数量遗传学预先设计杂交育种实验，获得杂种优势的杂交育种；综合单倍体育种、分子标记育种和转基因育种的现代生物工程育种[1]。Edward S Buckler[2]总结了过去的三个时代，并提出了“育种4.0”的概念。王向峰等[1]提出了在“育种4.0”时代深度融合生命科学、信息科学和育种科学的理念。人工智能设计育种是由人工智能与育种相结合，能够给传统育种带来革命性的改变。它包括利用深度学习和机器学习把基因组学、转录组学、蛋白质组学、表观遗传学、代谢组学和表型组学的

收稿日期： 2021-05-13；**接受日期：** 2021-06-16
联系方式： 辛志奇 E-mail：xzqlucky950417@163.com；*通信作者 路铁刚 E-mail：lutiegang@caas.cn

多组学数据结合，构建遗传调控网络，实现对作物表型的精准预测；深度学习指导基因编辑，实现对作物表型的控制和设计；深度学习在合成生物学的应用会使作物的设计育种更加自由高效。

1 人工智能及分支

人工智能这一概念最早在20世纪40年代被提出，但是受计算能力的限制，人工智能领域一直处于发展的低谷。进入21世纪后，计算机性能的大幅提升（尤其是GPU的发展）使得人工智能领域重新回到人们的视野。目前，人工智能已被应用于多个领域。

1.1 机器学习

人工智能领域最主要的研究方法是机器学习，机器学习按学习形式可以分为监督学习和无监督学习两种。监督学习是指在训练实例中学习输入变量数据和其因变量（或叫标签）之间的关系，然后以此在新实例中预测结果，主要应用于回归和分类问题。例如，可以用大量历史气象数据训练机器学习模型，该模型可以以过去的天气数据为预测因子，预测未来的天气。如果预测的目标变量为离散变量，则该机器学习任务称为分类问题（classification）；如果预测的目标变量为连续变量，则该机器学习任务称为回归问题（regression）。在机器学习中有很多监督学习算法及应用，例如结合统计学的隐马尔可夫模型（hidden Markov model, HMM）和机器学习的支持向量机（support vector machine, SVM）可以快速准确预测和区分DNA和RNA结合残基的方法，这有利于进一步预测蛋白质-DNA和蛋白质-RNA相互作用的序列[3-4]；用随机森林和支持向量机模型通过DNA甲基化数据精确诊断癌症[5]。无监督学习是指在训练实例中输入没有因变量（或标签）的数据，又称为归纳性学习，典型的无监督学习包括降维（dimensionality reduction）和聚类（clustering），适合学习高维度数据，例如组学数据[6-7]。

1.2 深度学习

深度学习是机器学习领域的一个相对年轻的分支，已经成为机器学习领域最流行和最强大的技术之一[8]。人工神经网络以数学模型模拟神经元活动，包括输入层、隐藏层和输出层三个部分（图1），其深度神经网络用多层的隐藏层使神经网络的性能大幅提高，同时需要的计算能力和数据量也大幅提升。

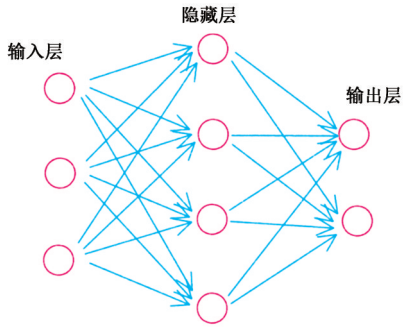

图1 人工神经网络层次

Fig1 Artificial neural network

卷积神经网络（convolutional neural network, CNN）是深度神经网络的一种，也是基础的深度学习模型，用卷积这一数学计算方式提取数据中的特征信息，再经深度神经网络处理，可以大大增加神经网络的性能。卷积神经网络更擅长提取结构信息。目前已经有很多利用CNN解决基因组学问题的例子。例如，Babak等[4]预测DNA和RNA与蛋白质的结合位点，Hashemifar等[9]预测蛋白质之间相互作用；Gao等[10]基于基因序列预测poly(A)位点；Zhou等[11]预测了人类基因组变异对基因表达调控和疾病的影响；Zhou和Wang等[12-13]预测了非编码基因突变的影响；Jost等[14]结合CRISPR技术实现调控基因表达量变化等。另一种监督学习模型，循环神经网络（neutral network, RNN）加入时间步（timestep）概念，使其具有记忆性和参数共享的特点，适合处理有时间信息的数据，广泛应用于自然语言处理领域。在生物学领域常被用来预测序列的功能。例如，Shen等[16]结合RNN和k-mer[15]预测转录因子识别位点；Li等[17]利用CNN和RNN从氨基酸序列预测酶的生化功能；Quang等[18]利用RNN和CNN预测非编码基因的功能等。值得注意的是，有报道指出，CNN在提取特征方面更高效，而释义DNA序列方面，来自自然语言处理领域的k-mer方法显得比CNN和RNN更优秀[19]。

