世界一流学会建设项目

中国人工智能系列白皮书
——智能生物信息处理 2019

中国人工智能学会

二〇一九年十月

《中国人工智能系列白皮书》编委会

《中国人工智能系列白皮书——智能生物信息处理》编委会

本书编写组

李　敏　　刘　琦　　宋晓峰　　王　颖

赵兴明　　张世华　　邹　权

全书统稿：赵兴明

前　言

近年来，伴随着生物技术的飞速发展，海量的生物医学大数据快速积累。这些数据包含了极其重要的信息，为研究生命奥秘、生命活动本质以及生命产生和发展规律提供了契机。生物医学研究也逐渐偏离传统的实验科学，开始向数据驱动的生物信息学交叉学科发展。但是，生物医学数据的解读速度远远滞后于数据的产出速度，生物信息的处理急需生物医学与数学和计算机科学等学科交叉。与传统数据相比，生物医学数据具有自己的鲜明特点，如样本少、维度高、数据非结构化、数据种类多、数据量庞大等特点。生物信息处理需要克服数据缺失、小样本高维、异源异构数据等困难，需要统计学、数学和计算机科学等学科的交叉融合。

随着信息化技术的发展，各类人工智能技术的出现大大促进了生物信息处理的速度和精度，也推动了生物信息学的发展。目前，人工智能技术进入了新的阶段。经过 60 多年的发展，在大数据、高性能计算、脑科学等新理论新技术驱动下，人工智能呈现出深度学习、跨界融合、人机协同等新特征。大数据驱动的知识学习已经成为人工智能的发展重点，相应的智能算法已经被广泛应用于生物信息处理。人工智能技术已被用于基因组注释、药物设计、结构预测等研究，帮助生物学家筛选处理、解释利用生物学研究中不断收集的庞大数据，提取重要的知识，并从大量的原始生物数据中找出有用的模式，帮助生

物学家解决未解决的问题，极大程度上推动了生物信息学的发展。

随着生物信息学和人工智能技术的飞速发展，精准医学的概念也开始被推广。在精准医学中，通过患者基因组学和临床表型等生物信息的处理，可以形成更加精准的诊断方案，并制定个性化的治疗方案，从而更加有效地预防和治疗疾病。精准医学涉及患者的临床表型、分子组学、医学影像、表观遗传、微生物组学等多种类型和结构的数据，疾病的亚型分类、预后分析，以及药物靶点识别、精准用药指导和药物风险等。人工智能技术通过融合和挖掘多模态大数据，可以加速实现疾病的精准治疗。

本白皮书收集了目前国内外人工智能与生物信息处理交叉的最新理论研究成果，并介绍了人工智能技术在生物信息学和精准医学等领域中的应用。编写过程中，参考了国内外从事相关研究工作者的资料，在此一并表示感谢。

目　录

第一章 人工智能与非编码基因 RNA

1.1 人工智能与非编码 RNA 概述

非编码 RNA 是一种由 DNA 转录，但不会翻译成蛋白质的 RNA 分子。越来越多的研究发现非编码 RNA 有着重要的生物学功能。非编码 RNA 种类繁多，其中包括 rRNA，tRNA，snRNA，snoRNA 和 microRNA 等多种已知功能的 RNA[1]。随着高通量技术的不断发展，我们对 RNA 分子功能的理解有了新的认识。

人工智能技术近年来发展迅速，机器学习与解决问题的能力大大提升。生命科学中通过高通量测序技术产生的大数据，正适合使用人工智能技术进行加工处理。非编码 RNA 的相关研究是目前生命科学重要前沿问题之一，相应的人工智能技术应用于研究非编码 RNA 的实践也愈加重要。非编码 RNA 的计算识别就是利用了人工智能技术来实现的。例如，microRNA(miRNA)最开始被认为是垃圾序列，但其调控功能被证实后，大量研究发现 miRNA 与多种生命活动相关。因此，miRNA 的识别有利于发现一系列生命过程的分子机制，但 miRNA 前体具有特殊的发夹结构、结构与功能保守性等特征，传统的计算识别方法存在着 miRNA 识别率低，敏感性差等问题，此外 miRNA 识别的实验方法成本高且用时长[2]。人工智能技术的出现，提供了更好的计算识别方法，其主要思路为：(1) 分析 miRNA 相关的物理化学特征并结合其生物意义，提取特征。(2) 选择数据集，通过机器学习的方法训练分类器模型。(3) 对候选的 miRNA 序列进行特

征提取，通过分类器模型进行预测，给出潜在的候选序列。可选择的分类器有：支持向量机(SVM)、随机森林、贝叶斯以及决策树[3]。很多非编码 RNA 通过 RNA 与蛋白质的互作实现它们的调控功能。因此，鉴定非编码 RNA 与蛋白质的相互作用对于理解非编码 RNA 的功能而言是必不可少的一步。然而，目前用于鉴定非编码 RNA 与蛋白质相互作用的生物实验技术花费相当昂贵且耗时，而利用人工智能技术对非编码 RNA 与蛋白质的相互作用进行预测，其结果较为准确且成本相对较低。

非编码 RNA 可以通过与蛋白质或者其他分子结合来开启或者关闭基因，进而达到控制基因表达作用的目的。并且，非编码 RNA 与多种疾病的发生有关，而目前对非编码 RNA 的研究了解甚少。随着科学技术的不断进步，将人工智能技术应用于非编码 RNA 具有广阔研究前景。

1.2 人工智能与非编码 RNA 识别

1.2.1 miRNA 的计算识别

miRNA 通过与靶基因 3’UTR 碱基互补配对来发挥调控功能[4]。虽然 miRNA 序列中部分碱基在进化中可能发生改变，但 miRNA 与靶基因结合的种子序列具有严格的保守性。基于机器学习的方法不依赖于序列的保守性，将已知的 miRNA 作为阳性集，非 miRNA 作为阴性集，选择包含序列和结构的特征，如发夹结构的最小自由能、茎序列、环的长度和序列重复等，训练分类器模型，将模型运用于不同的数据，从而预测新的 miRNA[5]。然而，基于机器学习的方法的准确度高度依赖于已知 miRNA，因此，阴性集和阳性集的选择至关重要。

RNAmicro 方法通过整合序列分析和结构预测来识别新的 miRNA。通过 RNAz 和 EvoFold 等工具在全基因组上生成的 ncRNA 的统计结果，预测 pri-miRNA，识别出潜在的发夹结构 pre-miRNA，分析考虑结构和热力学特性的特征，来预测新的 miRNA[6]。MiRFinder 方法比较了相关物种的序列，利用 18 种不同的特征，例如：最小自由能，成熟 miRNA 的碱基配对、二级结构元件的频率等，从候选的 miRNA 集合里识别出发夹结构。由于大量的序列可以形成类似于 miRNA 前体的发夹结构，因此该方法通过随机检验评估预测 miRNA 的统计学显著性，从而降低方法的假阳性率[7]。ProMiR 方法基于隐马尔可夫模型，通过序列比对来识别 miRNA[8]。上述方法都能比较准确地识别出 miRNA，但当前的研究无法判断选择的阴性集中是否包含 miRNA 行使功能的发夹结构，这在一定程度上限制了这类方法的准确性。

1.2.2 lncRNA 的计算识别

基于机器学习的方法是识别 lncRNA 最常见的手段，这些方法在识别 lncRNA 时都表现出比较高的准确度和灵敏度。CNCI 方法是通过分析序列的内在组成来区分蛋白编码和非编码转录本的分类器。根据两条序列中相邻核酸三联体的不均匀分布，构建 64×64 的三联体评分矩阵来评估序列，并使用滑动窗口作为补充以获得更加可靠的结果[9]。CNCI 旨在区分没有序列注释的编码和长非编码转录本，当序列缺乏注释时，CNCI 能够有效地解决这个问题。LncRNA-ID 方法是基于随机森林识别 lncRNA 的分类模型，使用随机森林来改进分类模型，使其能够处理不平衡的训练数据。LncRNA-ID 具有 11 个特征，这些特征可根据开放阅读框、核糖体和蛋白质的保守性进行分类。

LncRNApred 在构造分类器之前，采用自组织特征映射聚类来选择更具有代表性的训练数据集，这在一定程度上增强了 LncRNApred 的性能。LncRNApred 选择最长的 ORF 和覆盖范围、GC 含量、k-mer、序列长度等作为特征，然后构建随机森林模型。LncADeep 是一种新颖的 lncRNA 识别和功能注释工具[10]，将深度学习算法深度置信网络中的集成序列固有特征和同源特征集成构建深度学习模型以识别 lncRNA。LncADeep 是第一个考虑转录本全长和部分长度的工具，其构建的模型既针对全长的转录本又考虑部分长度的转录本。此外，该方法还整合 KEGG 和 Reactome 等生物通路数据，为候选的 lncRNA 的功能注释提供准确的通路和功能模块，更加准确地预测新的 lncRNA。PLIT 方法是一个新的比对工具，它使用 L1 正则化进行特征选择，使用随机森林分类器对序列进行分类[11]。该方法利用 LASSO 优化模块在随机森林中进行迭代，LASSO 优化模块从训练集和验证集特征中选择最优特征集，总共共选择了包括 ORF 长度、ORF 覆盖度、GC 含量和密码子偏性特征等 31 个特征。PLIT 方法基于 RNA-seq 数据集的转录序列来准确识别 lncRNA，提供了更优的特征，是一种强大的半监督优化方法[12]。

1.3 人工智能与非编码 RNA 调控网络建模技术

非编码 RNA 在生物体内发挥着重要的调控作用。多个不同的非编码 RNA 调控关系对可以形成复杂的调控网络。单个非编码 RNA 的失调能导致整个调控网络的功能紊乱。许多疾病，例如癌症，都与非编码 RNA 的失调有着密切关联。非编码 RNA 根据其长度、二级

结构的差异可以分为不同的类型，不同类型的非编码 RNA 在生物体内的作用机制及其功能具有差异性。随着人工智能技术的发展，高通量的进行非编码 RNA 调控关系的预测以及构建非编码 RNA 调控网络已经成为可能。然而目前的预测方法大多存在一定的假阳性，仍需后续实验对人工智能技术预测的调控关系进行验证。

1.3.1 miRNA 调控网络的构建

成熟 miRNA 是一类长度约 22nt 的核苷酸序列，通过靶向调控基因的表达参与到多种生物学过程中。例如，成熟的 miRNA 可以与 mRNA 的 3'UTR 互补结合，抑制 mRNA 的翻译或者降解 mRNA，从而达到抑制基因表达效果[13]。传统实验的方法可以检测 miRNA 与靶基因的调控关系，然而实验的方法通量低且花费较高。随着人工智能技术在生命科学领域的应用发展，目前已开发许多 miRNA 调控算法工具以及调控关系数据库，例如常用的 miRNA 靶预测算法 targetScan，它通过寻找基因的 3'UTR 与 miRNA 种子序列的匹配模式来识别 miRNA 结合位点，预测 miRNA 靶基因。PITA 预测算法根据位置特异的规则和物种之间保守性来评估 miRNA 靶向特征，该方法设计了一个用于 miRNA-靶基因互作的无参数模型，计算从 miRNA-靶基因复合物中获得的自由能与解除配对关系所耗费的能量差来评估靶向的可能性[14]。

此外，还有些算法可以利用 CLIP-seq 数据预测 miRNA 靶基因。micro MUMMIE 可以针对每个 miRNA 和每个可能的靶位点评估已知与 miRNA 诱导的沉默复合物(miRISC)和 miRNA 靶点之间的相互作

用的特征[15]。其结合了 miRNA 的进化保守性特点，miRNA 的种子匹配的类型，以及在 CLIP 数据和峰的序列组成，利用多变量马尔可夫模型识别 miRNA 靶点。PARma 算法利用 Ago-PAR-CLIP 实验的完整数据研发的分析软件，能够用来鉴定 microRNA 的靶点以及与这些靶点结合的 microRNA，其将实验的数据特征整合到模型中，将模型和新的匹配模式迭代地应用于数据来估计种子活动概率、聚类置信度得分并分配最可能的 miRNA[16]。

总之，不同的 miRNA 靶点预测方法基于不同的假设和模型，而预测 miRNA-mRNA 相互作用的工具旨在获得准确的结果。然而，目前 miRNA 靶预测算法仍具有高假阳性率，miRNA 靶点的相互作用的准确预测研究仍然具有挑战性。为此，必须权衡每个预测工具使用的生物学方面的特征，通过获取 miRNA 与基因的靶关系，构建在不同状态下的 miRNA 调控网络。

1.3.2 lncRNA 调控网络的构建

lncRNA 是一类长度大于 200nt 的长链非编码 RNA，虽然不具备蛋白编码能力，但是在细胞内发挥重要调控作用。lncRNA 可以与 mRNA 共享 MRE（miRNA 响应元件），作为 miRNA sponge 间接调控基因的表达。starbase 数据库利用人工智能技术，收集了 lncRNA 作为 ceRNA(competing endogenous RNA)的调控关系[17]。此外，lncRNA 还可以作为转录因子 scaffold 参与转录调控的过程中，调控基因的转录过程，从而调控基因的表达。

人工智能技术可以挖掘 lncRNA 作为 scaffold 的调控关系。

LncReg 数据库收集了 lncRNA 与基因的调控关系，该数据库收集 1081 个经验证的 lncRNA 相关调控关系，包括 258 个非冗余 lncRNA 和 571 个非冗余基因[18]，提供 lncRNA 调控网络和生物信息学研究的综合数据，这对理解 lncRNA 的功能具有十分重要的作用。目前有许多数据库可以调控 lncRNA，例如 lncRNA2target 收集了敲除或过表达 lncRNA 的表达谱数据，将 lncRNA 的靶基因视为差异表达的基因，并开发了一个名为 LncRNA2Target 的数据库来收集、存储实验验证的 lncRNA-mRNA 靶点之间的关联关系。这个数据库不仅有助于计算研究人员对公开的 lncRNA 靶点进行综合性的分析，而且还能让实验科学家们能在其他相关公共数据的背景下分析自己的数据，大大加快 lncRNA 的靶基因的研究进程。

目前大部分关于 lncRNA 与基因调控网络的研究均基于 ceRNA 假说，然而 lncRNA 许多其他的调控功能正在进一步的挖掘、研究，lncRNA 的功能多样性同时也使得调控网络变的复杂，lncRNA 作为调控元件参与到多个生物学过程的研究还有待深入探索。

