科学通报 2021 年 第 66 卷 第 3 期: 284 ~ 299
定量工程生物学专辑 评 述
《中国科学》杂志社 SCIENCE CHINA PRESS

面向合成生物学的机器学习方法及应用

胡如云[1†], 张嵩亚[2,3,4†], 蒙海林[2,5], 余函[2,3,4], 张建志[2,3,4], 罗小舟[2,3,4], 司同[2,3,4*], 刘陈立[2,3,4*], 乔宇[1*]

1. 中国科学院深圳先进技术研究院先进计算与数字工程研究所, 深圳 518055;
2. 中国科学院深圳先进技术研究院合成生物学研究所, 深圳 518055;
3. 深圳合成生物学创新研究院, 深圳 518055;
4. 中国科学院定量工程生物学重点实验室, 深圳 518055;
5. 广州中国科学院先进技术研究所, 生物工程研究中心, 广州 511458

† 同等贡献

* 联系人, E-mail: tong.si@siat.ac.cn; cl.liu@siat.ac.cn; yqiao@siat.ac.cn

2020-04-24 收稿, 2020-07-01 修回, 2020-07-02 接受, 2020-07-03 网络版发表
深圳市科技创新委员会项目(KQTD20150331172101 53)资助

摘要 机器学习的目标是设计可以根据先验知识和观测数据不断改进其性能的算法. 该算法可以帮助机器从大量的数据中提取知识, 从而提升其在特定任务上的性能. 作为数据驱动的方法, 机器学习可以有效利用高通量实验技术产生的大批量生物数据, 实现合成生物体的功能预测与智能化设计, 改变合成生物学的研究范式. 本文首先介绍机器学习在合成生物学领域广泛应用的几个模型及方法, 如支持向量机、神经网络、生成式对抗网络、深度强化学习等. 然后介绍机器学习方法在合成生物学领域的典型应用, 如启动子预测、酶催化设计、代谢途径构建、基因线路设计等. 本文综述面向合成生物学的机器学习方法及应用, 并试图启发读者如何选择和设计机器学习方法用于合成生物学的研究.

关键词 机器学习, 合成生物学, 生物元件设计, 生物网络设计

合成生物学采用工程化设计理念, 对生物体进行有目标的改造乃至从头设计并合成, 创建赋予非自然功能的“人造生命”. 目前, “设计-构建-测试”的闭环研究策略在合成生物研究中的应用, 主要关注“构建”和“测试”的实验试错环节, 存在效率低、设计非最优等弊端, 因此需要有机结合机器学习手段, 配合海量工程化试错来加速和优化生物体设计. 机器学习可以充分利用生物学实验产生的高通量数据, 如基因组数据、蛋白组数据、生物网络数据等, 构建从序列(包括DNA序列、RNA序列和氨基酸序列)到功能元件(蛋白质)的关系, 重构或设计高效的生物网络, 如调控网络、代谢网络等, 从而帮助人们从系统角度认识生物体的组成元件、行为和控制机理, 综合解析生物体的代谢与变化规律, 辅助规划并优化合成生物体的设计方案.

机器学习是一种数据驱动的模式识别方法, 尤其是深度学习. 学习本身基于一定的数学规则和统计假设. 机器学习的总体目标是从给定的数据集中利用特征之间的统计关联性学习判别式模型或生成式模型. 学得的模型可以预测如类别标签、聚类类别、连续实值等输出, 或生成如DNA序列、氨基酸序列、代谢途径等合成生物对象. 对于基因注释问题, 机器学习首先试图识别已标记基因组的关键特征, 如转录起始位点的构成、基因编码区C+G内容和密码子. 然后, 利用学得的特征在训练集上构建基因发现的预测模型. 最后,

引用格式: 胡如云, 张嵩亚, 蒙海林, 等. 面向合成生物学的机器学习方法及应用. 科学通报, 2021, 66: 284–299
Hu R Y, Zhang S Y, Meng H L, et al. Machine learning for synthetic biology: Methods and applications (in Chinese). Chin Sci Bull, 2021, 66: 284–299, doi: 10.1360/TB-2020-0456

© 2020《中国科学》杂志社

将学得的模型应用到新测序的基因组序列，标记其中的基因序列. 对于DNA序列生成问题，机器学习算法试图从天然DNA序列数据中学习到DNA序列数据的分布, 从而对学到的模型采样获得新设计的DNA序列.

机器学习在生物中的应用极为广泛，不止基因注释[1~3]，还有基因调控元件识别[4]、蛋白质结构[5]和功能预测[6]、蛋白质定向进化[7]、代谢途径构建[8]等. 总地来说，凡是可以从数据中提取模式并应用到新数据集的任务都可以应用机器学习. 在生物大数据时代, 机器学习可用于挖掘生物信息、发现生物机理、辅助生物元件和生物网络设计等.

本文主要面向拥有生物学背景希望利用机器学习技术的入门者，或者拥有机器学习背景希望将其运用在生物学领域的入门者. 因此，本文主要分为两部分: 机器学习和合成生物学应用. 首先介绍机器学习的基本概念、工作流程和分类，然后介绍生物学领域常用的一些机器学习模型的基本思想和典型应用. 之后定义合成生物学任务，并举例说明机器学习算法的应用. 最后总结并展望合成生物学与机器学习技术结合的发展方向.

1 机器学习

1.1 机器学习概述

机器学习利用已设计的算法在给定的数据集上学习判别式或生成式模型. 输入数据由多个样本组成, 每个样本通常由特征和标签构成. 然而, 数据集也可以不包含标签. 机器学习方法的构建过程分为三步. 下面以转录因子结合DNA序列的特异性[9]为例来说明. 第一步, 由机器学习算法研究人员提出有效的算法. 该算法实现机器学习模型及模型参数的估计. 第二步, 给算法提供蛋白质结合微阵列(protein binding microarray, PBM)数据，包括基因芯片上的基因序列及对应的亲和度. 一个基因序列-亲和度对为一个样本, 基因序列上的碱基为特征, 亲和度为标签. 算法处理PBM数据得到模型及其参数的估计值并保存. 第三步, 将未给定亲和度的基因序列提供给算法, 令其计算亲和度. 若所提供基因的亲和度已知, 则可立刻评估算法的性能; 否则, 算法的性能需要通过实验测定的结果进行评估，在合成生物学领域通常如此. 这三步分别为算法的设计、训练和测试. 设计-训练-测试过程提供了对机器学习算法的假设检验: 该算法能否准确预测亲和度. 另外, 该算法本身可以产生一个新的假设: 给定新的基因序列输出的亲和度值. 从这个角度来讲, 如何解释这个输出值至关重要, 尤其对于生物大数据分析, 我们需要试图从中获得生物学机理的启发.

机器学习发展至今, 已经提出了大量的算法, 从学习任务角度, 分为分类、回归、聚类. 分类是根据给定的样本, 由算法预测样本所属的类别, 且类别标签已知, 如对基因序列中转录因子起始位点的识别[10]. 回归是由给定的样本, 计算样本对应的一个数值, 如对蛋白质结合基因序列亲和度的预测[9]. 聚类是将训练样本集根据样本间的相似程度分割成不同的部分，且无事先指定的类别标签, 如基因表达数据的聚类[11].

从学习目标角度，机器学习分为判别式和生成式. 生成式方法[12,13]试图学习数据的分布，或学习变量的联合分布$p(x_1, \cdots, x_n)$, 关注模型对数据本身的解释, 任务迁移性好. 而判别式方法[9,14]则尽可能精确匹配学习任务, 或学习变量的条件分布$p(y|x_1, \cdots, x_n)$, 迁移性差但准确性高. 在转录因子结合位点识别的任务中, 生成式方法试图理解“序列的什么特性是识别其为结合位点的重要因素”，如位置权重矩阵；而判别式方法则只关注如何尽可能地提高识别准确度, 如支持向量机.

从训练数据角度, 根据数据标签的有无, 机器学习分为监督学习、无监督学习和半监督学习. 监督学习所有用于训练的数据都带数据标签，基因序列转录因子起始位点的识别[10]利用基因序列是否涵盖起始位点的标签训练识别模型. 无监督学习所有用于训练的数据都无标签, 如基因表达的层次聚合聚类[11]. 半监督学习利用部分带标签的数据训练模型，从无标签数据中学习数据分布, 并根据带标签数据适配学习任务[15].

从数据类型角度，机器学习模型分为结构模型和非结构模型. 结构化数据包括特征列表、序列和数组, 如k-mer特征[16]、DNA序列[9]、生物图像[17]. 非结构数据包括三角网格、点云和图，如生物网络[18]、基因表达图[19]等.

从模型复杂度角度，机器学习分为经典机器学习和深度学习[20]. 现实世界的一些(对人来说)简单问题很难用形式化的数学规则来描述，比如识别照片中的人脸等. 然而，这些直观的问题，可以通过层次化的概念体系来描述. 每个复杂概念由相对简单的概念之间的关系来定义, 由此形成一张很“深”的网络, 然后从训练数据学习这些概念及它们之间的关系. 这种机器学习范式即为深度学习，如深度神经网络[21]、深度信念

网络[22]和深度玻尔兹曼机[23]等.