自编码器（auto-encoder）是深度学习中的无监督学习的重要组成部分。自编码器分为编码和解码两部分。编码部分负责将输入数据低维化处理，也可以理解为特征提取；解码部分负责将编码

得到的结果恢复到原始输入的形式，它是理解复杂深度学习模型的关键，可以把数据中的关键数据提炼并展现出来，解决了深度学习模型训练过程的不可见问题。目前自编码器在图像识别、降噪、色彩化方面有广泛应用。Zhang等[7]用自编码器整合多组学数据，有效缓解了生物领域在运用人工智能模型时出现的“少样本，高维度特征”的问题；用自编码器解码深度学习模型并结合全基因组关联分析（genome wide association study，GWAS）的技术观察到未分类的基因在深度学习模型的不同深度中被有序的分类[20]。

生成模型技术作为深度学习领域的重要分支，它既不属于监督学习也不属于无监督学习。主要包括生成式对抗网络（generative adversarial network，GAN）和变分自动编码器（encoder）两种模型。

生成式对抗网络[21]分别建立并训练生成模块和判别模块，将生成模块生产的伪数据交由判别模块判断真伪，通过这种对抗学习的方式进行训练，可以生成真实度高的数据。目前在生物医药方面已经有相关的文章报道：基于生成式对抗网络设计蛋白酶抑制剂[22]；RamaNet模型从头设计合成螺线蛋白骨架[23]；基于生成式对抗网络设计合成大肠杆菌启动子序列[24]。

变分自动编码器[25]与生成式对抗网络同属生成模型家族成员，两种模型都致力于生成更接近真实的数据，但是二者的实现思路不同。变分自动编码器在结构方面与自动编码器有相似之处，也是由编码器和解码器组成（也被称作识别模型和生成模型），并且二者都是学习输入数据的潜在向量并试图重建输入数据。不同的是，变分自动编码器学习潜在向量的分布关系，在潜在空间中是连续的，再由生成模型构建输入数据；生成式对抗网络由生成器和判别器组成，生成器负责创造数据，而判别器负责评价生成器创造的数据是否能够以假乱真。Davidsen等[26]用变分自动编码器模型生成T细胞受体的蛋白质序列。

2 深度学习在作物基因组学中的应用

目前人工智能在农业上应用的报道主要是对图像和视频进行识别，如对玉米照片进行识别和对玉米干旱胁迫下的表型进行分类[27]；视频检测植物生长早期干旱胁迫[28]；视频识别水稻虫害和病害[29]；以拟南芥为例基于植物图像对植物表型分类[30-31]等。生物的遗传信息是沿着中心法则传递的，想对植物基因进行设计，表型精准预测就一定要对基因组、转录组、蛋白质组、表观遗传组甚至是代谢组规则有更深的认识。近年来，在基因组学领域，围绕各种分子表型发展出了一系列基于二代测序的高通量技术，如转录组技术、开放染色质分析技术、DNA-转录因子互作技术[32]等。深度学习技术可以对这些大规模数据集进行建模。

2.1 深度学习模型建立的过程

深度学习模型建立首先遇到的一个问题就是生物学数据该以何种形式输入到人工智能模型中，这个问题在基因组和转录组已经有了统一的答案。One-hot编码方式可以高效地将基因组和转录组数据储存在电脑中作为输入数据。将基因的A、T、G、C 4种碱基储存在一个4×N的矩阵中，每一列只储存1个碱基（图2），这个方法可以将N bp的基因数据输入模型。