1.4 人工智能与非编码 RNA 二级结构建模技术

RNA 的二级结构是核苷酸链弯曲、折叠自身互补配对形成的，RNA 通过特定的二级结构发挥生物学功能。非编码 RNA 的二级结构通常比序列本身保守性更好，可用于研究 lncRNA 在不同物种之间的进化特征。此外，RNA 二级结构中的一些模体(motif)，例如发夹结构，可以在某些生物学调控过程中发挥重要作用。我们常见的 tRNA 即转运 RNA，它拥有三叶草结构，在蛋白质翻译过程中发挥重要的

作用。常见的 RNA 二级结构类型有：茎环结构(stem-loop)、假结(pseudoknots)、发卡(hairpin-loop)等。准确的二级结构预测方法对于理解 RNA 功能起着重要作用。目前大部分二级结构预测方法工具主要根据进化保守性进行同源模建，然而由于非编码 RNA 的序列保守性较差，结果往往并不理想。随着人工智能技术的发展，我们可以利用人工智能进行非编码 RNA 二级结构预测。常用的非编码 RNA 二级结构预测方法是 RNAfold，该方法基于动态规划算法进行建模，能够快速并高效的预测 RNA 的二级结构，并计算预测二级结构下的最小自由能[19]。

RNA 在折叠自身互补配对形成复杂二级结构的过程中，往往会匹配更多的核酸达到一个稳定的状态，还可以通过最小自由能，根据序列特征预测 RNA 二级结构。计算方法一下两种，一是通过包括伪假结的结构对的预测。二是通过找到一组同源序列共有的二级结构来改进结构预测准确性。此外，集成学习的方法可以用来预测非编码 RNA 的二级结构，集成机器学习方法比使用单个的学习方法得到的结果更加准确。

除此之外，随着人工智能技术的发展，深度学习技术愈加成熟，同样可以应用于非编码 RNA 的二级结构预测。预测 RNA 的二级结构对于研究其功能至关重要，但确定 RNA 二级结构具有一定难度，特别是对有假结的 RNA 的二级结构进行预测。DMfold 是一种基于深度学习的预测方法，可以在已知结构中学习相似的 RNA 来预测 RNA 的二级结构。该方法在多序列中使用相似的 RNA 序列而不是高度同

源的序列，因此减少对辅助序列的需求。在 DMfold 中，只需要输入目标序列就可以预测二级结构，其折叠参数通过深度学习自动完全提取，可以避免单序列方法中缺少折叠参数的问题，也缩小了用户自定义参数导致的差异[19]。HotKnots 同样是一个启发式的算法预测非编码 RNA 二级结构，包括预测二级结构中的假结，它使用了动态规划算法组装假结的结构，使用自由能最小化算法用于假结的二级结构来识别候选茎环结构，对计算结果中的最低自由能进行排序，这样可以预测几个潜在的二级结构并进行深入筛选，更为准确的刻画 RNA 二级结构的潜在特征[20]。

总而言之，利用人工智能技术预测非编码 RNA 二级结构的方法众多，大多数是基于最小自由能和基于序列同源的方法。这些人工智能技术的方法可以从头预测非编码 RNA 的二级结构，研究分析对已知序列的非编码 RNA 的二级结构，为生物学家研究非编码 RNA 的分子功能提供参考。从头预测甚至突变后的非编码 RNA 的结构，研究突变对非编码 RNA 的结构的影响，使得科学家可以预测突变前后的 RNA 结构变化，进一步研究突变对 RNA 功能的影响，这对研究非编码 RNA 的功能有着十分重要的作用。

1.5 非编码 RNA 与蛋白互作建模技术

很多非编码 RNA 通过与蛋白质的互作来实现它们的调控功能。因此，识别非编码 RNA 与蛋白的相互作用对于理解非编码 RNA 的功能具有重要作用。然而，目前用于鉴定非编码 RNA 与蛋白质相互作用的生物实验技术花费相当昂贵且耗时。因此，建立一个准确的计

算预测模型已经成为识别非编码 RNA 与蛋白相互作用不可或缺的方法[21]。近年来人工智能技术在生命科学研究中应用广泛，使得很多生物学问题有了更好的解决方案或得以解决。

近年来，虽然监督学习和非监督学习两种方法在 RNA 与蛋白相互作用的研究上都取得了很好的效果，但是它们仍然有缺点以及可以改进的空间。大量研究发现，非编码 RNA 与蛋白质相互作用具有序列特异性[22]，这一研究表明该序列携带足够有用的信息用于预测非编码 RNA 与蛋白的相互作用。

1.5.1 提取互作模型的序列特征信息

目前用于研究预测非编码 RNA 与蛋白相互作用的第一步往往是分别对非编码 RNA 序列与蛋白序列进行特征提取，一般用奇异值分解(SVD)将非编码 RNA 序列从 k-mer 稀疏矩阵转化为特征向量。为了进一步提取隐藏的高级特征信息，可以使用深度学习中 SAE(Stacked Auto-Encoder)算法[23]。SAE 算法是一种无监督特征学习方法，它与大多数深度学习一样逐层学习原始数据的各种表达式，基于前一层的表达特征，每一层再提取出更抽象、更合适的复杂特征，以完成一些分类任务。SAE 能够从原始数据中自动学习高级特征，形成降维表示，已有科学家应用 SAE 与随机森林分类器完成非编码 RNA 与蛋白质相互作用的预测[24]。

1.5.2 构建互作模型的随机森林分类器

提取所需的特征后，先选择合适的分类器，对特征进行分类。目前常用且高效的分类器有支持向量机，神经网络，朴素贝叶斯以及随

机森林等。预测非编码 RNA 与蛋白的相互作用时，将对各种分类器进行选择，旨在选择最准确、性能最好的分类器。随机森林分类器是用多个决策树训练和预测数据集，很多用于预测非编码 RNA 与蛋白质的相互作用所用的分类器正是随机森林。IPMiner，一种计算方法，是利用 SAE 从蛋白和非编码 RNA 的序列组成特征中挖掘出隐藏序列的交互模式，再将学习过的隐藏特征输入随机森林分类器，得到 RNA 和蛋白的互作模型[25]。

1.5.3 构建互作模型的卷积神经网络

近几年来，深度学习在很多领域（如：语言识别、翻译、图像识别等）取得了巨大的成功，也在生命科学研究领域应用广泛。卷积神经网络是一种用来处理网格结构数据的特殊网络结构，该网络通过一系列的办法，将识别庞大数据量的问题进行降维，达到使其能够被训练的目的。卷积神经网络避免了显式的特征取样，采用隐式的方式从训练数据中学习。深度学习的研究方法在生物信息学领域也应用广泛。DeepBind 利用深度卷积神经网络训练相关序列，可以用来预测 RNA 与蛋白质结合序列的特异性[26]。基于深度学习的方法在解决各种生物学问题上有着不错的表现，实际生活中人们常常采用特定的深度学习方法实现非编码 RNA 与蛋白质的相互作用的预测。

1.5.4 贝叶斯方法识别非编码 RNA 与蛋白的相互作用

贝叶斯算法在统计学中是一种非常重要的分类方法，它是在已知对象先验概率的情况下利用贝叶斯公式计算它的后验概率。根据得到的后验概率选择其中最大概率的类作为该对象的类[27]。目前已经有两

种识别蛋白质与 RNA 相互作用的贝叶斯分类方法，分别是朴素贝叶斯(NB)分类方法与扩展朴素贝叶斯(ENB)分类方法。这两种分类方法均只需输入蛋白质与 RNA 的初级序列，不需要任何其它信息。朴素贝叶斯分类模型是一个能够快速、准确预测蛋白质与 RNA 相互作用的分类器，并且特征之间是独立的，这符合朴素贝叶斯分类器的假设[28]。而扩展朴素贝叶斯分类器考虑到了特征之间的相关性，这样能够提供具有相关特征的准确预测。朴素贝叶斯模型特征独立性的假设，极大地降低了分类器的复杂程度，提高了估计参数的可靠性，特别是当输入数据集维数和可用的数据集的大小相比较更高时，估计参数的可靠性增加[29, 30]。然而在实际情况当中特征之间存在着一定的相关性，而扩展朴素贝叶斯能够处理这种特征之间有着相关性的数据集，这种预测蛋白质与 RNA 的相互作用的扩展朴素贝叶斯分类器，已经通过了生物学实验的验证。

1.6 人工智能在非编码 RNA 中的发展前景

非编码 RNA 与蛋白之间的相互作用在很多生物过程中起着重要的作用，并且与多种疾病的发生息息相关。为此，识别非编码 RNA 与蛋白的相互作用对于进一步研究非编码 RNA 功能具有重要的意义。随着科技的发展，人工智能技术应用于计算预测非编码 RNA 与蛋白的相互作用变得方便快捷，逐渐取代了技术花费昂贵的生物实验的主导地位。选择最合适的人工智能技术预测非编码 RNA 与蛋白的相互作用非常重要，这需要人们不断努力探索，找到最佳的预测非编码 RNA 与蛋白质相互作用的方法，发挥人工智能技术在生物信息学这一领域应用的优势。

第二章 人工智能与宏基因组

2.1 人工智能与宏基因组概述

宏基因组(Metagenomics)，也称环境微生物基因组或元基因组，是指特定环境群落中全部微小生物 DNA 的总和。对于微生物而言，传统的研究方法是在实验室中对微生物进行繁殖和生长，但 99%以上的微生物无法在现有实验室条件下进行培养。因此宏基因组技术为观察微生物世界提供了一个强大的视角，高通量测序技术(High throughput sequencing technology)提供精细到碱基层面的分辨率，为人类对微生物世界的认知带来重要变革。技术的创新使得微生物群落的研究对象从最初的土壤迅速拓展到人体（肠道、口腔、皮肤等）、水体、大气、废水以及动植物体内的微生物。

高通量测序技术将碱基序列随机打断后扩增，并行对几十万到几百万条 DNA 分子进行序列测定，得到几百万条读段(read)，每条读段的来源以及之间的相互关系尚未可知。因此，若将单个基因组序列视为一本书，高通量测序则是将该书撕成碎片后得到的数据集合。而宏基因组测序则是将许多不同种类且不同数量的基因组书籍混合撕碎后得到的混合数据碎片，测定无法获取碎片的源头书籍信息和碎片之间的相互关系。

微生物群落的高通量测序有两种对象：早期出现的是以 16S rRNA 为代表的扩增子测序。16S rRNA 是微生物中核糖体 RNA 的一个亚基，由于其普遍存在于一切原核生物细胞内，生理功能重要且稳

定，因此可以用于研究群落的物种组成、物种间的进化关系以及群落的多样性，但分析精度常常无法达到种(species)的水平。但其测序成本低，且有较为成熟的基因标记(gene marker)数据库，在微生物群落的研究和分析中仍然占有重要的地位，类似的还有 18S rRNA 和 ITS 测序。真正的宏基因组测序数据是对环境样品中全部微生物的总 DNA 进行高通量测序，是目前获取微生物群落数据最主要的技术。该技术除了获取物种组分、关系、多样性外，还可以进行基因和功能层面的深入研究，达到种甚至株(strain)水平的分析精度。因此，本主题主要介绍智能领域的方法在宏基因组测序分析中的应用。

宏基因组测序数据可以视为从混合概率分布（例如，混合泊松分布）中抽取的观测值的集合，单个测序数据的大小在 100-101GB 之间。因此对微生物群落的多样性、物种组分、相互关系、功能及与宿主或者环境性状的关联性进行估计、预测与判断，是典型的大数据挖掘问题。科学家们广泛应用基于统计推断、机器学习、模式识别和深度学习等人工智能领域的技术和方法进行该方面问题的研究。

基于微生物群落的宏基因组高通量测序数据的研究内容如图 2-1 所示。其主要分析包括：(1) Beta 多样性分析，即多个微生物群落间的差异性度量和比较，可以理解为前文提到的不同堆的混合书籍碎片之间的差异性。(2) Alpha 多样性分析，即单个微生物群落内部物种的丰富度和组成成分，可以理解为前文提到的一堆混合书籍碎片数据中有多少种书、分别是什么书和数量有多少。(3) 根据微生物群落内部物种的组分估计不同微生物之间的相互作用关系。(4) 根据微生物群

落内测序数据拼装出的长序列进行基因预测，从而进行群落内部的功能分析。(5) 将微生物群落的组分、功能及序列信息与群落的某些表型特性相关联，进行宏基因组关联性分析，基于基因、物种和序列信息，识别不同组别的微生物群落间的组别特异标记物，例如特异基因、特异物种或者特异序列等。

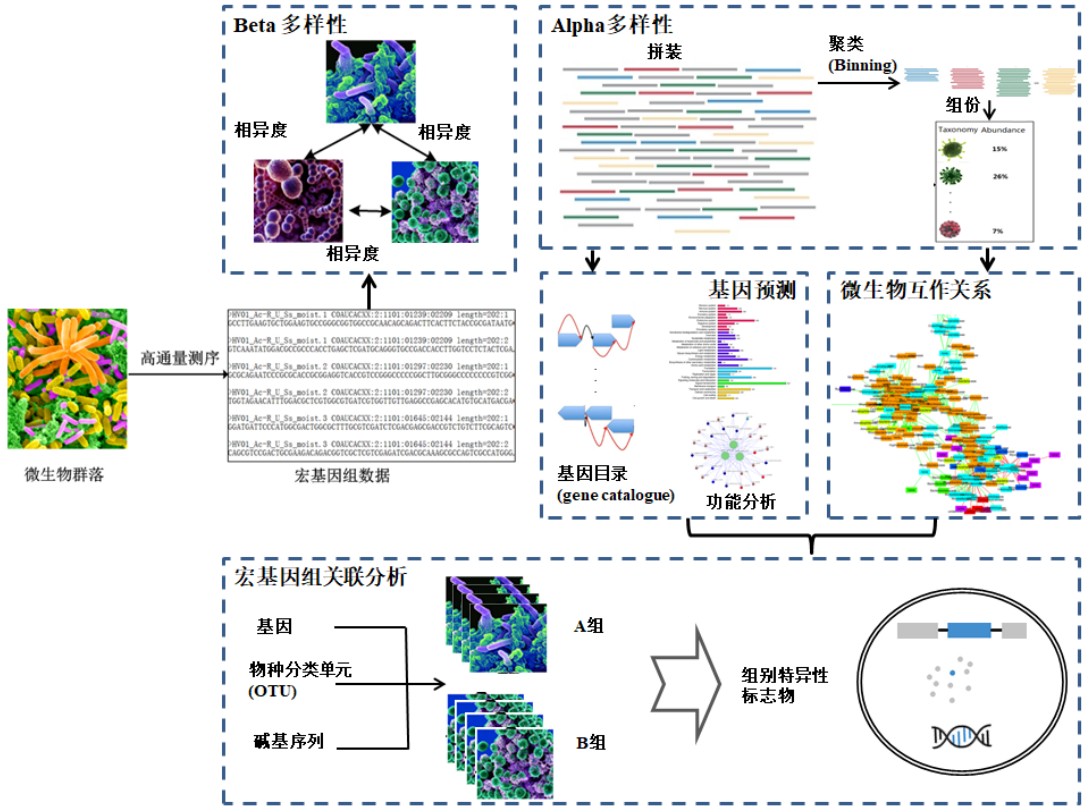

图 2-1 宏基因组高通量测序数据研究内容

以上提到的五个研究内容，本质都是建模为字符串的分类、聚类、预测或者优化问题，可通过统计模型、模式分类、智能优化和深度学习等方法进行研究和解决。但是在生物信息中，高通量测序数据文件大、样本数量相对少，因此不能直接套用自然语言处理、图像处理等领域现有的模型、方法和流程，需要针对具体的问题和数据，提出针对性的研究方案。