1.2 经典机器学习模型

1.2.1 线性最小二乘回归

线性最小二乘回归(linear least square regression)是最简单的一种机器学习模型. 它从训练样本集$D=\{(x_i,y_i)\}$中学习线性映射$y=\mathbf{a}x+b$, 其中$\mathbf{a}_j$和b是待定的模型参数, x_i是数据的输入特征, y_i是对应的值. 若模型参数已知, 则可以预测给定输入特征下的值. 模型参数的学习通过最小化训练集到直线的距离, 此最小化的解为最小二乘解. 模型本身是线性的, 不能直接对非线性问题建模. 对于非线性问题, 需要对原始数据进行特征组合或者特征变换. 对共线性问题, 需要对原始数据进行降维, 相关方法如主成分回归(principle component regression, PCR)和偏最小二乘回归(partial least square regression, PLSR)[24]. 该模型直观、可解释性强, 系数$\mathbf{a}$的大小表示特征$\boldsymbol{x}$对结果y的重要性.

1.2.2 支持向量机

支持向量机(support vector machine, SVM)[25], 是一种监督、二分类的机器学习模型. 基本思想: 给定训练样本集$D=\{(\boldsymbol{x}_i,y_i)\}$, $y_i\in\{-1,+1\}$, 在样本空间中寻找一个划分超平面$\omega^{\mathrm{T}}\boldsymbol{x}+b=0$, 将两类样本分开, 如图1所示. 显然, 这样的划分超平面不止一个, 为提高模型的泛化能力, 该划分应该使得两个类别之间的"距离"最远. 该距离为两个异类支持向量——到超平面最近的几个样本——到超平面的距离之和$\gamma=\frac{2}{\|\omega\|}$, 亦称为"间隔". 因此, 支持向量机的学习是在正确分类的条件下, 最大化间隔. 此约束最大化问题是一个凸二次规划问题, 可由拉格朗日乘子法求解. 该模型训练完成后, 只与少数的样本(支持向量)有关, 模型稀疏, 可解释性强.

由于模型寻找的是划分超平面, 若样本非线性可分, 则无法用一个超平面将两类样本分开. 此时, 可以利用"核化"技巧, 对样本空间做非线性变换, 使得变换后的特征空间线性可分. 核化后的划分超平面为$f(\boldsymbol{x})=\sum_{i=1}^{m}\alpha_i y_i k(\boldsymbol{x},\boldsymbol{x}_i)+b=0$, 常用核函数如高斯核$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$.

支持向量机可用于DNA序列剪接位点预测[14]、DNA甲基化预测[26]、蛋白质结构预测[27]、癌症分类[28]等.

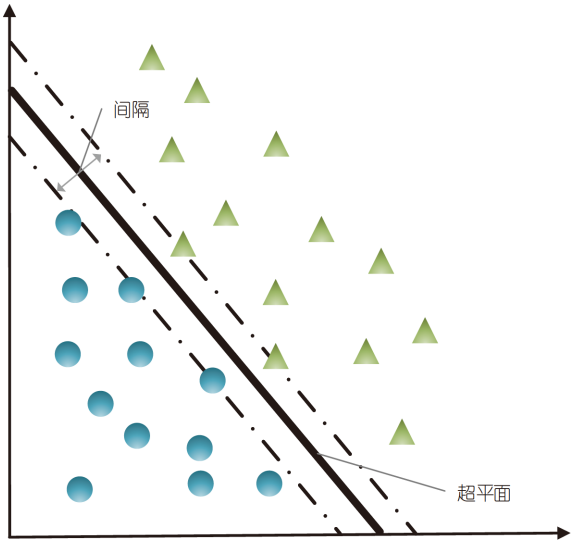


图 1 (网络版彩色)支持向量机原理示意图. 三角形代表正例, 圆圈代表负例. 位于点划线上的样本为支持向量, 两条点划线之间的距离为间隔, 实线代表划分超平面$\omega^{\mathrm{T}}\boldsymbol{x}+b=0$

Figure 1 (Color online) Illustration of support vector machine. Triangles stand for positive examples, and circles for negative. The support vectors are located on dash-dot lines, and margin is the distance between the dash-dot lines. The solid line is superplane, $\omega^{\mathrm{T}}\boldsymbol{x}+b=0$, for the division

1.2.3 层次聚合聚类

聚类(cluster)模型是一类非监督、聚类的机器学习模型. 基本思想: 给定训练样本集$D=\{x_i\}$, 依据样本点x_i之间的相异性(相似性), 将训练样本集分割成不同的类别, 保证类间差异最大、类内差异最小.

对于层次聚合聚类(hierarchical agglomerative cluster)[29,30], 首先将每个样本点视为一个类, 不断合并连接性(类间差异)最小的两个类, 直到所有样本点完全合并为一个类为止. 最终得到一颗层次聚类树, 在树的不同层截断就得到不同的聚类结果. 样本之间的相异性由某种距离, 如欧式距离$d_{ij}=d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2$度量, 从而得到相异性矩阵. 类$(I,J)$之间的连接性由某种类间距离, 如单连接$D(I,J)=\min_{i\in I,j\in J}\{d_{ij}\}$度量, 并依此进行类合并.

层次聚合聚类无须事先确定类数目, 但计算量大, 适合于不太大数据量的聚类. 层次聚合聚类可用于基因表达[11,31]和蛋白质-蛋白质相互作用[32].

1.2.4 贝叶斯网络

贝叶斯网络(Bayesian network, BN)[33]是一种有向无环的概率图模型. 概率图模型对变量间的联合概率

分布建模，而贝叶斯网络只对变量间的直接关系(条件概率)建模，是对变量间联合概率分布的一种紧致的表式．节点表示变量x_i，边表示变量之间的关系，边的权重由条件概率$p(x_i|pa_i)$给出，此概率表示在父节点pa_i条件下子节点x_i发生的可能性．贝叶斯网络的有向性和条件概率，代表了一种信念的传播，因此又叫信念网络(belief network)．

贝叶斯网络的学习包括网络结构的学习[34,35]和权重的学习[36,37]．其中结构学习常用方法是“评分搜索”，用评分函数评估贝叶斯网络与训练数据集的契合程度，评分函数通常基于信息论准则，如最小描述长度[38]．若网络结构已知，则权重的学习通过简单的训练样本“计数”即可．但在一些任务中，部分变量不可观测，即存在隐变量$\boldsymbol{z}$，此时常采用EM算法：首先假设网络参数已知，由可观测变量$\boldsymbol{x}$推断隐变量$\boldsymbol{z}$，即计算$p(\boldsymbol{z}|\boldsymbol{x})$，然后再最大化隐变量似然期望，更新网络参数．网络推断[12]是根据训练好的网络和证据变量的取值，计算查询变量的分布．精确推断——根据网络定义的联合概率分布精确计算后验概率——是“NP难”的，通常采用吉布斯采样进行近似推断．

贝叶斯网络的概率表示使得其能处理带噪声的数据．另外，网络还能对变量间的因果关系建模．贝叶斯网络可用于基因(表达)调控网络建模[12,39]、蛋白质-蛋白质相互作用建模[40]．

1.2.5 隐马尔可夫模型

隐马尔可夫模型(hidden Markov model, HMM)[41]，是结构最简单、含隐变量的动态贝叶斯网络．该模型由状态变量(隐变量)和可观测变量构成，不可观测状态和可观测状态均有限，如图2．状态之间的转移由状态转移矩阵决定，状态变量与可观测变量之间的关系由输出观测概率矩阵决定，初始状态由初始状态概率决定．

模型的3个基本问题是：(1) 已知模型，计算其产生某个观测序列的概率，即评估它们之间的匹配程度；(2) 已知模型和观测序列，推断出隐藏状态序列；(3) 已知观测序列，学习模型参数使得该序列出现的概率最大．第一个问题典型应用在基因预测或蛋白质数据库的搜索，第二个问题常用在序列对齐，如蛋白质序列对齐[42]，第三个问题典型用在蛋白质同源性建模[43]．

隐马尔可夫模型天然适用于序列建模，可应用于蛋白质同源性检测[44]、蛋白质结构预测[5]等．

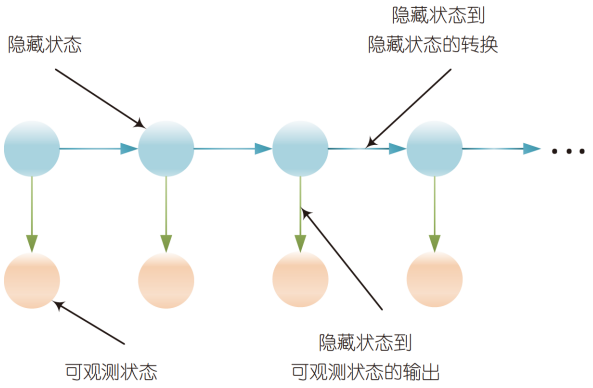


图 2 (网络版彩色)隐马尔可夫模型示意图．上排圆圈代表隐藏状态，下排圆圈代表可观测状态．横向箭头代表隐藏状态之间的概率转移，竖向箭头代表隐藏状态到可观测状态的概率输出

Figure 2 (Color online) Illustration of HMM. Circles in the upper row stand for hidden states, circles in the lower row for observable states. Horizontal arrows for state transition by probability, vertical arrows for output conditioned on hidden states