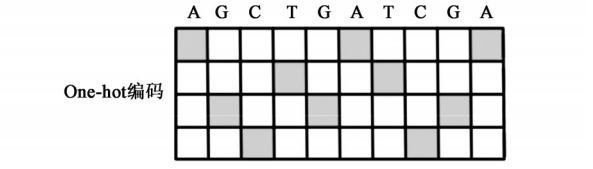

图2 One-hot编码

Fig.2 One-hot encoding

当建立机器学习模型时，观测数据通常被随机分为训练集（用于训练模型）、验证集（用于确定模型结构和超参数），以及测试集（用于评估模型的性能）。这种随机划分能够避免数据间存在规律性特征而得出准确率虚高的模型。训练集/测试集的划分应尽量保持数据分布的一致性，避免混杂因素（confounder）对最终结果的影响。最常用的训练集/测试集分割方法为交叉验证法。在训练集上的准确度高于在测试集上的准确度，这种现象被称为过拟合（over-fitting）。有几种情况会导致过拟合。一个通常出现的问题是特征空间中的维度有时大大超过观测值。例如，当从基因组变体预测一个表型时，检测到的基因组单核苷酸SNP数目几乎总是超过植物基因型的数目。在这种情况下，可以使用主成分分析(principal component analysis,PCA)和自动编码器[11-12]等降维技术

来减少特征的数目。然而，在处理基因组学中的问题时，过拟合有时候是隐藏的。例如，当一个基因家族的成员被划分为训练集和验证、测试集时，模型将学习家族特异性的分子特征，并高估预测准确性。

2.2 利用深度学习技术预测生物学序列

各个组学数据都有被人工智能挖掘有用信息的巨大潜力。在DNA层次上，Umarov等[33]利用CNN构建了启动子的预测模型，分析了几种原核和真核生物的启动子序列特征，包括人、老鼠、植物（拟南芥）和细菌（大肠杆菌和枯草芽孢杆菌）。DanQ是一种将CNN和双向长短期记忆循环神经网络（BLSTM）相结合的混合框架，用于从头预测非编码区的功能。DanQ学习了一种调节语法来改善预测准确性，并为非编码基因组区域提供了新的见解[18]。DanQ还结合CNN和BLSTM在序列中从头预测非编码区功能[18]。Sample等[34]使用CNN和遗传算法精准预测了人类5'UTR变体对核糖体装载的影响。

在RNA水平上，使用循环神经网络（neutral network，RNN）在人类mRNA和lncRNA序列上训练了一个门控RNN，然后用它来预测RNA分子是否编码蛋白质[35]。使用CNN预测人类5'UTR变异对核糖体装载的影响[34]。他们将28万个随机的5'UTR的多聚体分析与深度学习相结合，建立了一个模型，从人类5'UTR序列预测翻译效率。此外，DeepChrome是一个从组蛋白修饰数据预测基因表达量的CNN，能够自动提取重要特征之间的复杂交互作用[36]。为了预测组织特异性的基因表达，研究人员将CNN与空间特征变换和L2正则化线性模型相结合，建立了ExPecto模型[37]。

在蛋白质水平上，为了在从头生成的肽序列中提取重要的氨基酸特征，利用CNN方法开发了DeepNovo [38]。为了预测蛋白质的二级结构，使用了相对溶剂可及性和残基间接触映射数据训练了深度学习模型rawMSA[39]。最近，谷歌的AlphaFold利用深度学习模型预测蛋白质的三级结构，其精确度远超传统机器学习方法[40]。此外，深度学习模型也用来预测蛋白质—蛋白质的相互作用。DPPI是一种能够从蛋白质序列信息预测蛋白相互作用和蛋白二聚体的深度学习模型[41]。DEEPre可以从蛋白质序列预测酶的类别，利用该模型可以发掘在宏基因组、工业生物技术和人类疾病中起重要功能的蛋白质[42]。

除了用各组学数据分别预测之外，Ma等[7]将各组学数据整合，使生物学数据更立体，与表型相关的信息也会更丰富准确，同时也会有效缓解人工智能与生物学结合领域一直存在的问题，即生物学"数据特征维度高但样本少"的问题，Ma等[7]也指出这样做的难点在于各组学数据的信息不均匀。