2.2 智能算法在宏基因组数据中的应用

2.2.1 Beta 多样性：宏基因组间的相异度度量

Beta 多样性分析，又称生境间的多样性(between-habitat diversity)，是指生境群落之间物种组成沿环境梯度不同的相异性或物种沿环境梯度的更替速率，可以通过计算多个微生物群落间的差异性进行度量，其本质是基于高通量测序得到的百万短序列(reads)，度量不同测序数据之间的差异性。

2.2.1.1 基于序列配准的相异度度量

传统的序列比较主要基于序列配准，例如 Smith–Waterman 算法[36]和 BLAST[37]等，虽然相对精确，但是存在以下限制[38, 39]：(1) 依赖于参考基因组或者基因序列数据库。由于微生物中大量物种的基因组都未知或者不完整，会影响分析结果的准确性和完整性。现有研究表明，海水、人体肠道还有含有藻类的水域中的微生物群落，分别有 19–42%[40]、10–20%[41]和高达 50%[42]的测序读段无法配准到参考数据库。(2) 高通量测序数据得到的是短读段，需要拼装得到较长的序列(contigs)，由于微生物之间序列的重复或者接近，直接进行序列配准很难像完整基因组比对那样得到较精确的比对结果。(3) 多序列比对是 NP 难问题，同时要耗费大量的时间成本和计算资源，因此科学家们开始探索免于配准的宏基因组间相异度度量方法。

2.2.1.2 基于 kmer 统计的相异度度量

针对测序序列，可以计算得到 kmer 的计数信息。kmer（也称为 k-tuple, n-gram 等）表示 k 长度的字符串，在测序数据中为 k 长度的

碱基序列，适用于大规模测序数据集。基于 kmer 计数信息的序列比较是最具代表性的免于配准的相异度计算方法。其机理是相同基因组的不同区域的 kmer 相对频度分布是相似的，而不同基因组之间的 kmer 分布则有较大的不同[43]。

基于不同长度的 kmer 计数信息，采用的处理策略也具有多样性，主要分为以下几类。

1. 基于频度分布统计模型的序列比较

基于 kmer 的计数信息，最直接的模型是估计每个 kmer 在所有序列中出现的概率，直接基于归一化后的观测频率，利用欧式距离、曼哈顿距离、d_2 距离等计算两个/两组序列之间的差异度[44]。或者更加细致地考虑基因组中某些碱基序列出现的概率常常受到它前面序列的影响，因此利用上下文信息，同时从测序数据中估计基因组的背景模型，去除背景噪声，得到 kmer 频度的期望值。研究者通过不同阶次的马尔可夫链较精确地描述测序数据中 kmer 的分布情况[45]。例如，r 阶马尔可夫链假设碱基状态的分布只与它前面 r 个位置的碱基信息有关，即$P\left(X_t \middle| X_1,\cdots,X_{t-1}\right)=P\left(X_t \middle| X_{t-r},\cdots,X_{t-1}\right)$,其中$X_1,\cdots,X_t$是序列$X$中按顺序依次出现，长度为 k 的碱基序列。基于马尔科夫模型对碱基序列进行比较最具代表性的度量模型有郝柏林院士团队提出的 CVTree[46-48]以及 Fengzhu Sun 团队提出的d_2^S, d_2^*[49, 50]。CVTree 用于比较细菌基因组，能较精确地构建细菌的系统发生树[51-53]。d_2^S, d_2^*广泛应用于基因组测序数据[54]、宏基因组[55]和宏转录组[38, 56]的比较，以及病毒与宿主关系的匹配[57]等。

2. 基于长 kmer 交集(intersection)的序列比较模型

基于 kmer 的计数信息建立在统计模型的基础上，需要估计每个 kmer 的分布和发生频数期望，因此要求计数向量不能稀疏，即当长度 k 的取值较大时，无法对 kmer 建立精确地统计模型，但 kmer 越长，其包含的信息就越多，信息量也越大，作为特征描述每个宏基因组的特性就更加精确。因此出现了基于 kmer 匹配数目的序列比较，现有实验表明这类方法能更加精确地衡量不同的序列/序列集之间的相异度。

随着 kmer 长度的增长， kmer 的数量呈指数级(4^k)增加，对计算资源和计算时间的需求急剧增长。针对此问题，最具代表性的解决方案是 Mash[58]，它将网页和图像比较中用到的 Minhash[59]技术运用在序列比较中。以长 kmer (k≥20bp)的频度作为特征，针对每个测序样本，进行长 kmer 随机抽样，构建该样本的缩略(sketch)表示，在此基础上估计两个样本间的 Jaccard 指数，衡量两个测序样本之间的距离。基于 Mash 的缩略表示，kWIP[60]利用香农熵作为权重，对在大部分样本中存在或者仅在很少样本中出现的 kmer 赋予低权重，通过加权内积估计样本间的距离。Skmer[61]用 Mash 计算基因组略读数据(genome skimming)的 kmer 频数谱和两个样本间的交集，以此修正测序误差和低覆盖度的影响，更精确地估计两个样本间的相异度。长 kmer 交集的序列比较模型在宏基因组的相异度计算中表现出显著优势[62]，Mash 已应用在全球海域的 44 个海水宏基因组数据集以及人体 888 个不同部位的宏基因组数据中，相比于其他基于 Jaccard 指数的工具，

其展现出更高的精度和更快的计算速度[58]。kWIP 应用于不同土壤、地点的水稻根系的微生物群落 16S rRNA 测序数据中，其计算获得的距离相比于采用 UniFrac 和加权 UniFrac，更能清晰地显示土壤特性和地点的聚类特性[60]。

3. 基于深度网络的基因组距离的映射学习

随着深度学习的不断发展，图像、自然语言处理等领域的相关技术也被引入到生物信息领域，进行尝试和探索。虽然目前尚未有针对宏基因组相异度计算的应用被报导，但 SENSE[63]可基于不同微生物的 16S rRNA 测序数据，通过两个孪生卷积神经网络拟合不同物种之间的相异度。训练集中的相异度通过 16S rRNA 的序列比对计算得出，虽然该方法还没有应用到真正的宏基因组比较中，但为其继续探索提供了思路和启发。

4. 信息理论在宏基因组度量中的应用

由于信息理论中如熵或者复杂度可以度量序列内部所包含的信息，因此通过比较两个序列同时发生所包含的信息量与各自单独发生所包含的信息量的关系，来衡量两个序列各自包含信息的差异，表现为互信息熵、相对熵（即 KL 散度）以及复杂度等计算方式。因此熵或者复杂度通常被融入以上的比较模型中[60, 64]，作为权重或修正项精确建模结果。

2.2.2 Alpha 多样性：宏基因组混合片段数据的拼装

Alpha 多样性，指的是特定区域或生态系统内的多样性，常用物种的数目和丰富度来度量。宏基因组测序数据中，读段来自不同微生

物的混合基因组，因此读段的拼装是宏基因组许多后续研究包括 Alpha 多样性的第一步。拼装(Assembly)是根据读段之间的重叠关系将读段连接成较长的序列(contig)。由于存在来自不同基因组的读段的混合，宏基因组的拼装过程更加复杂。拼装的主要策略分为以下三类[65]。

1. 基于贪婪扩展(greedy-extension)的序列拼装

贪婪扩展算法利用一些读段为种子(seed)，通过比较其他读段的前缀或后缀序列的重叠长度对该种子进行扩展，然后将扩展后的序列与读段连接成为新的种子，如此往复，直至没有新的读段可以加入连接为止[66]。由于贪婪扩展算法在每个步骤都选择对当前来说一步最好的策略，所以容易陷入局部极值，导致大量不正确的拼装结果。

2. 基于重叠布局一致性(overlap-layout-consensus; OLC)的序列拼装

基于 OLC 算法[67]的序列拼装主要分为三个步骤：首先识别出所有读段之间的重叠(overlap)区域，接着用图(graph)表示所有读段的布局(layout)和彼此之间的重叠区域，最后根据读段之间的布局和关系识别出一致性(consensus)的序列，即在图中寻找一条经过每个顶点的哈密尔顿路径(Hamilton path)，这是 NP 难问题。因此 OLC 算法的计算成本很大，一般用于单基因组的拼装[67, 68]，若用于宏基因组的序列拼装处理，则要耗费极高的计算时间和计算资源。

3. 基于 de Bruijn 图的序列拼装

与 OLC 不同，de Bruijn 图首先将 reads 打断成长度为 k 的核酸片

段，即 kmer。利用 kmer 间的重叠关系构建 de Bruijn 图，节点(node)由 kmer 组成，如果两个节点 kmer 之间存在(k-1)mer 完全匹配，则这两个节点有边相连。例如 ACTG, CTGC, TGCC 在 k=3 时的 kmer 为 ACT, CTG, TGC, GCC，可以表示为 ACT -> CTG -> TGC -> GC，组装工具从 de Bruijn 图中寻找欧拉路径(Eulerian path)组装出可能的序列。因此基于 de Bruijn 图的拼装过程不需要搜寻重叠区域，而是搜索具有相同 kmer 的读段。由于其计算速度快、存储能力高，大部分宏基因组序列拼装工具都采用此策略，但重复区域和测序误差是拼接中较难处理的两个问题，重复区域在图中表现为分支，测序误差会产生额外的假阳顶点(false positive vertice)。针对测序误差、SNPs、重复、分支以及缺口等问题，发展出 metaSPAdes[69], Meta-IDBA[70], MetaVelvet[71]等具有代表性的宏基因组数据拼装工具。

2.2.3 Alpha 多样性：宏基因组混合片段的聚类

宏基因组是不同基因组序列的混合，因此要获得微生物群落内部的组成信息，需要基于片段数据估计哪些原始读段或者拼装序列属于相同的物种，分别有多少种物种。其本质就是基于混合概率分布的观测值估计原有的概率分布情况，表现为序列片段的聚类问题，在宏基因组中被称为 binning。在宏基因组的聚类工具中，采用的特征主要体现为以下三类。

1. 基于序列的构成(composition)为特征

该特征基于不同的分类(taxonomy)其序列构成也不相同的假设，最常用的特征形式是 kmer 频度，例如 4mer 的频度分布，或者 GC 含

量等，将属于相同物种单元(Operational Taxonomic Units, OUT)的序列视为来自相同分布的观测值，通过 MCMC 极大似然估计[72]或者插值马尔科夫[73]来估计原始分布的参数，或者通过基于马尔科夫模型衡量序列间的相异度对聚类结果进行调整[74]。但是这类方法常依赖初始聚类数目的确定和初始聚类中心的设定，因此应用起来具有一定的依赖性。因此，基于自组织映射(self-organizing maps SOMs) [75-77]和基于 Stochastic Neighbor Embedding (SNE) [78]的方法用于将高维特征降维到低维空间以确定聚类数目和用于聚类结果的可视化。

2. 基于序列在不同样本间的丰度为特征

基于丰度的特征有两种使用方式：第一种针对单一测序样本，假设属于同一物种或者物种单元的序列丰度应该具有一致性，因此问题可以转化为：来自相同参数泊松分布的序列属于同一个物种单元，通过 EM 算法进行极大似然估计[79, 80]。第二种针对一系列测序样本，假设来自相同物种的若干序列，其在不同样本中得到的丰度向量应该具有较高的相关性。Canopy[81]是首先引入基于不同样本间基因丰度的相关性进行聚类的工具，该工具的探索为后续将序列构成与序列丰度结合的特征的使用提供了重要的参考信息。

3. 基于序列构成与序列丰度的组合为特征

研究者在前期的探索中发现，基于样本丰度相似性的特征能为聚类提供很重要的参考信息，因此研究者们将 kmer 频度信息与丰度信息连结成新的向量，通过混合高斯模型对聚类数目进行估计，并通过基于欧式距离或者其他范数的距离度量，采用变分贝叶斯等方法对序

列的聚类分配进行决策[82, 83]或者基于新向量计算不同序列间的基于概率模型的距离[84]。单拷贝标志基因对于聚类数目的确定也有重要的参考意义。

此外，基于序列构成与丰度信息，将序列信息视为来自低维空间中正态分布的观测值，通过变分自编码网络，将序列非线性映射到均值和方差，实现序列的聚类和可视化[85]。同时，针对宏基因组数据，基于参考病毒基因组训练得到的卷积网络，可以识别出细菌与病毒混合序列中的病毒序列[86]。

2.2.4 宏基因组内部物种关系网络构建

对于微生物群落来说，网络的构建是通过微生物的丰度信息构建他们之间的生态关系，例如共生共栖、互养、捕食、寄生等。物种关系网络以物种或者物种单元为节点，物种之间的正负相关性为边，刻画出微生物群落中存在的物种及其之间的关系。预测微生物间生态关系的原理为如果两个物种，或者两个物种单元在多个样本中都表现出同增同减的相似的丰度模式，那么它们之间存在一个正相关关系；如果呈现出互斥的丰度模式，那么它们之间存在一个负相关关系。根据物种关系网络可以推断出微生物群落的生态模式，物种的两两关系和更复杂的网络关系，以及中心节点等重要微生物，对研究微生物群落结构关系有重要作用。

基于宏基因组测序数据得到的物种或物种单元的丰度信息，估计和构建物种关系网络的工具主要有 SparCC[87], MENA[88], CoNet[89]等，其主要的研究步骤基本如下。