1.3 深度学习模型

1.3.1 多层感知机

多层感知机(multi-layer perceptron, MLP)，又称人工神经网络(artificial neural network, ANN)，是仿照神经元的工作机制设计的深度学习模型．最简单的多层感知机由3层构成，分别是输入层、隐藏层和输出层，如图3(a)所示．输入层接受输入数据，然后对输入做线性变换$\omega_i^{\mathrm{T}}\boldsymbol{x}$．为了向模型引入非线性表示能力，变换后的特征加上偏置$\boldsymbol{b}$再经过激活函数，如整流线性单元(rectified linear unit, ReLU)，得到隐藏特征$\boldsymbol{h}$=ReLU$(\omega_i^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b})$．最后对隐藏特征做线性变换，得到输出$y=\omega_o^{\mathrm{T}}\boldsymbol{h}$．

相较经典机器学习模型，多层感知机可以直接处理原始数据．无须手动的特征工程可以实现端到端的训练．隐藏层的加入实现了对原始数据的非线性特征变换．多层感知机的训练采用基于梯度的优化算法，如随机梯度下降．训练时根据任务选择不同的准则，分类任务通常采用极大似然准则，回归任务通常采用均方误差准则．

多层感知机是万能函数近似器，表示能力极强，其性能超过一些经典的机器学习模型，如线性回归．但对于许多问题，其预测能力通常不及基于梯度提升的决策树模型．虽然如此，多层感知机已成为深度学习工具的基础模块，可与其他神经网络层，如卷积层高效结合．

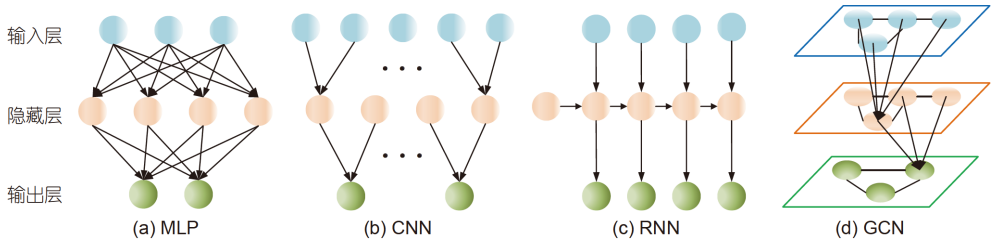


图 3 (网络版彩色)神经网络示意图. (a) 多层感知机; (b) 卷积神经网络; (c) 循环神经网络; (d) 图卷积神经网络. 圆圈代表特征, 箭头代表信息传播方向和参数

Figure 3 (Color online) Illustration of neural networks. (a) Multi-layer perceptron; (b) convolutional nerual net-work; (c) recurrent neural network; (d) graph convolutional neutral network. Circles stand for features, arrows for parameters and information flow

1.3.2 卷积神经网络

卷积神经网络(convolutional neural network, CNN)[21], 是一种包含广义卷积层作为隐藏层的人工神经网络. 广义卷积层通常包含卷积层、激活层、池化层.

(1) 卷积层(convolutional layer). 用卷积运算conv($\boldsymbol{x}$, *kernel*)代替全连接的线性变换$\omega^{\mathrm{T}}\boldsymbol{x}$, 如图3(b)所示. 卷积运算提取局部特征并引入了平移不变性. 卷积运算实现参数共享. 另外, 卷积核的大小远远小于输入数据的维度. 这使得卷积神经网络的存储和运算效率大大优于全连接的多层感知机. 一个卷积核可以从原始数据提取一个特征, 多个卷积核提取的特征组成一个卷积层.

(2) 激活层(activation layer). 向模型引入非线性, 将卷积层的输出作非线性变换. 增加模型的非线性表示能力.

(3) 池化层(pooling layer). 通过下采样(down sampling)降低数据维度, 再配合后续的特征提取操作, 实现对数据在不同尺度上的特征提取. 也是一种防止神经网络过拟合的手段.

广义卷积层作为卷积神经网络的特征提取层, 其输出特征经过全连接层(对特征进行放射变换)完成分类或回归任务.

DNA序列或蛋白质序列具有天然的局部特性, 卷积运算可以提取序列功能位点(motif), 有助于提高序列位点的预测准确性. 卷积神经网络广泛应用于合成生物学领域, 3个具有代表性的模型分别是DeepBind[9]、DeepSEA[45]和Basset[46].

1.3.3 循环神经网络

循环神经网络(recurrent neural network, RNN)[47,48], 是一类用于处理序列数据的神经网络. 卷积神经网络在空间共享模型参数: 序列每个位置的特征是其临近位置特征的线性组合, 每个位置共享组合系数. 循环神经网络在时间(序列)上共享参数: 序列t位置的特征(隐藏状态$\boldsymbol{h}^{(t)}$)是其前序特征$\boldsymbol{h}^{(t-1)}$和当前输入$\boldsymbol{x}^{(t)}$的函数$\boldsymbol{h}(t)=\tanh(\boldsymbol{W}\boldsymbol{h}^{(t-1)}+\boldsymbol{U}\boldsymbol{x}^{(t)}+\boldsymbol{b})$, 每个位置共享参数$\boldsymbol{W}$, $\boldsymbol{U}$和$\boldsymbol{b}$, 如图3(c)所示. 模型可以选择在每个位置上直接输出结果或者输出给后续网络. 这种参数共享方式引入了模型对前序特征的记忆功能, 理论上可以提取无限长序列的全局信息. 另外, 循环神经网络天然适合于处理多个不同长度的序列. 然而, 循环神经网络的序列化操作很难实现并行, 因此, 其计算速度远远低于卷积神经网络.

循环神经网络已被应用于聚合卷积神经网络预测的单细胞DNA甲基化状态[49]、RNA结合蛋白质[50]、转录因子结合位点和DNA可达性[51,52]中的输出.

1.3.4 图卷积神经网络

图卷积神经网络(graph convolutional neutral network, GCN)[53,54], 是卷积运算作用在图数据上的神经网络模型. 蛋白质-蛋白质相互作用网络和基因调控网络都属于图数据. 传统的卷积只能对结构化数据(如序列、图像)进行特征变换, 而图卷积则利用图节点特征或节点间连接信息来进行特征变换. 每层图卷积变换都利用相邻节点或边的信息, 生成新的图以表示一层新的特征, 如图3(d)所示.

图卷积神经网络的学习任务有三大类: 节点分类、边分类和图分类. 另外, 还有节点的嵌入表示. 图卷积神经网络已被应用于生物学和化学领域, 如从蛋白质-蛋白质相互作用网络预测组织功能[55], 在给定其他基因表达结果条件下, 预测待考察基因的表达情况[19]. 图卷积神经网络天然的适配图数据, 能从中学习有用的结构化模式, 在合成生物学的各种网络结构数据分析中将发挥重要作用.

1.3.5 变分自编码器

上述神经网络模型是典型的有监督模型. 下面讲述两类典型的无监督深度模型: 变分自编码器和生成式对抗网络. 这两类模型都是基于可微生成器网络(generator network)的模型. 可微生成器网络使用可微函数$g(\boldsymbol{z};\theta^{(g)})$将潜变量$\boldsymbol{z}$的样本变换成可观测变量$\boldsymbol{x}$样本或其分布, 可微函数通常由神经网络表示.

自编码器(autoencoder, AE), 将数据嵌入到低维的特征空间(编码), 然后再由低维特征(隐藏特征$\boldsymbol{z}$)重构原始输入数据(解码), 并利用原始数据和重构数据的差异作为自我监督的准则, 学习自编码器. 自编码器的重构过程可被看作对数据的去噪过程, 因为不重要的特征已经被自动忽略. 因此自编码器可以用于对数据进行去噪、缺失数据填补[56]、微阵列基因表达数据异常检测[57]等.

变分自编码器(variational autoencoder, VAE)[58], 对自编码器的隐藏特征附加额外的分布假设, 使得能从低维特征分布中采样隐藏特征, 然后经过重构网络(解码器)生成新的数据样本. 在训练期间, 可以通过嵌入网络(编码器)由训练样本推断低维特征. 训练的基本思想是: 最大化与样本相关(对数似然)的变分下界. 变分自编码器已被应用于学习单个细胞和批量RNA序列的低维表示[59]. 这些方法能显著改善单个细胞数据的可视化和聚类.

1.3.6 生成式对抗网络

生成式对抗网络(generative adversarial network, GAN)[60]的基本思想来源于博弈论场景, 其中生成器网络必须与对手竞争, 如图4所示. 生成器网络从随机特征生成样本$\boldsymbol{x}=g(\boldsymbol{z};\theta^{(g)})$, 其对手判别器网络(discriminator network)试图区分从训练数据抽取的样本和从生成器抽取的样本. 当训练收敛时, 生成器的样本和训练数据不可区分. 然后丢掉判别器网络, 只留下训练好的生成器网络. 生成式对抗网络已被用于生成蛋白质编码DNA序列[13], 以设计蛋白质结合微阵列DNA探针等.

1.4 现代机器学习方法

合成生物设计空间巨大, 如蛋白质序列空间至少为20^{110}, 随机试错难以解决如此巨大空间寻优问题. 强化学习由试错经验学习寻优策略, 已成为合成生物优化设计的重要方法. 另外, 机器学习模型需要大量的训练数据样本以获得精确的学习器. 然而, 在一些学习任务中, 很难获得足够的训练数据, 尤其是带标签的训练数据. 如何充分地利用有限的训练样本, 成为机器学习的一个重要问题, 特别是深度学习. 本文简单介绍合成生物学领域常用的3种机器学习方法及深度强化学习.