3 深度学习在育种4.0中的应用

作物自然群体中存在着海量的自然变异，其中能够影响作物表型的变异称为功能变异。功能变异位点的不同等位变异具有不同的表型效应，可以划分为有利等位变异和有害等位变异。作物育种很大程度上可以视为有利等位变异的富集（也可以从另一个方面看做有害等位变异的清除）。过去的30年被概括为育种3.0时代，在这一历史阶段，获取高通量基因型数据和表型数据的成本不断降低，同时通过关联分析和连锁分析克隆了大量控制重要农艺性状的关键位点。以此为基础，分子标记辅助选择技术、基因组预测技术在作物育种中逐渐成为常规技术。未来我们将进入一个新的育种历史阶段：育种4.0。在这一阶段，人工智能将主要从三个方面促进设计育种发展：①发掘功能变异，指导精准杂交育种。通过各生物组学数据和环境数据预测出作物的产量和表型性状，从而实现简单化精准化的预测作物复杂优良性状。②设计有利等位变异，指导基因编辑育种。从基因水平、转录水平，以人工智能模型指导基因编辑，进一步细致调控基因表达，从而改良性状。③设计具有特定功能的基因组元件，指导合成生物学。创造新的DNA元素、基因，甚至具有某种特定功能的调控通路，并将其应用于作物育种。

目前大多数研究都聚焦于人工智能进行分类和回归的能力。Wang等[19]的文章中提到人工智能的生成模型可以通过学习生成新的基因元件从而应用于合成生物学。生成模型技术与合成生物学结合，根据预测模型的指导，重新设计非自然的基因、蛋白质等应用已经被报道。如深度学习指导编辑gRNA实现基因表达量的调控[14]；结合生成式对抗网络设计大肠杆菌基因启动子序列[24]；设

计蛋白质序列以拓展蛋白质空间[43]；设计螺线蛋白质骨架[23]；生成T细胞受体的蛋白质序列[26]等。

深度学习模型存在迁移学习的性质，即可以用某一物种训练的预测模型预测相近物种，这种性质使得生物学中单一物种训练的模型有了更广泛的用处，如小鼠基因组训练的模型可以用在人类基因组上[50]，单一植物叶片胁迫表型的识别模型可以用来预测其他植物的叶片胁迫表现[44]。

4 展望

人工智能特别是深度学习出现之后，已经在多个领域掀起新的浪潮，现阶段已经在基因组学、转录组学、蛋白质组学和合成生物学等领域发挥了巨大作用，如完善基因组功能注释、挖掘新功能基因、预测植物表型、发现基因、RNA、蛋白质等物质的新分类模式，指导基因编辑。高通量技术的发展见证着植物基因组学的进步，它以较低的花费识别着多种分子表型。然而，基因组学也要求利用强大的数据挖掘工具来预测和解释这些分子表型，深度学习则可以预测任何基因组变异的分子表型效应，获得直接控制分子表型的功能变异。此外，在合成生物学中应用深度学习模型也有望创造具有理想功能的新基因。总之，深度学习在未来植物基因组学研究和作物遗传改良中将发挥中心作用，人工智能将会是未来农业发展不可或缺的一部分。

参 考 文 献

[1] 王向峰,才卓. 中国种业科技创新的智能时代——"玉米育种4.0"[J]. 玉米科学, 2019, 27(01): 1-9.

[2] WALLACE J G, RODGERS-MELNICK E, BUCKLER E S. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics[J]. Annu. Rev. Genet., 2018, 52: 421-444.

[3] YAN J, KURGAN L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues[J/OL]. Nucl. Acids Res., 2017, doi:10.1093/nar/gkx059[2021-07-10]. https://doi.org/10.1093/nar/gkx059.

[4] BABAK A, ANDREW D, MATTHEW T W, *et al*.. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning[J/OL]. Nat. Biotechnol., 2015, 33(8): 831 [2021-07-10].https://doi.org/10.1038/nbt.3300.

[5] MAROS M E, CAPPER D, JONES D T W, *et al*.. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data[J]. Nat. Protoc., 2020, 15(2): 479-512.

[6] XU C, SCOTT A J. Machine learning and complex biological data[J/OL]. Genome Biol., 2019, 20(1):76[2021-07-10]. https://doi.org/10.1186/s13059-019-1689-0.