(1) 由于测序覆盖度以及丰度估计误差等因素，认为观测到的丰度数据都来自真实丰度数据下的某种分布。

(2) 由于丰度数据是成分数据(compositional data)，因此对数比值变换(log ratio transformation)去除闭合效应的步骤必不可少。

(3) 两个丰度比值的对数包含有丰度的方差信息和其真实值之间的相关性，因此可通过包含观测值、真实值、噪声等不同的数学关系式拟合出丰度的真实值或者相关性。SparCC 建立观测值、方差和相关性之间的数学工具，拟合出两个物种丰度的真实相关性[87]，而 CoNet 通过不同物种单元丰度的回归拟合出当前物种的真实丰度[89]。

(4) 将相关性大于某个阈值的微生物节点连接起来，形成物种关系网络。MENA 通过随机矩阵理论自动选择合适的阈值[88]。

2.2.5 宏基因组的功能分析

对于一个微生物群落，当研究者知道群落中有哪些微生物，它们之间的关系如何之后，群落中微生物的功能就是接下来被关注的问题。该问题的研究更具挑战性，但是对真正深入理解群落对象极为重要。对于功能分析，最直接的切入点是蛋白质编码基因，因此从宏基因组测序数据中进行基因预测(gene calling)是宏基因组功能分析的关键步骤。

宏基因组基因预测一般包括同源预测和从头预测。同源预测是通过与基因的同源序列比对，识别可能的基因区域，其依赖已知的基因信息且需要耗费较大的计算资源和计算时间。从头预测是从给定的基因序列抽取出有效的特征，利用机器学习或者统计模型进行描述，构

建概率模型，对编码基因进行预测，能够预测出已知和未知基因，且计算资源消耗小，时间花费少，其本质是一个模式识别问题。

1. 基于统计模型的基因预测

假设基因编码区域与非编码区域的碱基分布不同，基于分布模型的不同，可以预测出基因区域。Glimmer-MG[90]通过插值马尔可夫模型实现 kmer 分布的可变阶次马尔可夫链的精确建模，并整合 ORF 区域长度、邻接基因的方向以及距离等特征，实现宏基因组的基因预测。

另一方面，对 DNA 序列进行基因预测也可以视为基于观察到的碱基序列估计最可能产生该序列的隐状态和路径，因此可以用隐马尔科夫模型来描述该问题，并通过 Viterbi 算法进行求解，其基本思路如图 2-2 所示。这就是宏基因组注解工具 MG-RAST[91]基因注释 FragGeneScan[92]的基本原理。

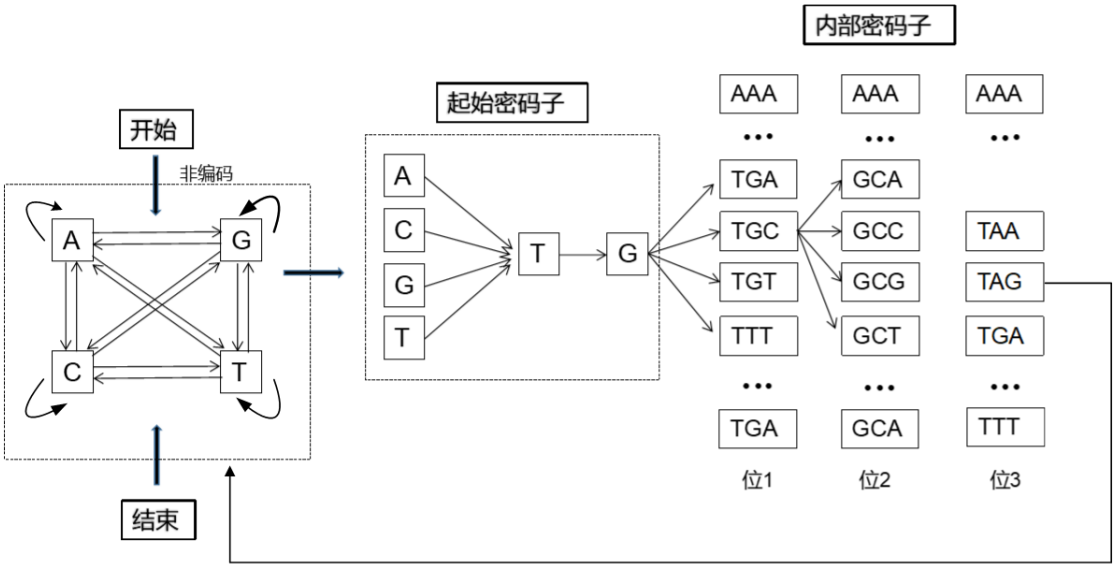

图 2-2　用隐马尔科夫模型预测原核生物的基因

2. 基于深度学习的基因预测

随着深度网络在分类中应用性能的提升，研究者们初步尝试将其

应用于宏基因组的基因预测中。CNN-MGP[93]将训练的基因序列用独热码(one-hot)形式表示，通过卷积神经网络(CNN)实现特征抽取，全连接的感知器网络(MLP)实现分类预测，进而对碱基序列进行基因预测。同时，DeepARG 通过深度网络将基于宏基因组数据对抗菌素抗性基因(ARG, Antibiotic resistance gene)进行预测[94]，输入从参考数据库中抽取的 4333 个特征，对长基因序列和测序数据建立两个含较多隐节点的全连接分类网络，进行 ARG 的预测。以上研究的基本架构如图 2-3 所示，仅较为初步地运用了深度网络的分类预测功能，离真正运用于宏基因组的基因预测还有一段距离。

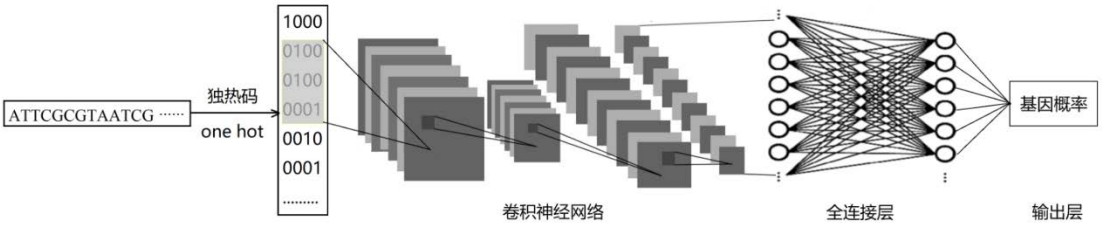

图 2-3 利用深度网络基于宏基因组测序数据进行基因预测

2.2.6 宏基因组关联性分析

宏基因组关联性分析(Metagenome-wide association studies, MWAS)旨在建立微生物群落的表型特性与基因型、功能或者碱基序列的关联。目前大部分研究集中在疾病与人体微生物群落之间的关联分析。在宏基因组中，关联分析是希望找到与疾病或者其他表型相关的标志物，因此可以建模为分类问题。

从上游的分析中，与表型建立关联关系的标志物可以是微生物的丰度、基因的丰度或者碱基序列的覆盖度。因此，将微生物群落按照不同的表型分组，将微生物丰度、基因丰度或者碱基序列的覆盖度作

为特征，通过选择可以将表型区分开的特征，从而获取与表型相关的标志物。也就是说，宏基因组关联分析的最终目的是获得与表型相关的标志物，寻找标志物的过程是分类效果进行特征选择。

传统的机器学习的分类算法和统计模型应用于关联分析中。在II型糖尿病(T2D)的宏基因组关联分析中[95]，研究者通过 PCA 降维，分析与疾病相关的微生物物种单元，并以基因的丰度为特征，一组丰度高度相关的基因的集合被定义为宏基因联结组(metagenomic linkage group,MLG)，识别出 T2D 相关的 MLG 单元和标志基因。后续在糖尿病，肠胃炎，肥胖症，肝硬化、肠癌以及类风湿性关节炎的研究中都以物种、功能以及基因的丰度为特征进行了宏基因组关联分析。同时，研究者还利用长 kmer (k=40bp)的频度为特征，通过 kmer 的频度在不同表型中的差异，识别单个 kmer 或者多个 kmer 的组合作为标志物，并通过 kmer 的拼接得到组别标志的长碱基序列[96]。

深度网络最直接的应用是模式分类，因此研究者们也尝试利用深度网络对不同类别的宏基因组数据进行分类。基于 OTU 丰度和 OTU 距离矩阵作为特征，研究者利用系统发生卷积神经网络(Phylogenetic convolutional neural networks)对正常人与病人的宏基因组进行分类[97]。同样基于 OTU 丰度，研究者尝试通过深度网络建立人体肠道宏基因组进行年龄预测和肠道菌群与年龄的关联关系，交叉验证中年龄的预测误差为 3.94 年[98]。

2. 3 智能算法在宏基因组分析中的应用

2.3.1 大型宏基因组项目

2007 年底，美国国立卫生研究院宣布投入 1.15 亿美元启动“人类微生物组计划(HMP)”研究微生物菌群结构变化对人类健康的影响。

HMP 已经进行了十多年，完成了两个阶段的工作，其第一阶段被称为 HMP1，第二阶段被称为 iHMP（也称为 HMP2）。

2008 年，欧洲斥资 2120 万欧元启动人体肠道宏基因组计划。

2016 年，美国投入 1.21 亿启动国家微生物组计划，意在开发微生物的潜在应用。

2016 年，欧洲“肠道微生物组学联合行动计划”出炉。

2017 年底，“中国科学院微生物组计划”，由中科院微生物研究所牵头，进行人体与环境健康的微生物组共性技术研究，聚焦人体肠道、动物肠道和活性污泥微生物组相关研究。

（就在本章内容准备发送给编辑的那天下午，2019 年 5 月 30 日，Nature 和 Nature Medicine 发表多篇论文和评述文章，介绍和解读人类微生物组计划二期 iHMP 的重大成果。iHMP 代表了未来人类微生物组多学科研究的范式，提供独特的数据资源，并已经开始逐步阐明宿主-微生物组互作的机制。文章包括整合人类微生物组计划 iHMP 综述[99]，孕妇[100]和产妇[101]的阴道菌群，炎症性肠病[102, 103]的肠道微生物系统，糖尿病前期的宿主与菌群特征分析[104]及时间序列数据与精准医疗[105]等。）

2.3.2 宏基因组与人体健康

人类共生微生物也是影响宿主健康非常重要的环境因素。人类共生微生物在人类营养、消化、神经营养、炎症、生长、免疫以及保护机体免受外源病原菌的感染中发挥主要作用，大约 95%的人类共生微生物位于肠道，微生物群落开始被认为是与个体健康有关的“第二基

因组”。人体的肠道、皮肤、口腔等的宏基因组是最早受到关注也是被研究得最深入的微生物群落。从 2010 年至今，nature, science 和 cell 出版过大量有关微生物群落研究的专刊、综述和其他文章。自 2008 年起，研究者们通过宏基因组技术系统探索人体微生物群落与健康之间的关系，肠道的微生物群落与肠道疾病[106]、II型糖尿病[95]、肝硬化[107]、抑郁症[108]、自闭症[109]、帕金森综合症[110]、阿尔兹海默病[111]和类风湿性关节炎[112]的相关性愈加明确。研究还发现许多种癌症的起因[113-115]和化疗效果[116]也与微生物群落有关，化疗、免疫和微生物群落三足鼎立的关系[116]得到研究者们的重视。

因此微生物参与治疗是一种重要的治疗手段并展开了广泛研究[117, 118]，研究明确了肠道菌群与癌症的发生发展密切相关且肠道菌群帮助肿瘤的免疫治疗，但菌群与免疫治疗制剂相互作用的具体机制尚不清楚。通过菌群移植改变肠道菌群将会成为癌症治疗的新手段，但是仍存在许多未知因素。同时，粪菌移植(human microbiota transplantation)可辅助多种疾病的治疗，包括艰难梭菌感染、炎症性肠病、糖尿病、肝硬化和肠脑相关疾病等疾病。粪菌移植治疗炎症性肠病已被写入由中华医学会消化病学分会炎症性肠病学组发布的 2018 年《炎症性肠病诊断与治疗的共识意见》。治疗复发性艰难梭菌感染(Clostridium difficile infection, CDI)已被列入临床治疗指南[119]。许多研究揭示了粪菌移植在治疗难治性溃疡性结肠炎[120, 121]、克罗恩病[122]、便秘[123]、肠易激综合征(IBS) [124]、肝病[125]、自闭症[126]中的潜在价值。

2.3.3 宏基因组与环境

人类对环境的微生物群落分析的关注程度也很高，从 2007 年人类开始研究海洋水域的微生物群落[127]到今年我国科学家对洪湖水域的生态研究[128]；从全球表层土壤中微生物组的结构和功能研究[129]到人类发现土壤或许能成为最好的抗生素来源[130]；从微生物群落组成可能对气候的影响到[131]到深海极端环境的不寻常微生物群落可能解开地球早期生命之谜[132]。随着宏基因组技术在环境中的不断应用，微生物群落对自然界、对环境的影响能力被不断的发掘并受到重视。

2.3.4 宏基因组的其他影响

随着微生物群落研究的不断深入，宏基因组逐渐应用于农业、工业及渔业等领域中。目前基本采用 16s rRNA 测序数据开展农业、工业和渔业等领域的研究，应用大规模宏基因组测序进行研究的还不多见。

在农业中，土壤、根际及周边环境的微生物群落与植物的生长状况密切相关。中国微生物组计划中将农作物微生物组列为跨越转化临界点的现代生物技术[133]，用来研究农作物-微生物组-土壤环境之间的相互关系，包括益生菌及其功能基因对作物生长发育的影响，微生物组与农作物氮、磷、铁等元素的吸收，微生物组与植物先天免疫反应和抗多种环境胁迫的关系等。我国科学家还对水稻全生育期根系微生物组的变化规律进行了系统研究，为水稻根系益生菌的施用提供理论支撑[134]。研究者们还对柑橘根际微生物群落宏基因组[135]、豆科植物根瘤菌微生物群落[136]及植物整体微生物组[137]进行研究和探讨。

在工业中，针对来自不同国家的不同啤酒，进行微生物多样性和

野生酵母菌的类型分析[138]；针对沼气产生过程中微生物群落的变化不同阶段，对温度、所需的能量等因素进行分析[139]。宏基因组还应用于工业酶、生物活性化合物和抗生素、异性生物质（杀虫剂、致癌物等）的降解等工业中[140]。