1.4.1 多任务学习

多任务学习(multi-task learning)为试图充分利用多个相关任务的有用信息来改善机器学习模型在所有任务中的泛化能力. 多任务学习方法[61]包括特征学习方法、特征子空间方法、任务聚类方法、任务关系学习方法和分解方法等. 每个方法采用不同的任务间信息共享方式, 例如特征学习方法假设所有任务共享所有隐藏特征, 而特征子空间方法假设所有任务共享部分隐藏特征. 多任务学习需要同时处理多个损失函数, 通常是将每个任务的损失函数加权求和. 因此, 训练多任务模型的一大挑战就是如何平衡每个学习任务. 已有一些策略来解决这个问题[62].

生物信息学数据通常包含相关生物活性的测量结

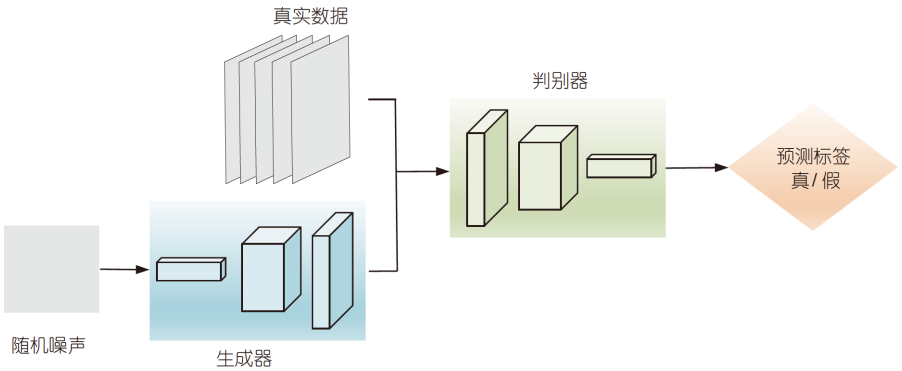


图 4 (网络版彩色)生成式对抗网络示意图. 随机噪声经过生成器合成图片, 与训练照片一起供判别器来区别"真实"与"合成"

Figure 4 (Color online) Illustration of GAN. Discriminator decides whether the input pictures, mixed with synthetic ones provided by generator, are "real" or "synthetic"

果, 例如存在多个活性的共调控基因. 多任务学习可以充分学习这样的共调控关系. 再如对多转录因子结合亲和度的预测问题, 多任务学习可以同时预测多个转录因子的结合. 模型共享大部分的特征层, 只有最后少数几层用于特定的任务. 这得益于转录因子的共结合特性和拥有相同的蛋白质结构域.

1.4.2 多视角学习

多视角学习(multi-view learning)[63]是充分利用任务不同侧面的信息来提高机器学习模型的泛化能力. 例如, 人可以通过脸、指纹、虹膜等不同的信息源来区别; 图像可以通过颜色或者纹理特征这两个不同的特征子集来表示; 同一个生物活性, 可以通过不同的实验测定, 由此得到不同的数据表示, 如ChIP-seq、微阵列等. 多视角学习可分为三类: 共学习——最大化两类数据的共性; 多核学习——学习不同视角各自的核然后将之集合起来; 子空间学习——学习多视角共享的潜变量子空间. 这些方法要么采取一致性原则学习, 要么采取互补性原则学习.

基于神经网络的方法首先将各视角的数据编码, 然后将各自特征拼接起来, 作为任务的输入特征. 例如将DNA序列、基因表达微阵列数据和染色体可达性数据结合起来预测所有细胞类型的转录因子结合性质[52].

1.4.3 迁移学习

迁移学习(transfer learning)[64]着眼于知识迁移问题. 在关注的领域数据量稀少, 而相关领域已有现成训练好的模型的情况下, 如何将已有模型迁移到当前领域是有现实意义的问题. 然而, 机器学习的主要假设就是训练数据和测试数据必须在同一特征空间且具有相同的分布, 因此相关领域间只能迁移部分信息. 可迁移的信息包括: 部分实例(数据重用)、特征表示(共性特征)、模型参数共性和变量关系网络.

最简单的是在目标模型训练过程中, 源模型作为目标模型的一个模块, 其模型参数不做任何改动. 这相比于从最初训练模型, 训练过程会缩短很多[65]. 在生物图像分析中, 可以将在ImageNet上训练的模型迁移过来进行皮肤损伤分类[55].

为了发挥迁移学习和多任务学习的潜能, 已训练的模型应该能够自由的分享. 在计算机视觉和自然语言处理领域, 已有代表性的模型库, 如Tensorflow模型库、Pytorch模型库等. 在生物学领域, Kipoi作为基因组模型库已经提供了约2000个基因组预测模型.

1.4.4 深度强化学习

在强化学习(reinforcement learning)中, 智能体向环境施加作用并接受环境的反馈, 然后根据反馈调整作用策略, 并由试错经验学习最优策略[66]. 强化学习最终目标是通过调整作用策略, 最大化预期奖励. 经典的强化学习方法为Q-learning[67], 然而传统的强化学习难以处理高维数据. 与深度学习结合的深度强化学习[68], 赋予了强化学习处理高维数据的能力, 使得其能够处理, 如蛋白质序列这类数据. DQN[68]采用深度神经网络建模Q值, 通过经验回放的方式从经验中学习策略. 在药物设计中, 采用RNN网络建模小分子SMILE格式数据, 将RNN作为策略网络, 学习小分子SMILE数据的生成策略[69].

深度强化学习已经在生物学领域得到应用[70], 包括代谢物生物合成路径设计[71]和药物的设计[69,72,73]等.

2 合成生物学应用

合成生物系统的设计和构建需要获得形式多样且具有预期功能特征的底层生物元件, 如调控元件(启动子、终止子、增强子、核糖体结合位点等)、功能元件(酶蛋白、抗菌肽等). 底层生物元件的设计可抽象为一类共性问题, 即序列是如何决定其功能? 或者说, 如何有效解析序列-功能之间极为复杂的映射关系? 近年来, 机器学习的迅速发展为该类问题的解决提供了全新的思路, 它可以利用生物大分子在改造过程中产生的突变文库数据, 先转换为计算机模型可以理解的输入, 然后选取合适的算法, 以数据为驱动, 训练合适的参数, 然后对未知的结构进行预测. 在该设计流程中, 无须详细的基础物理或生物学途径, 就可以构建(定性)或者预测能力(定量)的模型, 建立起映射关系. 这样的方法通过从表征的变体特性中学习, 并使用该信息来选择可能表现出改进特性的序列, 来帮助完成合成生物学方面的应用. 目前, 利用机器学习人工智能方法提取序列信息帮助完成合成生物方面改造的例子主要集中在启动子、多肽、蛋白质工程等研究领域. 下面着重从这几个方面来阐述合成生物在该领域遇到的问题和机器学习可以提供的解决思路和具体应用.

2.1 机器学习在调控元件设计上的应用

调控元件的设计在基因合成线路设计与代谢工程中都发挥着重要作用, 在药物设计[74]、特定产品生产[75]等方面都有巨大的潜力. 按照功能特性, 调控元件

通常包括启动子、增强子、沉默子等，它们对基因的表达起不同的调节作用，以满足生命活动的需要. 因此，如何合理设计调控元件的序列结构以定量调控基因的表达已经成为了我们面临的重要挑战.

以启动子为例[76]，传统的研究策略主要包括核苷酸水平的突变筛选[77]和元件水平的重组筛选[78]，这在一定程度上获得了一些成功. 然而，受到生物实验高人力成本、物力成本的限制，使得高表达、特异性表达的启动子设计变得十分困难，而且，生物体调控机制的复杂性也使得我们很难定量描述这些元件的功能[79]. 因此，基于机器学习、深度学习等人工智能方法的调控元件的合成与设计应运而生. 这类方法[80]通常以序列数据作为驱动，训练一个合适的从序列到活性之间的映射模型，然后对未知的结构进行预测，并用生物实验对结果进行验证. 该模型可以从已有的序列中提取出足够的结构信息，既可以加强我们对天然调控元件的理解，也可以帮助我们设计符合特定要求的调控元件[81].

目前，随着新一代测序技术[82]的发展，基于这类新技术对启动子进行设计和改造的思路已经获得了广泛的应用，其中以深度学习最为突出[83]. 目前已经应用的方法涵盖CNN[84,85]、RNN[86]、EP2vec[87]、GAN[80]等. 以Oubounyt等人[88]提出的基于深度学习的鲁棒的启动子预测器为例，他们分析了真核生物的启动子片段的特点，实现了对人类和小鼠启动子序列的准确识别. 在这项工作中，他们选取了包含或者不包含TATA框的人类和小鼠的四类数据集，序列长度取转录起始位点上游的250个碱基和下游的50个碱基序列，独热编码后作为输入数据，通过卷积神经网络和长短期记忆模型的组合网络进行训练分类，并将每类数据集按照五折交叉验证的方法对模型进行了评估. 结果发现，该模型在准确率、召回率等相关指标平均都达到了90%以上的准确率，较之前流行的模型有20%~30%左右的提升，由此可见，这类深度学习模型在对启动子的预测方面有着较好的性能. 此外，该研究着重强调了负样本的设计，提高了预测器提取特征的能力，降低了误识率. 总之，该预测器不仅在启动子的预测性能上优于之前的模型，而且也克服了高误识率的问题，为启动子序列的分析设计提供了一条新思路.