[7] MA T, ZHANG A. Integrate multi-omics data with biological interaction networks using multi-view factorization autoEncoder (MAE)[J]. BMC Genom., 2019, 20(S11): 944[2021-07-10]. https://doi.org/10.1186/s12864-019-6285-x.

[8] HINTON G E,SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313 (5786): 504-507.

[9] HASHEMIFAR S, NEYSHABUR B, KHAN A A, *et al*.. Predicting protein-protein interactions through sequence-based deep learning[J]. Bioinformatics, 2018, 34(17): 802-810.

[10] GAO X, ZHANG J, WEI Z, *et al*.. DeepPolyA: a convolutional neural network approach for polyadenylation site prediction[J]. IEEE Access, 2018, 6: 24340-24349.

[11] ZHOU J, THEESFELD C L, YAO K, *et al*.. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk[J]. Nat. Genet., 2018, 50(8): 1171-1179.

[12] ZHOU J, TROYANSKAYA O G. Predicting effects of noncoding variants with deep learning-based sequence model[J]. Nat. Methods, 2015, 12(10): 931-934.

[13] WANG M, TAI C, E W, *et al*.. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional noncoding variants[J]. Nucl. Acids Res., 2018, 46(11): e69[2021-07-10]. https://doi.org/10.1093/nar/gky215.

[14] JOST M, SANTOS D A, SAUNDERS R A, *et al*.. Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs[J]. Nat. Biotechnol., 2020, 38: 355-364.

[15] MARÍA K M, EDWARD S B. A k-mer grammar analysis to uncover maize regulatory architecture[J/OL]. BMC Plant Biol., 2019, 19(1):103[2021-07-10]. https://doi.org/10.1186/s12870-019-1693-2.

[16] SHEN Z, BAO W, HUANG D S. Recurrent neural network for predicting transcription factor binding sites[J]. Sci. Rep., 2018, 8(1): 15270[2021-07-10]. https://doi. org/10.1038/s41598-018-33321-1.

[17] LI Y, WANG S, UMAROV R, *et al*.. DEEPre: sequence-based enzyme EC number prediction by deep learning[J]. Bioinformatics, 2018, 34(5): 760-769.

[18] QUANG D, XIE X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences[J/OL]. Nucl. Acids Res., 2016, 44(11): e107[2021-07-10]. https://doi.org/10.1101/032821.

[19] WANG H, CIMEN E, SINGH N, *et al*.. Deep learning for plant genomics and crop improvement[J]. Curr. Opin. Plant Biol., 2020, 54: 34-41.

[20] DWIVEDI S K, TJARNBERG A, TEGNER J, *et al*.. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder[J]. Nat. Commun., 2020, 11 (1): 856[2021-07-10]. https://doi. org/10.1038/s41467-020-14666-6.

[21] 王坤峰,苟超,段艳杰,等. 生成式对抗网络GAN的研究进展与展望[J]. 自动化学报, 2017, 43(03): 321-332.

[22] ALEX Z, VLADIMIR A, ALEXANDER Z, *et al*.. Potential 2019-nCoV 3C-like protease inhibitors designed using generative deep learning approaches[J/OL]. Theor. Comp. Chem., 2020,doi:10.26434/chemrxiv.11829102.v2[2021-07-10]. https://doi.org/10.26434/chemrxiv.11829102.v2.

[23] SARI S,MIKHAIL M. RamaNet: computational *de novo* helical protein backbone design using a long short-term memory generative adversarial neural network[J/OL]. bioRxiv, 2019, doi: 10.12688/f1000research. 22907.1[2021-07-10]. https://doi. org/10.12688/f1000research.22907.1.

[24] WANG Y, WANG H C, WEI L, *et al*.. Synthetic promoter design in *Escherichia coli* based on generative adversarial network[J/OL]. bioRxiv, 2019, doi: 10.1093/nar/gkaa325[2021-07-10]. https://doi.org/10.1093/nar/gkaa325.

[25] CARL D. Tutorial on variational autoencoders[J/OL]. arXiv, 1606[2021-07-10]. https://arxiv.org/pdf/1606. 05908v3.pdf.

[26] DAVIDSEN K, OLSON B J, DEWITT W S, *et al*.. Deep generative models for T cell receptor protein sequences[J/OL]. Elife, 2019, 8: e46935[2021-07-10]. https://doi. org/10.7554/elife. 46935.