在渔业中，研究者们关注鱼类[141]、虾类[142]和贝类[143]的肠道微生物对这些生物的各种性状（如耐温性、生长率、长肉率、抗病性及营养性和观赏性等）的影响。

2.4 人工智能在宏基因组中的发展前景

从现有的工作中，我们可以感受到人工智能、机器学习的相关方法和策略在宏基因组的分析和应用中起着举足轻重的作用。目前流行的深度学习等智能算法还处于较为初级和直接的应用，其性能与传统的机器学习和统计模型相比，本质上并未提升。大部分研究仍处在利用智能算法单独解决某个具体问题的阶段，如何构建完整的人工智能体系从而完成某个领域或者某个应用的宏基因组分析和决策任务，还需要进一步探索。不过，随着科学技术的发展和科学家研究的深入，相信 AI 在宏基因组科学问题的解决和工程技术的应用方面会有更加广阔的发展空间和应用前景。

第三章 人工智能与生物网络

3.1 人工智能与生物网络概述

人工智能是一门利用计算机模拟人类智能行为科学的统称。人工智能的很多目的是使计算机“更像人”。比如，科学家和工程师经常训练计算机使其能完成归纳、判断、总结、决策等人类行为的范畴。目前，实现人工智能最主要的方法是机器学习。机器学习是使计算机能够自动处理数据，从中学习规律，并对真实世界的任务进行预测和决策的一门学科。近年来，随着技术的发展，深度学习这一概念正在兴起。深度学习是实现机器学习的一种技术，其利用许多层次化的神经网络模型来解决更为复杂的问题。相比于传统的机器学习技术，深度学习有着层次化，可拓展，准确度高等一系列优点，被普遍视为下一代机器学习技术。而在人工智能这一大领域，深度学习也是最被寄予厚望的技术。

以应用范围来分类，人工智能可以分为通用人工智能与专业人工智能。通用人工智能指具备自主快速学习，达到甚至超过人类智慧的人工智能产物。由于受限于技术的发展，这类人工智能大多出现在科幻作品中，目前也看不到其成为现实的技术路径。专业人工智能是指在某一特定领域进行应用的人工智能，比如谷歌开发的会下围棋的AlphaZero，OpenAI 开发的会合作的电子竞技机器人，或者是支付宝刷脸识别中的高精度人脸识别器。在人工智能发展的浪潮中，专业人工智能技术成为主角，因此本章重介绍了专业人工智能在一个重要的

领域——生物网络方面的应用。

生物网络是一个非常宽泛的概念，任何适用于生物系统的网络都可以称之为生物网络。随着生物实验和数据的积累，越来越多不同种类的生物网络被定义和挖掘，例如基因转录调控网络，是指细胞内基因和基因之间相互作用关系所形成的网络。众多相互作用关系中，又特指基于基因调控所导致的基因间作用，例如蛋白质相互作用网络。蛋白质-蛋白质相互作用是指两个或两个以上的蛋白质分子通过非共价键形成蛋白质复合体的过程，将这些相互作用关系以蛋白为结点，是否相互作用为边就构成了蛋白蛋白相互作用网络，例如信号传导网络，是指参与信号传导通路的分子和酶以及其间所发生的生化反应所构成的网络。使用生物网络看待问题，许多情况下是为了研究一簇分子间的相互作用关系，例如基因调控网中，给定基因表达的量不仅由其他基因单独调控，更多的时候由一组行使生物功能的基因集合调控。图 3-1 为基因调控网络的工作示意图。使用基因调控网络不仅可以看清单个基因之间的关系，还能看出上下游基因的协同作用，例如在蛋白与蛋白相互作用网络中，不仅可以确定两个蛋白是否相互作用，还可以研究一组蛋白复合体之间的相互作用。从网络的视角看问题，为阐述生物过程背后的机理和研究新的生物功能提供了许多新的见解。同时，生物网络技术也为下游的生物实验提供了十分丰富的候选集合。

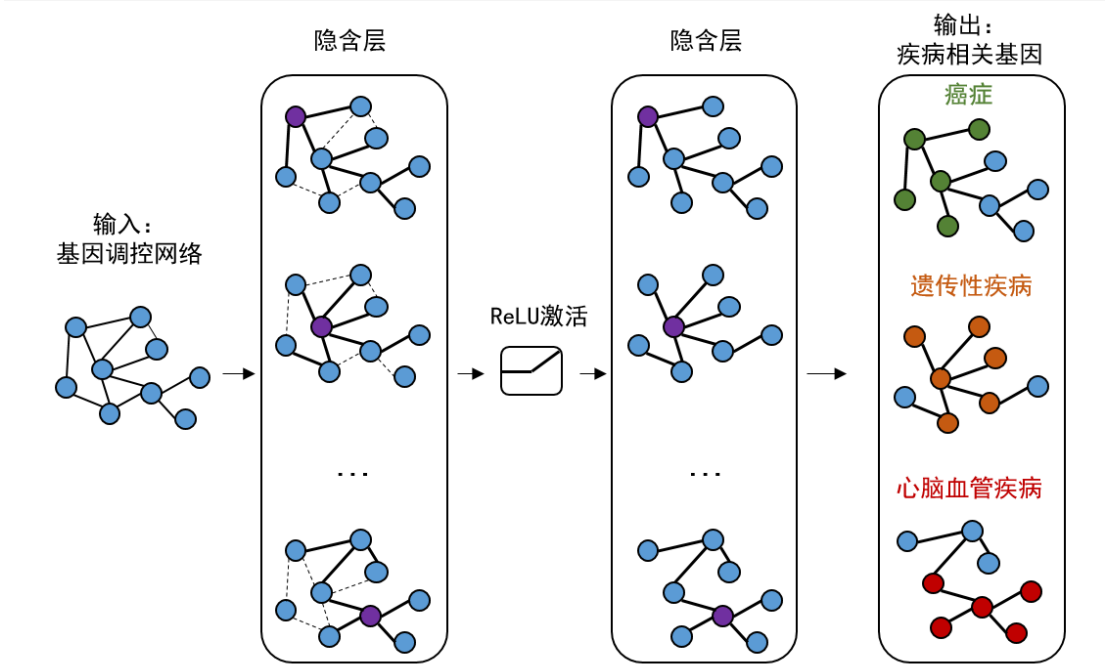

图 3-1 基因调控网络的工作示意图

近年来随着人工智能的快速发展，下一代机器学习技术如深度学习和表示学习，对生物网络的发展产生了巨大的影响。例如结合深度卷积神经网络和非线性降维等技术，越来越多的研究和应用将网络结构编码为低维表示。这些表示学习方法的本质是学习一个数据转换函数，其将网络的节点映射到低维向量空间中的点，也称为嵌入。基于深度非线性嵌入的表示学习方法，已经显著提升了网络科学中很多任务的水平。我们正处于人工智能和生命健康大数据的时代，实验数据的积累极大促进了生物网络中人工智能技术的应用。结合人工智能技术，从生物网络的角度看待问题，对理解生物功能背后的机理，开发药物，精准医疗等都有非常大的帮助。目前，结合人工智能技术来解决生物网络的问题正逐渐成为一个新兴并且有前景的研究方向。

3.2 人工智能在生物网络中的应用

3.2.1 加权基因共表达网络

寻找协同行使同一生物功能的基因集合对理解生物机理和疾病的治疗有重要的意义。因此，寻找行使特定功能的基因集合就成为了一个重要的问题。

在这样的背景下，利用基因与基因之间表达的相似性来定义共表达网络成为了一种较为通用的方法。其中，最具有代表性的是加权基因共表达网络分析(WGCNA, Weighted correlation network analysis)。WGCNA 是用来描述不同样本之间基因关联模式的生物网络方法，它可以用来鉴定高度协同变化的基因集，并根据基因集与表型之间的关联，鉴定候补生物标记基因或治疗靶点。因此，在疾病以及其他性状与基因关联分析等方面的研究中广泛应用。更具体地，WGCNA 算法首先假设基因网络服从无尺度(scale free)分布，然后根据基因共表达相关矩阵计算基因网络的邻接矩阵，再基于此邻接矩阵进行层次聚类，并划分出不同的基因模块。每个模块内的基因共表达程度高，而不同模块间的基因共表达程度低。由于模块内的基因共表达程度高，所以它们更倾向于协同行驶生物学功能，当这样的基因模块被定义后，可以用来开展下游的分析工作，如关联性状分析和代谢通路建模等。

WGCNA 简单易用，相比于神经网络等黑匣子算法，WGCNA 的参数具有更明确的生物意义，并且其自动做出的层次聚类图能清晰地反应基因模块之间的相似性与差异性。未来如何将人工智能技术与

WGCNA 相结合是一个值得探索的问题。

3.2.2 网络节点嵌入

生物网络的机器学习方法的主要挑战在于如何提取有关点和边的信息，并将并将这些信息合并到机器学习模型中。经典的机器学习方法通常依赖于汇总统计（例如结点的度）以及精心设计的特征，因此难以推广，传统汇总统计显示出了一定的局限。在此背景下，如何自动地将图的节点映射到一个低维的数值向量（称为节点嵌入），就成了一个非常重要的问题。一个优良的嵌入对解决后续的预测问题有着重大帮助。

直观上，我们希望将网络的节点映射到低维的数值向量，并且两个节点在低维空间上能保持他们在原网络中的相似性。根据以上简单想法，我们首先需要定义一个编码器，把图的节点映射到一个数值向量。其次，我们需要定义一个相似函数，用以衡量图中节点两两之间的相似性。编码器本质上是一个可以学习的函数即可（比如浅层的神经网络），相似函数则表明了对问题的定义。比如，我们可以通过两个节点是否相连；是否共享很多邻居节点；有相似的局部网络结构等来定义其相似性，这样导出的模型也有较大的差异。在此背景下，节点嵌入领域涌现出了一大批优秀的方法，比如 node2vec, DeepWalk, LINE, struc2vec 等等。

同时，基于节点嵌入的生物网络应用也纷纷涌现。总体的思路是，首先对给定的生物网络进行节点嵌入，然后基于嵌入的数值向量进行后续任务的预测。比如，Agrawal 等人成功使用节点嵌入技术预测了

蛋白和疾病的关系。他们首先假设相互交互的蛋白会对相似的表型有影响，然后使用 node2vec 把蛋白质相互作用(PPI)网络中的节点嵌入到低维向量。最后，他们使用嵌入的低维向量来预测蛋白和疾病的关系，取得了非常巨大的成功。同期，斯坦福大学的团队也使用 node2vec 技术成功预测了蛋白与蛋白间的交互关系。

总之，节点嵌入作为最近几年发展的人工智能技术，正在生物网络中逐渐崭露头角。由于生物网络的巨大数量和丰富的多样性，我们认为节点嵌入的自动特征学习将非常契合大部分的生物网络。所以我们预期这项技术会有非常光明的前景。

3.2.3 图神经网络

在 3.2.2 中，如果我们把节点到低维空间的编码器用适当的神经网络来表示，就近似得到了一个简单的图神经网络。虽然图神经网络和网络节点嵌入有一定相似性，但随着技术的发展，图神经网络也展现出越来越多的优点。图神经网络的表示法最早由 Gori 等人提出，早期的研究通过迭代的方式，利用循环神经结构传播邻居信息，直到达到一个稳定的不动点，来学习目标网络节点的表示。由于该过程计算复杂性大，难以推广，所以进展稍显缓慢。

近年来，图神经网络受到卷积神经网络在计算机视觉领域应供的成功经验，也得到了快速发展，大部分方法主要用于定义和改进网络中的卷积算子。这些方法大多属于图卷积网络，图卷积网络大致分为基于谱的做法和基于空间信息的做法。2013 年，Bruna 等人首次基于谱图理论设计了一种图卷积的变体，其通过对归一化图拉普拉斯矩阵

进行操作，在许多任务上有非常好的表现。自此，基于谱图的卷积网络的改进、扩展和逼近越来越多。基于谱的做法一般同时处理整个图，计算上难以并行和拓展，所以也有一系列的局限。另一方面，受到图片上的卷积启发，基于空间的图卷积神经网络迅速发展，其中一种具有代表性的方法是基于递归的方法，其主要思想是递归地更新节点的潜在表示，直到达到稳定的不动点。相比于之前的做法，基于递归的方法结合了规范化、门循环单元架构、异步和随机更新节点等技术，大大简化了计算过程。另一种代表性的方法是基于组合的空间图卷积网络，其使用聚合函数，通过聚合一个结点的邻域信息，不断更新迭代得到最终的表示。同时，此类图卷积结合抽样策略，可以不用在整个图上进行计算，只需要在批节点上进行计算，能够显著减少计算复杂度，展现出了很大的计算优势。除了图卷积神经网络，其他著名的方法还有图注意神经网络、图自动编码器、图生成网络和图时空网络等等。

得益于图神经网络的快速发展，生物网络中的问题也得到了一定解决。Zitnik 和 Leskovec 结合图卷积神经网络的相关技术，成功预测了组织特异性的蛋白功能。首先在每个组织里表示出一个蛋白质相互作用网络，然后使用图神经网络 GraphSAGE 来得到每个图中蛋白的数值嵌入。通过使用这些得到的嵌入，在下游的蛋白功能预测上得到了非常好的结果。

图神经网络的应用范围十分广泛，因此相关的研究大量开展，作为一个新兴的技术，图神经网络技术本身还有很大的发展空间，生物

网络本身就非常复杂，每个网络的数据背景不同、形态各异。如何将图神经网络的技术应用到更多的生物网络问题中，是一个很有前景的研究的方向。此外，如何结合生物网络本身的特点来设计生物网络特异的图神经网络，也是一个值得研究的问题。图神经网络作为目前最活跃的研究领域之一，我们期待见到其在生物网络中更多的应用。

3.2.4 异质性网络的嵌入

除了同质性的网络，生物网络中也有非常多的异质性网络，比如基因共表达网络以及其对应的蛋白质相互作用网络或者药物靶向网络等等，很多生物过程都是多个异质性网络共同作用的结果。研究不同异质性网络之间的协同作用，对理解生物机理、寻找药物靶标等具有重大意义。在此背景下，如何将一系列异质性的网络整合并映射到一个数值表示并进行下游任务分析，是一个非常重要的研究方向。