此外，基于机器学习预测增强子、沉默子等调控元件[89]也有广泛的应用，如基于CNN以端到端的方式预测增强子[55]、基于SVM模型以MPRA数据预测沉默子[90]等. 这表明，基于机器学习方法对调控元件进行设计和改造，有着重要意义.

2.2 机器学习在酶催化元件设计上的应用

蛋白质工程旨在改善甚至改变蛋白酶对天然与人工底物的催化性能，促进酶催化反应在化合物合成中的应用，对于生物转化、微生物发酵等产业的发展具有十分重要的意义. 然而，传统的蛋白质工程策略往往需要进行多轮突变文库构建与高通量筛选，耗时耗力. 近年来，计算机辅助蛋白质设计，以及人工智能技术的应用为蛋白质工程领域发展提供了新思路. 通过机器学习指导的定向进化进行蛋白质工程，有望在短时间内极大地优化蛋白质功能，并且显著减少繁琐的实验[91].

目前，通过对特定功能的蛋白质催化酶进行有选择性的突变迭代，定向进化可以帮助研究人员获取显著优化功能的蛋白质. 在进行蛋白质定向进化过程中产生的大量数据可以通过机器学习的方法进行数据模型构建，用于指导后续的蛋白工程化改造[92]，如图5所示. 蛋白质序列和功能之间的关系可以通过很多机器学习的统计模型进行分析. 例如，可以通过偏最小二乘回归算法来鉴定细菌卤代醇脱卤酶的有益突变，进而提高该蛋白的产量[93]. 通过人工智能计算技术，研究人员运用势能算法，近似反应态几何尺度限定等计算方法，对芽孢杆菌来源的天冬氨酸进行了分子重构，设计了具有特异分子选择性和立体选择性的β-氨基酸合成酶[94]. 为了提高蛋白质的热稳定性能，Romero等人[7]通过重组细菌细胞色素P450蛋白质CYP102A1、CYP102A2和CYP102A311的血红素结构域的序列片段，利用T50(在10 min的温育后导致酶丧失一半活性的温度)来检测细胞色素P450的热稳定性，并运用机器学习算法建立了用于训练T50的初始高斯过程(Gauss processes，GP)模型. 在收集到足够的训练数据后，作者随后使用贝叶斯优化来搜索更多的热稳定变异体. 首先使用批处理GP上限算法产生了热稳定P450的各种变体，通过筛选预计高度稳定的序列来验证其序列功能模型，之后使用GP-UCB算法进行另外两个迭代，最终挖掘到更加稳定的嵌合P450蛋白酶.

2.3 机器学习在功能多肽设计上的应用

许多小分子多肽能有效抑制细菌、真菌、病毒、支原体、衣原体、螺旋体、原虫及肿瘤细胞的增殖，甚至具有激素调节、免疫调节、刺激伤口愈合及促进动物生长等多种生物活性，其研究开发已成为当前一

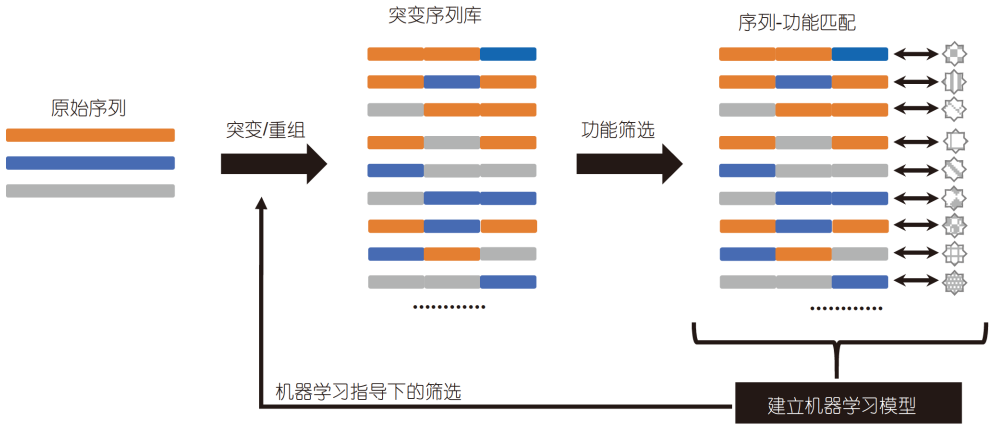


图 5 (网络版彩色)采用上一轮定向进化过程中收集的数据来构建机器学习模型, 用于预测本轮突变体的性能
Figure 5 (Color online) The function of variants are predicted by the machine learning model build on the dataset collected during directed evolution

大热点. 研究人员已从自然界中发现和分离了数千种活性多肽, 并建立了相关数据库[95~98]. 传统分离鉴定方法虽然有效但过程却十分繁杂, 需要耗费大量的人力、财力和时间, 尤其是筛选环节充满了重复性劳动. 除了筛选和鉴定天然抗菌肽, 人们亦期待能够按照预定目标进行新肽的人工设计或改造. 随着高通量实验技术的兴起, 多肽数据库的数据量正在快速扩增, 通过机器学习对多肽数据(氨基酸序列、二级结构、电荷、疏水性、生物活性等)进行分析来挖掘出其序列结构特征和生物活性之间的关联已成为可能. 根据机器学习的特点, 这种预测序列结构-活性间的关联问题可分为两类: 一是预测给定的多肽序列是否具有生物活性, 或者具有哪一类活性, 属于典型的分类问题(定性问题); 二是根据给定的多肽序列, 预测其具体活性的高低, 则属于典型的回归问题(定量问题). 机器学习在该问题上的应用, 将着力于解决: 多肽序列的哪些特征决定了其是否具有活性或者具有哪类活性? 多肽序列的哪些关键特征决定了其活性的高低? 这些问题的回答将决定我们能否设计出预期的目标多肽分子.

利用机器学习技术, 通过对已知的多肽序列数据信息进行学习, 建立合理的预测算法从中找出最优序列的特点和规律, 有望推广至未知多肽的预测、发现或设计. 例如, 在抗菌肽研究上, 机器学习可应用于建立定量构效关系(QSAR)模型, 即利用理化性质描述符来预测其生物活性. 目前已有多种机器学习方法得到应用, 包括人工神经网络(ANN)[99~102]、支持向量机(SVM)[99,103]、定量矩阵(quantitative matrix, QM)[99]、隐马尔可夫模型[104]、随机森林(random forests, RF)[105]、k-近邻(k-nearest neighbor, kNN)[6,106]、进化算法[107]、自组织映射(self-organizing mapping, SOM)[108]等. 为了设计出更高活性的新抗菌肽, Lee等人[109~111]基于支持向量机构建了针对α-螺旋抗菌肽的分类器, 可根据抗菌肽的理化性质发现并设计新的α-螺旋结构抗菌肽. 为实现该目标, 研究人员以线性SVM作为"白箱"机器学习技术, 根据Bi等人[112]报道的特征选择方法, 从1588个包含了抗菌肽的长度、电荷、疏水性等信息理化描述符中, 鉴别出12个最具预测性的描述符, 建立了12个描述符的QSAR模型, 并以286个α-螺旋抗菌肽和286个非抗菌肽数据进行训练. 训练好的模型可识别出决定抗菌肽活性的关键特性, 并以人们可理解的方式展示其抗菌肽分类流程. 由此建立的线性SVM模型对测试集的分类精度高达91.9%, 特异性和敏感性也分别达到93.0%和90.7%, 甚至比非线性或者包含更多描述符的SVM模型具有更优的性能. 尽管该方法较为简单, 但仍获得了很高的分类精度, 而且还揭示了抗菌肽特定理化过程与抗菌活性之间的相互关联. 研究人员由此建立了一个精准且可解释的α-螺旋抗菌肽活性模型, 并通过实验发现了数种通过自然进化或点突变难以产生的α-螺旋结构抗菌肽.