[27] AN J Y,LI W Y, LI M S,*et al*.. Identification and classification of maize drought stress using deep convolutional neural network[J/OL]. Symmetry, 2019, 11(2): 256[2021-07-10]. https://doi.org/10.3390/sym11020256.

[28] LI H H, YIN Z Z, MANLEY P, *et al*.. Early drought plant stress detection with bi-directional long-term memory networks [J]. Photogram. Engin. Remote Sens., 2018, 84(7): 459-468.

[29] LI D, WANG R, XIE C, *et al*.. A recognition method for rice plant diseases and pests video detection based on deep convolutional neural network[J/OL]. Sensors (Basel), 2020, 20(3): 578[2021-07-10].https://doi.org/10.3390/s20030578.

[30] UBBENS J R, STAVNESS I. Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks[J/OL]. Front. Plant Sci., 2017, 8: 1190[2021-07-10]. https://doi.org/10.3389/fpls.2017.01190.

[31] JORDAN R U,IAN S. Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks[J/OL]. Front. Plant Sci., 2017, 8(1190): 1190[2021-07-10]. https:// doi. org/10.3389/fpls.2017.01190.

[32] ZAMPIERI G, VIJAYAKUMAR S, YANESKE E, *et al*.. Machine and deep learning meet genome-scale metabolic modeling[J/OL]. PLoS Comp. Biol., 2019, 15(7):e1007084[2021-07-10]. https://doi.org/10.1371/journal.pcbi.1007084.

[33] UMAROV R K, SOLOVYEV V V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks[J/OL]. PLoS ONE, 2017, 12(2): 0171410 [2021-07-10]. https://doi.org/10.1371/journal.pone.0171410.

[34] SAMPLE P J, WANG B, REID D W,*et al*.. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay[J/OL]. Nat. Biotechnol., 2019, 37(7): 803 [2021-07-10]. https://doi.org/ 10.1038/s41587-019-0164-5.

[35] STEVEN T H,RACHAEL K,AMY T, *et al*.. A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential[J]. Nucl. Acids Res., 2018, 46 (16): 8105-8113.

[36] SINGH R, LANCHANTIN J, ROBINS G,*et al*.. DeepChrome: deep-learning for predicting gene expression from histone modifications[J]. Bioinformatics, 2016, 32(17): 639-648.

[37] ZHOU J, THEESFELD C L, YAO K,*et al*.. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk[J]. Nat. Genet., 2018, 50(8): 1171.

[38] TRAN N H, ZHANG X L L, XIN L,*et al*.. De novo peptide sequencing by deep learning[J]. Proc. Natl. Acad. Sci. USA, 2017, 114(31): 8247-8252.

[39] CLAUDIO M,BJÖRN W. rawMSA: end-to-end deep learning using raw multiple sequence alignments[J/OL]. PLoS ONE, 2019, 14(8): 0220182[2021-07-10]. https://doi. org/ 10.1371/journal.pone.0220182.

[40] EVANS R, JUMPER J, KIRKPATRICK J,*et al*.. De novo structure prediction with deeplearning based scoring[J]. Annu. Rev. Biochem., 2018, 77:363-382.

[41] HASHEMIFAR S,NEYSHABUR B, KHAN A A,*et al*.. Predicting protein-protein interactions through sequence-based deep learning[J]. Bioinformatics, 2018, 34(17): 802-810.

[42] LI Y, WANG S, UMAROV R,*et al*.. DEEPre: sequence-based enzyme EC number prediction by deep learning[J]. Bioinformatics, 2018, 34(5): 760-769.

[43] DONATAS R,VYKINTAS J,LAURYNAS K, *et al*.. Expanding functional protein sequence space using generative adversarial networks[J/OL]. bioRxiv, 2019, doi: 10.1038/s42256-021-00310-5[2021-07-10]. https://doi. org/10.1038/s42256-021-00310-5.

[44] DAVID R K. Cross-species regulatory sequence activity prediction[J/OL]. bioRxiv, 2019, doi: 10.1101/660563[2021-07-10]. https://doi.org/10.1101/660563.

[45] SAMBUDDHA G,DAVID B,ASHEESH K S, *et al*.. An explainable deep machine vision framework for plant stress phenotyping[J]. Proc. Natl. Acad. Sci. USA, 2018, 115(18): 4613-4618.