2017 年，Zitnik 等人提出了一种用来处理多层次网络的无监督节点表示学习方法 OhmNet，他们将 node2vec 的思想拓展到多层网络中，成功预测了多种组织特异性的细胞功能。也有一些异质性网络嵌入方法选择了更为通用的解决办法，例如 Dong 等人在 2017 年提出 Metapath2vec，一种可拓展的异质性网络表示学习方法。这种方法用元路径指导随机游走的邻居节点的选择，可以同时捕获不同网络之间的相似性，这种方法简单易用，在预测疾病相关的基因方面具有良好的前景。2018 年，Zitnik 等人提出 Decagon 网络，通过异质性网络的深入嵌入，成功地对复方药物的副作用进行了建模。更具体地，他们将蛋白质相互作用网络、药物蛋白靶向网络和药物-药物交互网络进

行预处理，通过图卷积网络自动学习并精准预测了复方用药的副作用，该做法相比于基准结果的准确度提升了 69%。

同时考虑多个异质性的网络一直是一个很实际的想法，然而生物网络本身具有复杂性，因此解决此类问题十分困难。近年来发展的异质性图嵌入技术为解决此类问题提供了一个良好的范例。如何结合生物问题，将此类技术推广应用于解决更多的生物网络问题，则是一个亟待发展的新方向。

3.3 人工智能在基因网络中的发展前景

我们认为，在人工智能和生物网络的交叉点上，存在着巨大的机遇的同时也面临着同样重大的挑战，解决该问题的关键的是需要高质量且大规模的数据集。虽然我们生活在一个生物医学大数据时代，可以收集许多不同层次的数据，但是这些数据对生物系统的描述过于详尽，以至于许多数据集在样本层面的数量级太小，无法被许多先进的算法（比如人工神经网络）利用。此外，由于生物系统本身的复杂性和生物实验的系统误差，收集到的数据的质量往往难以保证，其中数据信噪比低是生物大数据和其他数据相比差异性较大且难以克服的困难。

此外，许多人工智能模型的“黑匣子”特性为生物网络带来了额外的挑战。通常，人们很难从生物学的角度解释给定模型的结构、规则以及输出，极大限制了模型在提供对底层生物机制和功能网络架构的洞察方面的效用。尽管一些传统的机器学习方法如线性回归（岭回归，LASSO 和弹性网）可以告知研究人员每个特征的相对“重要性”，但

相比于神经网络等更精准的方法，这些可解释性是以预测精度的降低为代价的。所以，对于更精准的人工智能方法，例如深神经网络，隐含层的生物意义，或是输入正规化对结果的影响，都更加难以捉摸。或许，模型的可解释性和预测的精度本身就是难以调和的矛盾—复杂模型总是更难解释。当然，模型特征的解释是当前人工智能中的一个公开的困难的挑战。如何发展出新的方法，将深度学习的“黑匣子”转化为“白匣子”，从生物学的角度来说意义重大。

更进一步的，利用人工智能对生物网络的预测结果，并不总是可以进行实验验证的。首先，对网络的预测结果可能是以节点（基因）为单位的，然而为每一个预测都进行生物实验是极其昂贵甚至无法实现的。其次，由于生物网络节点之间的相互作用，行驶功能的可能是一组节点而不是一个节点，这进一步增加了实验的复杂性。这就导致了我们不能完全评估人工智能方法在生物网络相关方面的性能，而只能得到一个近似的估计。如何结合实验的需求来应用人工智能的模型，也有着分非常巨大的前景。

如何将人工智能更好地应用于基因编辑，首先在资源允许的情况下，应该尽可能地提高收集到的数据的质量。比如使用更纯净的生物样本，或者在测序实验中进行更深的测序。数据的质量会直接影响到最后的结果。另一种可能是生成具有真实属性的数据。在深度学习背景下的图像应用中，这通常是通过生成对抗网络实现的，该网络学习创建类似于训练数据的数据集，对抗生成网络由一个生成模型和一个识别模型组成。两个神经网络之间通过竞争不断学习，直到无法区分

生成的数据集与训练数据集。将这种方法拓展应用于生物组学数据前景广阔。然而由于生成对抗网络训练的不稳定性，目前还没有与生物数据相关的通用大规模的对抗生网络应用。其次，由于模型的可解释性和预测的精度本身就是难以调和的矛盾，建议区分生物任务来看待模型的可解释性。在可解释性需求比较弱的任务中，尽量用精度高的模型；在可解释性需求强的任务中，可以适当降低模型的复杂度，牺牲结果的精确度进行解释。最后，建议根据生物网络的任务需求来设计人工智能模型，结合真实的应用需求，平衡计算复杂性、模型可解释性和预测精度来设计模型。

探究和利用生物网络背后的原理，我们还有很长的路要走。在利用人工智能技术研究生物网络的过程中，我们会更深入地理解这些复杂的生物网络，期待网络生物学的美好未来。

第四章 人工智能与基因编辑

4.1 人工智能与基因编辑概述

随着高通量测序技术以及生物信息学的发展，人类对基因的认识不断加深。2001 年，人类基因组工作草图发表，标志着人类基因组计划取得了里程碑式的成功[169]。到 2003 年，人类基因组计划的测序工作全部完成，测定了组成人类染色体中所包含的 30 亿个碱基对组成的核苷酸序列，绘制了人类基因组图谱，并对其中载有的基因及其序列进行注释，从此人类进入了后基因组时代或功能基因组时代[170]。在这个时代，研究目标相对于前基因组时代有了重大变化，在已知基因组结构的情况下，探究基因组的功能及调控机制并将这些研究应用于临床医学、药物开发等下游领域成为了新的重要目标[171]，这时的核心科学问题主要包括了基因组的多样性、基因组的表达调控以及蛋白质产物的功能等。这些研究将为人们深入理解人类基因组的逻辑构架，基因结构与功能的关系、个体发育、生长、衰老和死亡机理、神经活动和脑功能表现机理、细胞增殖分化和凋亡机理，信息传递和作用机理、疾病发生发展的基因及基因后机理（如发病机理、病例过程）以及各种生命科学问题提供共同的科学基础。因此，功能基因组学方面的研究成果不仅具有巨大的科学意义，而且具有广泛的应用前景[172]。

重组 DNA 技术发展于 20 世纪 70 年代，是一项划时代的生物学技术。这项技术使得研究者获得了修改 DNA 分子的能力，使得研究

基因和利用基因开发新型医药和生物学技术成为可能[173]。近年来，一种新型基因组工程技术地进步正在逐步推动一次生物研究领域的革新，该工程技术不同于以往的 DNA 修饰需要先将其从基因组中提取出来。目前结合人类基因组图谱以及功能基因组学研究所产生的一系列成果，研究者几乎可以在所有生物体的内生环境中直接编辑或调控 DNA 序列的组成和功能，阐明其基因组层面的功能组成并确定其遗传学意义上的因果联系。这种技术就是基因编辑技术，一种在活体基因组中进行 DNA 插入、删除、修改或替换的一项生物工程技术[6][174]。该技术与早期的重组 DNA 技术的不同之处在于，早期的重组 DNA 技术是在宿主的基因或基因组中进行遗传物质的随机插入，而基因编辑技术则是在特定的位置插入、删除或修改基因片段。在基因编辑技术中，早期以锌指核酸内切酶(zinc-finger nucleases, ZFN) [175-178]和类转录激活因子效应物核酸酶(transcription activator-like effector nucleases, TALEN) [179-181]为代表的序列特异性核酸酶技术因其能够高效地进行基因组定点编辑，在基因研究、基因治疗和遗传改良等方面展示出巨大的潜力[175-177, 182-184]。但锌指核酸内切酶和类转录激活因子效应物核酸酶这两项基因编辑技术仍存在成本高、制作繁琐、效率低下和特异性不强等诸多劣势。CRISPR-Cas 基因编辑系统是继锌指核酸内切酶和类转录激活因子效应物核酸酶这两代基因编辑技术之后发展出的第三代基因组定点编辑技术[174, 185-188]。与以锌指核酸内切酶和类转录激活因子效应物核酸酶为代表的前代技术相比，CRISPR-Cas 基因编辑系统具有成本低、制作简单、效率高以及特异

性较高等优点。这些优势使得 CRISPR-Cas 基因编辑系统成为进行功能基因组学研究和阐明基因组功能的有力技术手段。CRISPR-Cas 基因编辑系统已成为一种重要的生物学研究技术，在科研、医疗等领域有着广阔的应用前景[174, 187, 189-191]。

尽管 CRISPR-Cas 基因编辑系统相较于锌指核酸内切酶和类转录激活因子效应物核酸酶这两代基因编辑技术在效率和特异性方面均有了显著的提升，但这种提升仍然有限，即 CRISPR-Cas 基因编辑系统在实际应用中依旧会因为在特异性方面的不足导致基因编辑实验失败或引入噪音，大大限制了 CRISPR-Cas 基因编辑系统在生物医学研究等领域的应用。因此，CRISPR-Cas 基因编辑系统的优化对于拓宽 CRISPR-Cas 基因编辑系统在功能基因组学、临床医学和生物医药开发等研究方面的应用有着重大意义。

4.2 CRISPR-Cas 基因编辑系统概述

CRISPR-Cas 系统于 1987 年在细菌中被发现[192]，在 2007 年被证实为一种细菌及古细菌中的适应性免疫系统[193]。近几年 CRISPR-Cas 系统的多个亚型被逐渐改造为具有高可用性的基因编辑工具[194]，可以达到较上一代基因编辑技术 ZFN 以及 TALEN 而言，效率和特异性更高。因此，CRIS=PR-Cas 基因编辑系统已成为目前一项重要的生物技术，在功能基因组学研究、基因治疗等多个领域有广泛的应用。因此，优化该系统效率和特异性将会对生物学以及临床医学等相关领域的研究产生积极的影响。

随着数据科学的不断发展，机器学习已成为目前计算机科学的热

门技术。将机器学习理论应用于 CRISPR-Cas 系统中向导 RNA 的设计优化是可行的，通过收集基于 CRISPR-Cas 基因编辑系统的实验数据建立机器学习模型，利用这些数据训练机器学习模型，进而获得最终的模型。利用这些模型预测候选的向导 RNA 的效率或特异性，最终获得优化的向导 RNA，即可达到优化设计的目的。这种数据驱动方法能够直接从已有的数据中通过算法自动归纳对预测结果产生影响的重要特征进行量化预测，而且随着实验数据收集量的增加，模型预测的准确度也能不断提高。

通过机器学习模型进行特征选择所发现的 CRISPR-Cas9 基因编辑系统的影响因素将在理论上为该系统的进一步优化提供参考，而预测模型将为实验中向导 RNA 的设计提供优化方案，并使得需要高效率、高特异性的 CRISPR-Cas9 基因编辑实验成为可能。

4.2.1 CRISPR-Cas 基因编辑系统的来源与发展

CRISPR 序列是细菌及古细菌中的一种基因结构，这种基因结构在早期研究中被认为与细菌和古细菌的物种多样性相关。整个 CRISPR 基因座通常是由一组 CRISPR 相关基因以及标志性的 CRISPR 阵列构成，其中 CRISPR 相关基因为 Cas 基因，这种基因会被转录翻译成核酸酶从而产生功能，而标志性的 CRISPR 阵列是由一系列有间隔的重复序列所构成，而重复序列之间的间隔部分则是一种可变序列。研究发现这些可变序列是由外源性的遗传物质片段构成[195, 196]，CRISPR-Cas 系统的整体发展历程参见图 4-1[174]。

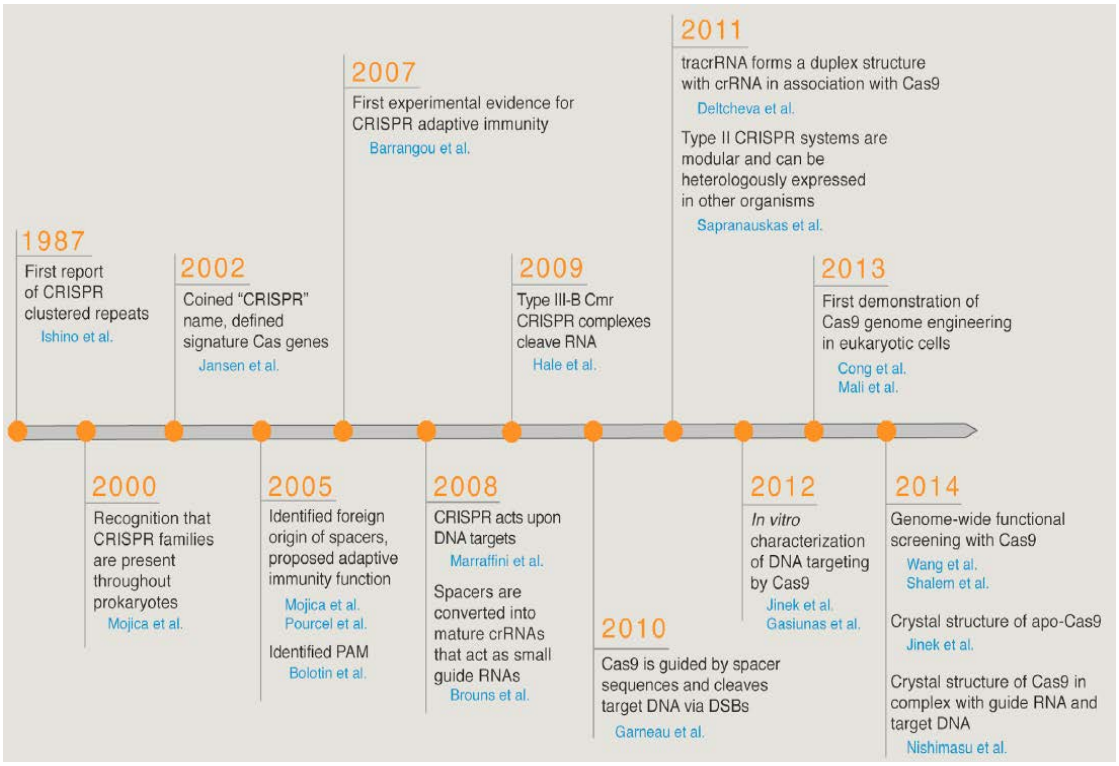

图 4-1 CRISPR-Cas 基因编辑系统的发展历程

1987 年，Ishino 等人在进行 iap 酶参与大肠杆菌碱性磷酸酶同工酶转化的研究工作中报道了在 iap 基因的下游发现以 29nt 为重复单元的重复序列。与一般的串联重复序列不同，这种重复序列含有五个中间插入的 32nt 非重复序列[192]。随着测序微生物基因组的增加，研究者在多种细菌和古细菌菌株的基因组中发现了类似的间隔重复序列。2000 年，Mojica 等人对这种出现在超过 40%的细菌以及 90%的古细菌中的间隔重复序列进行了分类，并把其归类为一种特定的间隔重复序列元件家族[197]。2002 年，Jansen 和 Mojica 正式将 CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)统一用于描述这种出现在微生物基因组的由间隔性的重复序列所组成的基因座[198]。后来有研究者将 CRISPR 基因座重复序列附近的若干被认为与 CRISPR 功