2.4 机器学习在代谢工程中的应用

1991年, 美国加州理工学院Bailey[113]在*Science*发表了综述性文章"Toward a science of metabolic engineering", 标志着代谢工程概念的诞生. 代谢工程结合酶学、化学计量学、分子生物学、数学和组学等研究手段, 对细胞代谢途径进行修饰和调控, 实现特定目的

产物的高效合成. 在随后的几十年中, 代谢工程领域发展迅速, 不仅在生物发酵, 同时在医药、植物、动物、环境和疾病治疗等领域均有重要应用[114,115]. 但是随着组学等技术的发展, 人工技术手段已无法高效处理代谢工程领域产生的海量生物数据. 近年来, 机器学习通过数据收集、学习和模型预测等方法在代谢工程中的应用越来越广泛, 包括用于基因注释、宿主微生物筛选的机器模型DeepRibo[1]、PEPARATION[2]、DeepEC[3]; 用于代谢途径构建的3N-MCTS[8]; 用于代谢流调控的GEMs[116]、MiYA[117]及Dcell[118]; 用于扩大化培养的SVM[119]、Fuzzy logic[120]和GPR[121]. 将代谢途径引入外源宿主仍然存在许多问题. 例如, 外源代谢途径的引入会导致宿主生长缓慢、代谢中间产物积累导致代谢失衡甚至致死等问题. 机器学习可以通过对基因表达水平、代谢流、培养条件等进行分析调控, 从而得到最高效的目的产物生产条件. 此外, 由于机器学习在代谢工程领域已有较多的应用, 在该领域产生了内容、类型和格式不一致的海量生物数据限制了机器学习在该领域的进一步发展. 同时, 机器学习应与传统的机械模型相结合, 进而提高机器学习预测的准确性. 外源代谢途径在宿主中的重构会影响宿主的代谢平衡, 对宿主的生长特性、代谢平衡等产生不可预知的影响. 因此, 对外源代谢途径组合的优化可以有效地提高目的产物的合成效率. 近期, MiYA[117]成功地将标准化构建模块与机器学习结合, 并应用于优化酿酒酵母中紫罗兰苷代谢途径. MiYA以β-胡萝卜素的代谢途径为训练集, 通过测试2%~5%的组合并与人工神经网络相结合可以精确地调整每个基因的表达水平, 而且可以避免在训练样本较少时可能出现的过度拟合问题. 进一步, 通过收集24株生产紫色杆菌素生产菌株的基因组等数据, 准确地从3125种不同代谢途径组合中预测并经实验验证得到了一株紫色杆菌素产量提高2.42倍的目的菌株. 同时, MiYA可以有效地降低目的菌株副产物的合成, 将代谢流更多的导向紫色杆菌素合成.

2.5 机器学习在合成生物其他领域的应用

除上述4个领域外, 近年来机器学习在其他重要的合成生物学领域中也得到广泛应用, 如基因线路设计、RNA合成生物学、基因编辑等.

2.5.1 在基因线路设计中的应用

基因线路在生物体内可实现各种重要的生物学功能, 如振荡器、双稳态、对环境信号的超灵敏响应等. 高质量的基因线路的设计已成为合成生物学研究中的热点领域. 基因线路设计通常包括两种相反的设计思路: 正向设计需要对特定的线路进行模拟和功能预测, 而反向设计则先定义所需的功能, 再寻找一个基因线路来实现它. 反向设计问题是一个复杂系统多参数寻优的问题, 其计算量往往十分巨大而难以求解. 机器学习尤其是深度学习为解决此类问题提供了新思路. 例如, Hiscock[122]利用梯度下降优化算法快速地筛选和设计出可实现不同功能的基因线路. 在此基础上开发了一个易用的Python组件“GeneNet”, 可为合成生物学应用提供自动化的基因线路设计服务. 而在另一项研究中, Wang等人[123]为了克服海量参数寻优的计算问题, 训练了一组高性能的神经网络模型来预测随机基因表达和新图案的形成, 从而实现了对该类生物学模型的快速寻优.

2.5.2 在RNA合成生物学中的应用

具有预期生物学功能的工程RNA分子在合成生物学中扮演愈发重要的角色, 尤其是作为对小分子、蛋白及核酸的可编程响应元件, 如核开关(riboswitch)、核调控子(riboregulator)、核酶(ribozyme)等. 然而, 这些RNA分子的设计极富挑战性, 主要存在的问题在于其功能的易变性难以得到有效预测. 为解决该难题, Angenent-Mari等人[124]以支点开关(toehold switch)为例, 利用深度神经网络(DNN)来预测其功能. 为便于训练模型, 他们合成并鉴定了91534个支点开关, 该数据集涵盖23个病毒基因组和906个人类转录因子. 优化后的DNN模型易于理解且可以实现二级结构的可视化, 其性能也远优于传统热力学及动力学模型. 该研究提出的模型和方法有望推广至其他功能RNA分子的设计, 进而推动RNA合成生物学的发展.

2.5.3 在基因编辑中的应用

基因编辑作为一种生物学研究方法, 通过对现有基因组序列进行删除、替换、插入等操作可以有效地改写生命体的遗传信息, 是合成生物学领域发展的重要工具. 近些年发展起来的基于细菌适应性免疫系统的CRISPR-Cas9等基因编辑技术极大地推动生物制造、疾病治疗等生物科技领域的发展, 具有巨大的应用价值. 然而, CRISPR-Cas9等基因编辑技术在编辑效率、精确度及脱靶效应等方面仍然面临诸多挑战[125]. 通过机器学习方法可以有效提高预测CRISPR编辑系统的准确性[126]. Cas9核酸酶是高度特异性基因组编辑工具, 其应用受到向导RNA(gRNA)在靶效率和脱靶问

题的影响. 为了优化基因编辑效率, Wang等人[127]对两个酿脓链球菌来源的SpCas9变异体eSpCas9(1.1)、SpCas9-hf1, 以及一个野生型SpCas9(WT-SpCas9)进行基因组级筛选, 以测量其gRNA活性, 并获得每个核酸酶超过50000个gRNAs的插入或缺失频率, 覆盖了约20000个基因. 通过评估获取到的1031个特征对gRNA效率的影响结果开发活性预测模型. 分析数据显示, RNN与生物特征结合测活性预测方式显著优于其他Cas9核酸酶活性预测模型, 证明通过深度学习方法有助于Cas9核酸酶的gRNA设计优化及计算机模型的开发. Kim等人[128]使用基于包含单导RNA编码和目标序列对的人类细胞库的高通量方法, 评估了12832个目标序列的SpCas9活性. 通过深度学习的技术, 对SpCas9诱导的插入或缺失频率的大数据集进行了训练, 开发了名为DeepSpCas9的SpCas9活性预测模型. DeepSpCas9对独立生成的数据集进行测试时显示出很高的泛化性能. 另一个方面, 许多抗CRISPR蛋白被发现可以抑制基于CRISPR-Cas体系的适应性免疫系统. 然而, 将已知的所有抗CRISPR蛋白序列进行比较, 并不能完全揭示抗CRISPR家族的共有特性. 为解决这个问题, Eitzinger等人[129]开发了一个基于机器学习的算法-AcRanker, 该工具可以仅仅基于蛋白序列信息直接鉴定新的潜在的抗CRISPR蛋白.

3 总结与展望

合成生物学以现有生物元件为基础, 通过设计构建特定用途的生物或者生化体系来实现生物体的特定功能. 合成生物学集成了多学科的技术方法和理念, 通过遗传分子水平的设计, 生物体代谢网络的构建和生物体元件的重组装来创造新的应用价值. 在这个过程中, 如何从生物过程流数据中获取有关知识, 抽提隐含在生物过程大数据中的一系列本体特性, 实现从感知到认知的过程是一个有待解决的重要问题. 人工智能可以帮助从生物过程获得的流数据获取隐含在生物过程大数据中的一系列本体特性, 真正实现从感知到认知部分, 从感知到反馈. 最终帮助人们从系统角度认识生物体的组成元件、行为和控制机理, 综合解析生物体的代谢与变化规律.

现代合成生物学的发展采用“设计-构建-测试-学习”的策略, 机器学习人工智能技术可以帮助从大量的实验数据中进行数据挖掘分析, 有助于设计更高效的网络体系并改造生物. 合成生物过程的智能感知与再认知学习是生物过程中可实现人工智能化的核问题. 另外, 鉴于生物反应器, 自动化机器产出的庞大数据量, 所有可能的排列组合、参数、数学模型, 对于人脑来说将难以把控, 需利用计算机辅助进行数据分析. 通过深度机器学习, 由计算机反复分析多循环、海量的实验结果, 可以辅助规划并优化新合成生物的设计方案. 在这些合成生物学领域项目中, 机器学习技术有很大的运用空间.

目前, 通过机器学习研究方法构建基因序列与生物体具体功能间的因果链条还存在诸多问题. 对于简单的线性模型来说, 其模型参数代表了输入特征对预测结果的重要性. 若输入特征相互独立, 线性模型就具有强可解释性. 然而, 对于复杂模型, 如神经网络, 特征冗余、特征与预测结果之间存在非线性关系, 模型可解释性极差. 另外, 对一些生物网络, 如基因调控网路, 基因之间存在调控关系的混淆效应. 需要借助因果推断, 解除调控关系之间的混淆[37]. 目前随着基因组方面的对照实验数据的不断丰富, 因果关系建模在机器学习领域得到越来越多的关注, 相信因果关系建模具备很大发展潜力. 另外, 蛋白质、启动子等序列功能建模是机器学习研究中越来越受欢迎的领域, 然而获得有监督标签的样本量数据需要付出比较高昂的价格成本. 怎样获得有效的序列模型数据, 减少资源浪费, 将筛选到无功能蛋白、启动子等序列样本的概率降低, 同时, 怎样通过有限的样本数据构建筛选模型是机器学习技术在未来需要重点解决的问题.

参考文献

1 Clauwaert J, Menschaert G, Waegeman W. DeepRibo: A neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. Nucleic Acids Res, 2019, 47: e36–e36

2 Ndah E, Jonckheere V, Giess A, et al. REPARATION: Ribosome profiling assisted (re-)annotation of bacterial genomes. Nucleic Acids Res, 2017, 47: e36

3 Ryu J Y, Kim H U, Lee S Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci USA, 2019, 116: 13996–14001

4 Oubounyt M, Louadi Z, Tayara H, et al. Deepromoter: Robust promoter predictor using deep learning. Front Genet, 2019, 10: 1–9

5 Sonnhammer E L, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol, 1998, 6: 175–182

6 Wang P, Hu L, Liu G, et al. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS One, 2011, 6: e18476

7 Romero P A, Krause A, Arnold F H. Navigating the protein fitness landscape with Gaussian processes. Proc Natl Acad Sci USA, 2013, 110: e193−e201

8 Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 2018, 555: 604–610

9 Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 2015, 33: 831–838

10 Ohler U, Liao G, Niemann H, et al. Computational analysis of core promoters in the *Drosophila* genome. Genome Biol, 2002, 3: 1–12

11 Domany E. Cluster analysis of gene expression data. J Stat Phys, 2003, 110: 1117–1139

12 Yu J, Smith V A, Wang P P, et al. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics, 2004, 20: 3594–3603

13 Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions. Nat Mach Intell, 2019, 1: 105–111

14 Zhang Y, Chu C, Chen Y, et al. Splice site prediction using support vector machines with a Bayes kernel. Expert Syst Appl, 2006, 30: 73–81

15 Weston J, Leslie C, Ie E, et al. Semi-supervised protein classification using cluster kernels. Bioinformatics, 2005, 21: 3241–3247

16 Ghandi M, Lee D, Mohammad-Noori M, et al. Enhanced regulatory sequence prediction using gapped *k*-mer features. PLoS Comput Biol, 2014, 10: e1004035

17 Kutsuna N, Higaki T, Matsunaga S, et al. Active learning framework with iterative clustering for bioimage classification. Nat Commun, 2012, 3: 1032

18 Camacho D M, Collins K M, Powers R K, et al. Next-generation machine learning for biological networks. Cell, 2018, 173: 1581–1592

19 Dutil F, Cohen J P, Weiss M, et al. Towards gene expression convolutions using gene interaction graphs. 2018, arXiv: 1806.06975

20 Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press, 2016

21 LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput, 1989, 1: 541–551

22 Palit S. Studies on ozone-oxidation of dye in a bubble column reactor at different pH and different oxidation-reduction potential. Int J Environ Sci Dev, 2010, 18: 341–346

23 Salakhutdinov R, Hinton G. Deep Boltzmann machines. J Mach Learn Res, 2009, 5: 448–455

24 Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression). WIREs Comp Stat, 2010, 2: 97–106

25 Cortes C, Vapnik V. Support-vector networks. Mach Learn, 1995, 20: 273−297

26 Bhasin M, Zhang H, Reinherz E L, et al. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett, 2005, 579: 4302–4308

27 Sun X D, Huang R B. Prediction of protein structural classes using support vector machines. Amino Acids, 2006, 30: 469–475

28 Furey T S, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 2000, 16: 906–914

29 Johnson S C. Hierarchical clustering schemes. Psychometrika, 1967, 32: 241–254

30 Müllner D. Modern hierarchical, agglomerative clustering algorithms. 2011, arXiv:1109.2378

31 Kerr M K, Churchill G A. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci USA, 2001, 98: 8961–8965

32 Arnau V, Mars S, Marín I. Iterative cluster analysis of protein interaction data. Bioinformatics, 2005, 21: 364–378

33 Pearl J. Bayesian networks: A model of self-activated memory for evidential reasoning. In: Proceedings of the 7th Conference of the Cognitive Science Society. University of California, Irvine, CA, USA, 1985. 15−17

34 Tsamardinos I, Brown L E, Aliferis C F. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn, 2006, 65: 31–78

35 Cussens J. Bayesian network learning with cutting planes. 2012, arXiv: 1202.3713

36 Richardson T, Jensen F V. An introduction to Bayesian networks. J Am Stat Assoc, 1997, 92: 1215

37 Friedman N, Linial M, Nachman I, et al. Using Bayesian networks to analyze expression data. J Comput Biol, 2000, 7: 601–620

38 Barron A, Rissanen J, Bin Yu J. The minimum description length principle in coding and modeling. IEEE Trans Inform Theor, 1998, 44: 2743–2760

39 Perrin B E, Ralaivola L, Mazurie A, et al. Gene networks inference using dynamic Bayesian networks. Bioinformatics, 2003, 19: 138−148

40 Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 2003, 302: 449–453

41 Eddy S R. What is a hidden Markov model? Nat Biotechnol, 2004, 22: 1315–1316

42 Eddy S R. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol, 1995, 3: 114–120
43 Eddy S R. Profile hidden Markov models. Bioinformatics, 1998, 14: 755–763
44 Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics, 1998, 14: 846–856
45 Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods, 2015, 12: 931–934
46 Kelley D R, Snoek J, Rinn J L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res, 2016, 26: 990–999
47 Elman J L. Finding structure in time. Cogn Sci, 1990, 14: 179–211
48 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780
49 Angermueller C, Lee H J, Reik W, et al. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol, 2017, 18: 1–13
50 Pan X, Rijnbeek P, Yan J, et al. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics, 2018, 19: 1–11
51 Quang D, Xie X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res, 2016, 44: 1–6
52 Quang D, Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods, 2019, 166: 40–47
53 Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. In: 2005 IEEE International Joint Conference on Neural Networks, Volume 2. 2005. 729–734
54 Bruna J, Zaremba W, Szlam A, et al. Spectral networks and deep locally connected networks on graphs. 2013, arXiv: 1312.6203
55 Min X, Zeng W, Chen S, et al. Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics, 2017, 18: 478
56 Scholz M, Kaplan F, Guy C L, et al. Non-linear PCA: A missing data approach. Bioinformatics, 2005, 21: 3887–3895
57 Brechtmann F, Mertes C, Matusevičiūtė A, et al. OUTRIDER: A statistical method for detecting Aberrantly expressed genes in RNA sequencing data. Am J Human Genets, 2018, 103: 907–917
58 Kingma D P, Welling M. Auto-encoding variational bayes. 2013, arXiv: 1312.6114
59 Grønbech C H, Vording M F, Timshel P, et al. scVAE: Variational auto-encoders for single-cell gene expression data. Bioinformatics, 2020, https://doi.org/10.1093/bioinformatics/btaa293
60 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems. 2014, 2: 2672–2680
61 Zhang L, Tan J, Han D, et al. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. Drug Discov Today, 2017, 22: 1680–1685
62 Chen Z, Badrinarayanan V, Lee C Y, et al. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 793–802
63 Xu C, Tao D, Xu C. A survey on multi-view learning. 2013, arXiv: 1304.5634
64 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345–1359
65 Kornblith S, Shlens J, Le Q V. Do better imagenet models transfer better? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 2656–2666
66 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey. J Artif Intell Res, 1996, 4: 237–285
67 Watkins C J C H, Dayan P. Q-learning. Mach Learn, 1992, 8: 279–292
68 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
69 Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Sci Adv, 2018, 4: eaap7885
70 Mahmud M, Kaiser M S, Hussain A, et al. Applications of deep learning and reinforcement learning to biological data. IEEE Trans Neural Netw Learn Syst, 2018, 29: 2063–2079
71 Koch M, Duigou T, Faulon J L. Reinforcement learning for bioretrosynthesis. ACS Synth Biol, 2020, 9: 157–168
72 Liu X, Ye K, van Vlijmen H W T, et al. An exploration strategy improves the diversity of *de novo* ligands using deep reinforcement learning: A case for the adenosine A_{2A} receptor. J Cheminformatics, 2019, 11: 35
73 Zhou Z, Kearnes S, Li L, et al. Optimization of molecules via deep reinforcement learning. Sci Rep, 2019, 9: 1–10
74 Lynch S A, Gill R T. Synthetic biology: New strategies for directing design. Metab Eng, 2012, 14: 205–211
75 Oborne M. The Bioeconomy to 2030: Designing a policy agenda. Organisation for Economic Cooperation and Development. OECD Obs, 2010, 278: 35
76 Meng H, Wang Y. Cis-acting regulatory elements: From random screening to quantitative design. Quant Biol, 2015, 3: 107–114
77 Du J, Yuan Y, Si T, et al. Customized optimization of metabolic pathways by combinatorial transcriptional engineering. Nucleic Acids Res, 2012,