能相关的保守基因，即 Cas 基因作为不同类型 CRISPR 系统的主要分类依据[199]。2005 年，多个研究工作者系统分析了 CRISPR 基因座中的 CRISPR 阵列，得出了 CRISPR 阵列中间隔于重复序列之间的可变序列是出自于微生物体外或是噬菌体相关的来源的结论[195, 196]。而其他的相关研究则发现若 CRISPR 基因座在微生物中被转录，且当细菌的 CRISPR 阵列中间隔的可变序列含有病毒基因组时，使用同样基因组的病毒便无法有效感染该细菌[195]。这些发现表明 CRISPR 基因座在微生物中起到了记忆性的免疫防御作用，而这种免疫作用可能是通过 CRISPR 阵列中的间隔可变序列与噬菌体中的相关 DNA 进行碱基配对发生的[195, 196]。就 CRISPR 基因座的作用机制而言，研究者提出了若干猜想，包括 CRISPR 阵列中的间隔可变序列经过某些生物学过程进入 RNA 干扰的相关通路，从而对抗并降解外源性病毒基因转录本，以及 CRISPR 阵列中的间隔可变序列通过与外源性病毒 DNA 进行碱基配对来引导由 CRISPR 相关基因表达的 Cas 核酸酶对外源性病毒 DNA 进行切割[200]。

2007 年，Barrangou 等人报道了首个在嗜热链球菌中实验验证的 II 型 CRISPR 天然适应性免疫系统，该研究验证了这种基于 CRISPR 基因座中 CRISPR 阵列的间隔可变序列决定其打靶特异性的免疫系统[193]，大大加速了 CRISPR 天然适应性免疫系统的研究进程。Brouns 等人通过研究大肠杆菌中的 I 型 CRISPR 系统基因座，在 2008 年报道了 CRISPR 阵列被转录、转化为含有单个间隔可变序列的小 RNA 序列，又称 crRNA，这种 crRNA 能够使 Cas 核酸酶产生活性[201]。与

此同时，Marraffini 等人发现表皮葡萄球菌中的 III-A 型 CRISPR 系统所介导的适应性免疫机制可以阻断相关质粒的结合，这项工作证明了 Cas 核酸酶的活性对象为 DNA 而非 RNA[202]。激烈火球菌的研究中发现其含有的 III-B 型 CRISPR 系统可以实现由 crRNA 引导的 RNA 切割活性[203]。随着 CRISPR 基因座研究的深入，CRISPR 系统的结构和作用机制被进一步揭示。到 2008 年，Deveau 等人通过将噬菌体基因组中的 PAM 序列突变从而抑制 CRISPR 系统切割的实验证实了 PAM 序列对 CRISPR 系统的重要性[204]。此外，在 I 型和 II 型 CRISPR 系统中未发现 CRISPR 阵列的重复序列内含有 PAM 序列，表明微生物中的 CRISPR 系统避免了自我靶向的错误切割。

直到 2010 年，CRISPR-Cas 系统在细菌及古细菌等微生物中作为适应性免疫系统的基本功能和机制已经逐渐被阐明，这时也出现了多种基于 CRISPR-Cas 系统的生物工程技术，但这些技术主要应用于微生物研究，如抗噬菌体培养基以及细菌谱系进化树分型等，尚未用于基因编辑，然而这种情况并没有持续太长时间。Garneau、Deltcheva 等研究者对细菌中 II 型 CRISPR 系统的功能性机制研究证明了 II 型 CRISPR 系统的基础性组件对其行使基于由 RNA 介导的内源性 DNA 核酸酶活性的基因编辑的必要性。其中 Garneau 等人在嗜热链球菌中进行的相关遗传学研究表明 Cas9 蛋白是 II 型 CRISPR 系统所包含的 cas 基因簇中唯一介导目标 DNA 切割的核酸酶[205]。而 Deltcheva 等人的研究工作则揭示了 crRNA 在合成和处理中所产生的关键组分，即 tracrRNA (trans-activating crRNA),一种与 crRNA 结合的结构性非编码

RNA，用以促进基于 crRNA 引导的 Cas9 核酸酶的靶向性[206]。tracrRNA 与 crRNA 的结合、Cas9 核酸酶以及 RNA 酶 III 构成了将 CRISPR 基因阵列转录产物处理为成熟 crRNA 的必需条件[206]。这两项研究表明对于 II 型 CRISPR 系统而言，成熟的 crRNA、tracrRNA 以及 Cas9 核酸酶是其行使基本功能的三个必要组件。鉴于这种可定制的 RNA 引导的内源性 DNA 核酸酶可以类比锌指核酸内切酶和类转录激活因子效应物核酸酶在真核生物细胞基因组中进行基因编辑，研究者开始将 CRISPR-Cas 系统应用于基因编辑领域。

2011 年，Sapranauskas 等人报道了 II 型 CRISPR 系统的可移植性，他们将来自嗜热链球菌的 II 型 CRISPR 系统移植到大肠杆菌中，成功重构出具有活性的 II 型 CRISPR 系统[207]。2012 年，Jinek 等与 Gasiunas 等研究者发表的研究成果阐述了来自嗜热链球菌以及酿脓链球菌的 Cas9 核酸酶可以在体外由人工设计的 crRNA 引导切割对应的目标 DNA[194, 208]。更为重要的是，Jinek 等人的研究成果进一步修改了天然 II 型 CRISPR 系统进行了，即将天然 II 型 CRISPR 系统中的 tracrRNA 和 crRNA 融合为一种单链 RNA，又称向导 RNA (sgRNA), 能够成功行使 tracrRNA-crRNA 双 RNA 复合体的作用，介导 Cas9 核酸酶切割目标 DNA[194]。这种基于向导 RNA 的 CRISPR-Cas9 系统是目前 CRISPR-Cas9 基因编辑系统的主要结构模式。

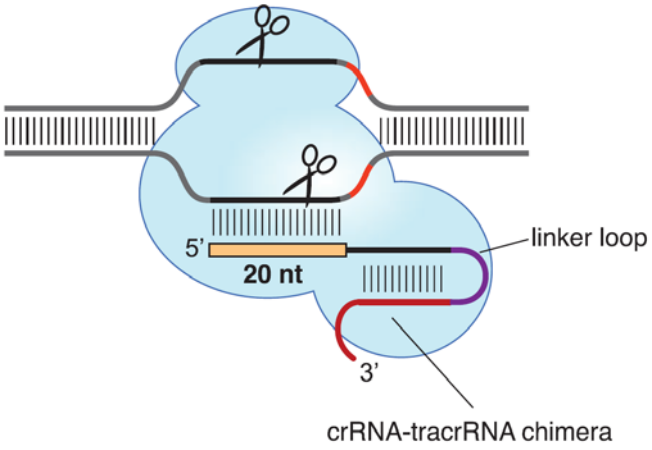

图 4-2 基于向导 RNA 的 CRISPR-Cas9 基因编辑系统

图 4-2[194]展示了 CRISPR-Cas9 基因编辑系统的经典模式。这种基因编辑系统主要包含三个部分：向导 RNA，即 crRNA 与 tracrRNA 链接而成的单链 RNA；Cas9 蛋白，即将 DNA 切割的核酸酶；PAM 序列，即目标 DNA 中紧邻在靶向区域后方的一小段特定序列，用于 CRISPR-Cas9 系统与目标 DNA 的初始识别和接触。

2013 年，Cong 等与 Mali 等研究者报道了将来源于嗜热链球菌以及酿脓链球菌的天然 II 型 CRISPR-Cas9 系统、不同来源的成熟 tracrRNA 与 crRNA 组合成的双 RNA 复合体以及由 tracrRNA 与 crRNA 融合的向导 RNA(sgRNA)所介导的人工 CRISPR-Cas9 系统成功用于哺乳动物细胞的基因编辑实验。实验包括切割目标 DNA 以产生非同源末端连接通路介导的 DNA 修复以及对目标 DNA 进行切割而产生同源重组介导的 DNA 修复[209, 210]。2014 年，Sander 等人使用了多个向导 RNA 同时靶向于多个目标基因，实现了在细胞中同时对多个基因的基因编辑的操作[211]。至此，CRISPR-Cas 基因编辑系统已经逐渐取代了锌指核酸内切酶和类转录激活因子效应物核酸酶这两

代基因编辑技术成为了新一代的常用基因编辑技术。目前CRISPR-Cas 基因编辑系统已经被全球大量研究机构使用，并用于多个研究领域，成为功能基因组学的热点生物学技术。

4.2.2 CRISPR-Cas 基因编辑系统的主要类型

自从 II 型 CRISPR-Cas9 系统被实验验证可以用于基因编辑以来，研究者又接连从不同细菌和古细菌中提取、改良了多种 CRISPR-Cas 系统用于基因编辑。以下重点介绍 CRISPR-Cas 系统的主要分型、分型依据以及每种分型相应的特点。

细菌和古细菌中的 CRISPR-Cas 基因座是不断进化的，从进化的角度而言,这些 CRISPR-Cas 基因座的编码行使了适应性免疫的作用。目前将大多数 CRISPR-Cas 系统的分型依据主要是标记的蛋白家族以及 CRISPR-Cas 基因座的特征，通过这些方法将 CRISPR-Cas 基因座划分为不同的大类、类型和子类型。现在这种分类方法已经囊括了 CRISPR-Cas 基因座的两个大类、五种类型以及十六种子类型。尽管这种分型方法已经十分有效，但仍有少数 CRISPR-Cas 基因座不属于这种分型方法的类型，因此该分型仍然有进一步完善的空间[212]。

进行 CRISPR-Cas 系统的分型时，对 cas 基因所产生的 Cas 蛋白进行鉴别是至关重要的。Cas 蛋白家族可以分为四个不同的功能模块：适应模块，用于获取间隔可变序列插入 CRISPR 阵列；表达模块，用于 crRNA 的加工以及目标 DNA 的结合；干涉模块，用于目标 DNA 的切割；辅助模块，用于调控以及其他 CRISPR-Cas 系统相关的功能。近年来，研究者在核心 Cas 蛋白(Cas1-Cas10)结构及功能方面获得了

大量的研究成果，这使得 Cas 蛋白在上述四个功能模块中的分类得以阐明。

对于适应模块而言，该模块在不同 CRISPR-Cas 系统中基本上是一致的，且是由 Cas1 和 Cas2 蛋白组成，可能存在的额外补充部分包括属于限制性内切酶超家族的核酸酶 Cas4 和在 II 型 CRISPR 系统中出现的 Cas9 蛋白[213, 214]。Cas1 蛋白是一种用于调节新的外源间隔可变序列插入到 CRISPR 阵列中的整合酶，其二级结构为阿尔法螺旋，其通过将 CRISPR 阵列中的特定重复序列位点切割，从而将外源间隔可变序列插入到 CRISPR 阵列中。而 Cas2 蛋白所扮演的角色则可能是 mRNA 干扰酶的同源蛋白质，其作用目前并不是非常清楚，但 Cas2 确实在大肠杆菌 I 型 CRISPR 系统中被发现与 Cas1 蛋白组合成复合体以完成外源 DNA 的整合适应过程。经过一系列研究发现尽管 Cas2 蛋白具有 RNA 酶以及 DNA 酶的活性，但其催化亚基对于整个适应性形成的过程并非是必需的，这表明 Cas2 蛋白并不是直接参与到适应性形成的过程[215]。

表达模块和干涉模块为具有多个亚单元的 crRNA-效应器复合体[216, 217]，但在 II 型 CRISPR 系统中则是一个例外。在 II 型 CRISPR 系统中，Cas9 蛋白同时起到了表达模块和干涉模块的作用。在表达阶段，crRNA 前体通常锚定在具有多个亚单元的 crRNA-效应器复合物或者在 II 型 CRISPR 系统中的 Cas9 蛋白上面，并在这里被处理为成熟的 crRNA。其处理方法有两种，一种是使用 RNA 内切酶 Cas6，这种情况通常出现在 I 型和 III 型 CRISPR 系统中；另一种是 tracrRNA

与 RNA 酶 III 相互配合实现同样的效果[206]。在干扰模块中，crRNA-效应器复合体（出现在 I 型和 III 型 CRISPR 系统中）或 Cas9（出现在 II 型 CRISPR 系统中）将核酸酶活性与特定的 RNA 锚定区域结合起来，使得目标的识别依赖于间隔可变序列的碱基构成。而目标 DNA 的切割在不同类型的 CRISPR-Cas 系统中是有所不同的，例如在 I 型 CRISPR-Cas 系统中，Cas3 核酸酶催化[218]目标 DNA 的切割，而在 III 型 CRISPR-Cas 系统中，同样的功能是由 Cas7 与 Cas10 蛋白共同组合完成的[219]。与之相应的，在 II 型 CRISPR-Cas 系统中，目标 DNA 的切割是由 Cas9 蛋白独立完成[194]。

辅助模块是多种蛋白质与结构域的组合，在这些蛋白质中除了 Cas4 蛋白，其他 CRISPR-Cas 系统的核心蛋白均非常少见。Makarova 等人的研究工作表明 Cas4 蛋白在 CRISPR-Cas 耦联的程序性细胞死亡中发挥一定的作用[220]。辅助模块还包括其他重要的组件，分别为用于调节 CRISPR-Cas 活性的含有 CARF (CRISPR-associated Rossmann fold)结构域的蛋白质[221]以及失活的 P 环 ATP 酶 Csn2[222]。Csn2 可以形成一个同源四聚体环，在环的中心孔中容纳线状的 DNA 双链，这种情况主要发生在 II 型 CRISPR 系统中[222, 223]。从上述描述中可以看出，II 型 CRISPR-Cas 系统中的 Cas9 蛋白在 Cas 蛋白家族中是一类非常特别的蛋白，这种蛋白兼具 RuvC 样核酸酶结构域以及 HNH 核酸酶结构域，这两个结构域均可以对 DNA 双链中的一个单链进行切割，更重要的是 Cas9 蛋白具有多个功能域，能够同时行使多项功能，包括适应、表达和干涉模块的相关功能。因此，Cas9 蛋白