40: e142
78 Blazeck J, Garg R, Reed B, et al. Controlling promoter strength and regulation in *Saccharomyces cerevisiae* using synthetic hybrid promoters. Biotechnol Bioeng, 2012, 109: 2884–2895
79 Gilman J, Love J. Synthetic promoter design for new microbial chassis. Biochem Soc Trans, 2016, 44: 731–737
80 Wang Y, Wang H, Liu L, et al. Synthetic promoter design in *Escherichia coli* based on generative adversarial network. 2019, bioRxiv: 563775
81 Siwo G, Rider A, Tan A, et al. Prediction of fine-tuned promoter activity from DNA sequence. F1000Research, 2016, 5: 158–172
82 Behjati S, Tarpey P S. What is next generation sequencing? Arch Dis Child-Educ Pract, 2013, 98: 236–238
83 Angermueller C, Pärnamaa T, Parts L, et al. Deep learning for computational biology. Mol Syst Biol, 2016, 12: 878
84 Umarov R K, Solovyev V V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLoS One, 2017, 12: 1–12
85 Umarov R, Kuwahara H, Li Y, et al. Promoter analysis and prediction in the human genome using sequence-based deep learning models. Bioinformatics, 2019, 35: 2730–2737
86 Quang D, Xie X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res, 2016, 44: e107
87 Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics, 2018, 19: 13–22
88 Oubounyt M, Louadi Z, Tayara H, et al. DeePromoter: Robust promoter predictor using deep learning. Front Genet, 2019, 10: 286
89 Li Y, Shi W, Wasserman W W. Genome-wide prediction of *cis*-regulatory regions using supervised deep learning methods. BMC Bioinformatics, 2018, 19: 202
90 Doni Jayavelu N, Jajodia A, Mishra A, et al. Candidate silencer elements for the human and mouse genomes. Nat Commun, 2020, 11: 1061
91 Lutz S. Beyond directed evolution—Semi-rational protein engineering and design. Curr Opin Biotechnol, 2010, 21: 734–743
92 Yang K K, Wu Z, Arnold F H. Machine-learning-guided directed evolution for protein engineering. Nat Methods, 2019, 16: 687–694
93 Fox R J, Davis S C, Mundorff E C, et al. Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol, 2007, 25: 338–344
94 Li R, Wijma H J, Song L, et al. Computational redesign of enzymes for regio- and enantiose-lective hydroamination. Nat Chem Biol, 2018, 14: 664−670
95 Wang Z. APD: The antimicrobial peptide database. Nucleic Acids Res, 2004, 32(Suppl1): D590−D592
96 Wang G, Li X, Wang Z. APD2: The updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res, 2009, 37: D933–D937
97 Wang G, Li X, Wang Z. APD3: The antimicrobial peptide database as a tool for research and education. Nucleic Acids Res, 2016, 44: D1087–D1093
98 Zhao X, Wu H, Lu H, et al. LAMP: A database linking antimicrobial peptides. PLoS One, 2013, 8: e66557
99 Lata S, Sharma B K, Raghava G P S. Analysis and prediction of antibacterial peptides. BMC Bioinformatics, 2007, 8: 263
100 Fjell C D, Jenssen H, Hilpert K, et al. Identification of novel antibacterial peptides by chemoin-formatics and machine learning. J Med Chem, 2009, 52: 2006−2015
101 Cherkasov A, Hilpert K, Jenssen H, et al. Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. ACS Chem Biol, 2009, 4: 65–74
102 Torrent M, Andreu D, Nogués V M, et al. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. PLoS One, 2011, 6: e16968
103 Giguère S, Laviolette F, Marchand M, et al. Machine learning assisted design of highly active peptides for drug discovery. PLoS Comput Biol, 2015, 11: e1004074
104 Fjell C D, Jenssen H, Fries P, et al. Identification of novel host defense peptides and the absence of α-defensins in the bovine genome. Proteins, 2008, 73: 420–430
105 Bhadra P, Yan J, Li J, et al. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci Rep, 2018, 8: 1
106 Xiao X, Wang P, Lin W Z, et al. IAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem, 2013 , 436: 168−177
107 Maccari G, Di Luca M, Nifosí R, et al. Antimicrobial peptides design by evolutionary multiobjective optimization. PLoS Comput Biol, 2013, 9: e1003212
108 Schneider P, Müller A T, Gabernet G, et al. Hybrid network model for “deep learning” of chemical data: Application to antimicrobial peptides. Mol Inf, 2017, 36: 1−8
109 Lee E Y, Fulan B M, Wong G C L, et al. Mapping membrane activity in undiscovered peptide sequence space using machine learning. Proc Natl Acad Sci USA, 2016, 113: 13588–13593

110 Lee E Y, Wong G C L, Ferguson A L. Machine learning-enabled discovery and design of membrane-active peptides. Bioorg Med Chem, 2018, 26: 2708–2718

111 Lee E Y, Lee M W, Fulan B M, et al. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? Interface Focus, 2017, 7: 20160153

112 Bi J, Bennett K P, Embrechts M, et al. Dimensionality reduction via sparse support vector machines. J Mach Learn Res, 2003, 3: 1229−1243

113 Bailey J E. Toward a science of metabolic engineering. Science, 1991, 252: 1668−1675

114 Burgard A P, Maranas C D. Review of the enzymes and metabolic pathways (EMP) database. Metab Eng, 2001, 3: 193−194

115 Carroll A L, Case A E, Zhang A, et al. Metabolic engineering tools in model cyanobacteria. Metab Eng, 2018, 50: 47−56

116 O'Brien E J, Monk J M, Palsson B O. Using genome-scale models to predict biological capabilities. Cell, 2015, 161: 971−987

117 Zhou Y, Li G, Dong J, et al. MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in *Saccharomyces cerevisiae*. Metab Eng, 2018, 47: 294–302

118 Ma J, Yu M K, Fong S, et al. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods, 2018, 15: 290–298

119 Chari R, Yeo N C, Chavez A, et al. SgRNA scorer 2.0: A species-independent model to predict CRISPR/Cas9 activity. ACS Synth Biol, 2017, 6: 902–904

120 Lee J, Lee S Y, Park S, et al. Control of fed-batch fermentations. Biotechnol Adv, 1999, 17: 29−48

121 Masampally V S, Pareek A, Runkana V. Cascade gaussian process regression framework for biomass prediction in a fed-batch reactor. In: Proceeding of the 2018 IEEE Symposium Series on Computational Intelligence, 2018

122 Hiscock T W. Adapting machine-learning algorithms to design gene circuits. BMC Bioinformatics, 2019, 20: 214

123 Wang S, Fan K, Luo N, et al. Massive computational acceleration by using neural networks to emulate mechanism-based biological models. Nat Commun, 2019, 10: 1−9

124 Angenent-Mari N M, Garruss A S, Soenksen L R, et al. Deep learning for RNA synthetic biology. 2019, BioRxiv: 872077

125 Pickar-Oliver A, Gersbach C A. The next generation of CRISPR-Cas technologies and applications. Nat Rev Mol Cell Biol, 2019, 20: 490−507

126 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: 13−22

127 Wang D, Zhang C, Wang B, et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat Commun, 2019, 10: 4284

128 Kim H K, Kim Y, Lee S, et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. Sci Adv, 2019, 5: eaax9249

129 Eitzinger S, Asif A, Watters K E, et al. Machine learning predicts new anti-CRISPR proteins. Nucleic Acids Res, 2020, 48: 4698–4708

Summary for “面向合成生物学的机器学习方法及应用”

Machine learning for synthetic biology: Methods and applications

Ruyun Hu[1†], Songya Zhang[2,3,4†], Hailin Meng[2,5], Han Yu[2,3,4], Jianzhi Zhang[2,3,4], Xiaozhou Luo[2,3,4], Tong Si[2,3,4*], Chenli Liu[2,3,4*] & Yu Qiao[1*]

[1] *Institute of Advanced Computing and Digital Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China;*
[2] *Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China;*
[3] *Shenzhen Institute of Synthetic Biology, Shenzhen 518055, China;*
[4] *CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen 518055, China;*
[5] *Center for Biological Engineering, Guangzhou Institute of Advanced Technology, Chinese Academy of Sciences, Guangzhou 511458, China*
† Equally contributed to this work
* Corresponding authors, E-mail: tong.si@siat.ac.cn; cl.liu@siat.ac.cn; yqiao@siat.ac.cn

Traditional synthetic biology takes a trial-and-error approach, suffering from inefficiency and local optima. Recent advances in high-throughput experimental techniques generate a huge amount of biological data, which enables the use of machine learning to close the “design-build-test-learn” loop. Machine learning, especially deep learning, is a data-driven modeling method, which extracts useful patterns from big data and then leverages learned knowledge to tackle specific tasks. In this review, we aim to provide a brief primer of machine learning to synthetic biologists. Starting with common taxonomy, we introduce representative methods, pipelines, and underlying principles of machine learning that can be applied in synthetic biology. We include typical methods such as support vector machine, deep neural networks, generative adversarial nets, transfer learning and reinforcement learning. In particular, discriminative models, including convolutional neural networks and support vector machine, are appropriate for predicting sequence-function relationship. Generative models, including generative adversarial nets (GANs) and deep generative models for graph generation, are suitable for sequence or network design.

Next, we review the recent applications of machine learning in studying synthetic biology parts and modules, including promoters, bioactive peptides, enzymes, metabolic pathways, and genetic circuits. For example, DeePromoter combined a convolutional neural network and a long-short term memory to achieve an accuracy as high as 90% when predicting promoter sequences. For enzyme design, a Gauss Process model was proposed with Bayesian optimization by upper confidence bound method, which resulted in the engineering of thermostable P450 enzymes. For antimicrobial peptides, a generative GAN model enhanced with a feedback mechanism was trained to design peptide sequences with new functions.

Finally, we conclude with future challenges and directions. Particularly, interpretable machine learning models are desirable to guide mechanistic investigation. Moreover, it is necessary to develop new machine learning methods that are more compatible with biological data, which are heterogeneous, multi-modal (such as sequence, network, image, and structure), and lack of proper labels. With the increasing availability of big biological data and development of machine learning methods tailored for synthetic biology, we envision a paradigm shift towards a closed cycle of “design-build-test-learn” in creating artificial life with predictable functions.

machine learning, synthetic biology, synthetic biology parts design, bio-networks design

doi: 10.1360/TB-2020-0456