已成为基因编辑中最常使用的 Cas 蛋白之一[174, 185, 186, 189-191]。Cas 蛋白的分型见图 4-3[212]。

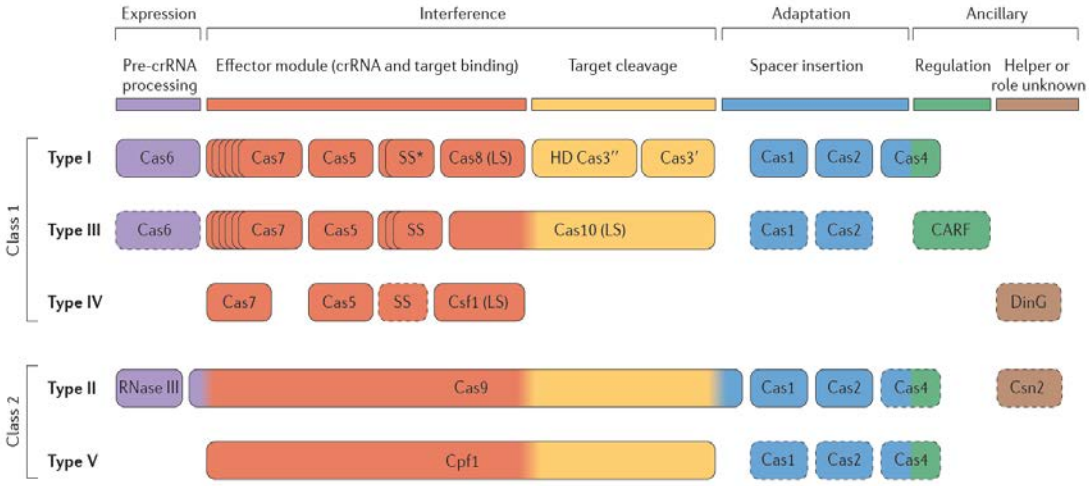

图 4-3 Cas 蛋白分型

CRISPR-Cas 系统有两个大类[212]，这两类系统的主要区别在于一类 CRISPR-Cas 系统的主要作用元件为由多个亚单元构成的 crRNA-效应器复合体，而二类 CRISPR-Cas 系统的主要作用元件为由单个单元构成的 crRNA-效应器复合体。一类 CRISPR-Cas 系统主要包括了 I 型 CRISPR-Cas 系统、III 型 CRISPR-Cas 系统以及 IV 型 CRISPR-Cas 系统。相对应地，二类 CRISPR-Cas 系统主要包括了 II 型 CRISPR-Cas 系统和 V 型 CRISPR-Cas 系统。不同类型乃至不同子类型的 CRISPR-Cas 系统通常具有不同的 crRNA 长度、PAM 结构以及 Cas 蛋白。下面将对五种类型的 CRISPR-Cas 系统作以介绍。

1. I 型 CRISPR-Cas 系统

所有的 I 型 CRISPR-Cas 系统都含有标志性的 cas3 基因，这种基因编码了一种受激单链 DNA 超家族 II 型解旋酶用于将双链 DNA 或 RNA-DNA 异体双螺旋解旋[224]。这种解旋酶的结构域通常会和来自

HD 家族的核酸内切酶相互结合，而这种 HD 结构域位于 Cas3 蛋白的 N 端末尾或者被另一种 cas 基因的变体编码[224]。

2. III 型 CRISPR-Cas 系统

相比于 I 型 CRISPR-Cas 系统标志性的 cas3 基因，III 型 CRISPR-Cas 系统的标志性基因是 cas10 基因。cas10 基因编码了一种多结构域的蛋白质，该蛋白质含有一个 Palm 结构域。这种 Palm 结构域是 RNA 识别基序的一个变体，同时也与多种核酸聚合酶与环化酶核心结构域同源。此外，这个结构域也构成了 III 型 CRISPR-Cas 系统中的 crRNA-效应器复合体中最大的亚单元。Cas10 蛋白在多种 III 型 CRISPR-Cas 系统中也显示了广泛的序列变异，因此可以用于进一步分类 III 型 CRISPR-Cas 系统。除了 Cas10 蛋白，III 型 CRISPR-Cas 系统的另一个标志是 Cas5 蛋白以及旁系同源来源的 Cas7 蛋白质[225]。

3. IV 型 CRISPR-Cas 系统

与 I 型 CRISPR-Cas 系统和 III 型 CRISPR-Cas 系统不同，IV 型 CRISPR-Cas 系统还处于推论中。IV 型 CRISPR-Cas 系统是在某些细菌菌株中发现的，与 III-B 型 CRISPR-Cas 系统相似，但具有不同的 CRISPR 基因座结构。目前的研究推测 IV 型 CRISPR-Cas 系统编码了一种最小的多亚单元 crRNA-效应器复合体。这个核糖核蛋白复合体很可能是由一个大型亚单元的部分降解产物、Csf1、Cas5 以及单个 Cas7 蛋白构成的，其中 csf1 基因是这种 IV 型 CRISPR-Cas 系统的标志基因。鉴于这类 CRISPR-Cas 系统中 crRNA-效应器复合体与 I 型 CRISPR-Cas 系统、III 型 CRISPR-Cas 系统的差异性，这类 CRISPR-Cas

系统被定义为 IV 型 CRISPR-Cas 系统[212]。

4. II 型 CRISPR-Cas 系统

II 型 CRISPR-Cas 系统与 I 型、III 型和 IV 型 CRISPR-Cas 系统有着很大的不同，其标志基因为 cas9。该基因如前所述，编码了一种多结构域的蛋白质，能行使外源 DNA 适应整合及目标 DNA 切割等作用的多功能蛋白质[213]。目前已经鉴别出的 II 型 CRISPR-Cas 系统均含有 cas1 和 cas2 基因，大多数 II 型 CRISPR 基因座能够编码 tracrRNA，与 crRNA 构成部分互补的复合体。就编码必需蛋白质的基因数量而言，II 型 CRISPR-Cas 系统是目前最简单的 CRISPR-Cas 系统，这也是其成为目前基因编辑主流系统的重要原因[174]。

5. V 型 CRISPR-Cas 系统

V 型 CRISPR-Cas 系统是一种与 II 型 CRISPR-Cas 系统不同的 CRISPR-Cas 二类分型，其主要代表是 cpf1 基因，一种在少数细菌和古细菌基因组中发现的基因。cpf1 基因常与 cas1，cas2 以及 CRISPR 基因阵列相邻。cpf1 基因编码的 Cpf1 蛋白是一种较大的蛋白质，一级序列包含了 1300 个氨基酸，其与 Cas9 蛋白相比，也包含了与 Cas9 蛋白相似的 RuvC 样核酸酶结构域用于 DNA 切割[226]。因此，继 II 型 CRISPR-Cas 系统成为目前主流基因编辑工具之后，V 型 CRISPR-Cas 系统的研究也在不断进行，目前也出现了基于 Cpf1 系统的基因编辑工具[226, 227]。

4.2.3 CRISPR-Cas 基因编辑系统的作用机制

上节介绍了 CRISPR-Cas 系统的主要分型，本节将详细介绍由 CRISPR-SpCas9 系统改造而来的 CRISPR-SpCas9 基因编辑系统的作用机制，其中 SpCas9 是指酿脓链球菌(Streptococcus pyogenes)来源的 Cas9 蛋白。如前所述，CRISPR-SpCas9 基因编辑系统包含三个主要组件分别为向导 RNA、PAM 序列以及 SpCas9 核酸酶。严格意义上来说，向导 RNA 是指 CRISPR-Cas 基因编辑系统中的整个 RNA 部分，但为了方便下文叙述，从这里开始，向导 RNA 单指与目标 DNA 结合的 RNA 部分而不包括行使结构稳定作用的其他 RNA 部分。对于 CRISPR-SpCas9 基因编辑系统而言，向导 RNA 的长度通常为 20nt，PAM 序列的主要模式为 NGG，少数情况可使用 NAG[228]。CRISPR-SpCas9 基因编辑系统的作用机制可由图 4-4[228]所表示。

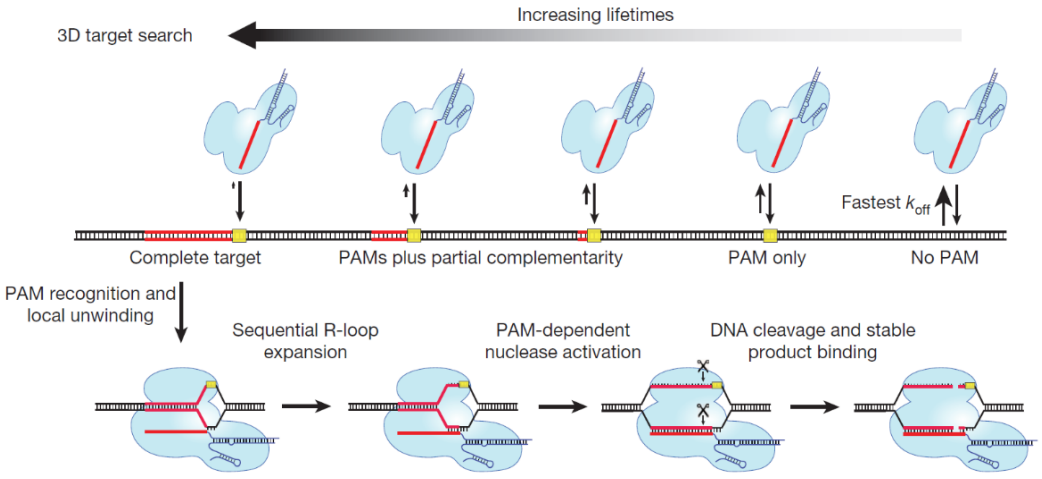

图 4-4　CRISPR-SpCas9 基因编辑系统作用机制

CRISPR-SpCas9 基因编辑系统在细胞中通常表现为 RNA-SpCas9 核糖核蛋白复合体。在 DNA 附近游荡的 RNA-SpCas9 核糖核蛋白复合体依据分子动力学原理与 DNA 分子在三维空间发生随机碰撞。当

碰撞区域未含有 PAM 序列时，RNA-SpCas9 核糖核蛋白复合体会迅速与发生碰撞的 DNA 脱离，但当碰撞区域含有 PAM 序列时，RNA-SpCas9 核糖核蛋白复合体中的 SpCas9 蛋白会尝试将碰撞区域的 DNA 双链解螺旋，同时使用向导 RNA 与解螺旋后的单螺旋 DNA 链进行互补。如果碰撞区域恰巧为该 RNA-SpCas9 核糖核蛋白复合体的目标 DNA 区域，则解螺旋后的单螺旋 DNA 链能够与其中的 20nt 向导 RNA 互补形成异种双螺旋，这种异种双螺旋会稳定整体结构，最终引导 SpCas9 蛋白切割目标 DNA 区域，在目标 DNA 区域产生双链断裂[228-230]。

CRISPR-Cas 基因编辑系统可以在目标 DNA 处引入 DNA 的双链断裂，之后靶细胞将可能进入两种 DNA 双链断裂修复途径。其一为非同源末端连接途径(Non-homologous end joining, NHEJ)，是 DNA 双链断裂修复的主要途径[176]，这种途径不依靠其它 DNA 序列直接修复双链断裂的 DNA。由于没有其它 DNA 序列作为模版，这种修复方式将在 DNA 的双链断裂处附近随机引入碱基的插入或者删除。当目标 DNA 是某个基因的外显子时，这些随机引入的碱基插入或者删除将有可能产生移码突变，造成该外显子表达出的蛋白质的失活，从功能上表现为该基因的失活。RNA 干扰是由用于干扰的小 RNA 与基因转录出的功能 RNA 进行双链结合，阻止功能性 RNA 发挥应有的功能，例如翻译或者调控功能。因此，RNA 干扰是在转录后的层面减少有效的功能 RNA 的产生，对目标基因进行下调[225]。而产生非同源末端连接途径修复的 CRISPR-Cas 基因编辑系统，会在 DNA 层面直接改

变原始基因，使目标基因永久性失活，这便是基因敲除。与不使用其它 DNA 序列作为模版的非同源末端连接途径相反，CRISPR-Cas 基因编辑系统产生出 DNA 双链断裂后的第二种修复方式将借助其他 DNA 序列对目标 DNA 的双链断裂位点进行修复。这种作为模板的 DNA 序列既可以是内源性 DNA 序列，也可以是来源于外部的 DNA 序列[231, 232]。内源性 DNA 序列模版，如二倍体细胞的另一个未断裂的同源染色体，可以作为修复模版，修复双链断裂的同时不引入随机的碱基插入或删除，而外源性 DNA 序列模版也可以产生类似的效果。值得注意的是，外源性的 DNA 序列模版可以与原始 DNA 序列有所不同，使其修复结果是与注入的外源性 DNA 序列模版相同或部分相同，这便是基因编辑。

4.3 常见 CRISPR-Cas 基因编辑系统优化工具

CRISPR-Cas 基因编辑系统经过近几年的发展已广泛应用于多种细胞系中，这种基因编辑技术主要用于以下几个基因层面的操作：基因敲除(gene knock-out)、基因插入(gene knock-in)、基因表达抑制(CRISPRi)以及基因表达激活[189, 190, 233]。如前所述，在 CRISPR-SpCas9 基因编辑系统中，向导 RNA 的主要作用是引导 SpCas 核酸酶至目标 DNA 位点进行切割，而 SpCas9 核酸酶对目标 DNA 区域的初始识别是由 PAM 序列决定的，其对应的 PAM 序列模式是 NGG。不同类型的 Cas9 核酸酶则会具有不同的 PAM 序列模式，表 4-1 展示了多种常见的 Cas9 核酸酶及其变体所对应的 PAM 序列[234]。

表 4-1　常见 Cas9 核酸酶及其变体所对应的 PAM 序列

Cas9 核酸酶及其变体	PAM 序列模式
SpCas9	NGG
SpCas9 D1135E	NGG
SpCas9 VRER	NGCG
SpCas9 EQR	NGAG
SpCas9 VQR	NGAN 或 NGNG
SaCas9	NNGRRT 或 NNGRR
Nm	NNNNGATT
St	NNAGAAW
Td	NAAAAC

目前，随着 CRISPR-Cas 基因编辑技术的不断发展，CRISPR-Cas 基因编辑实验的数据也在不断增多，引发的计算生物学问题也随之而来。CRISPR-Cas 基因编辑系统在进行基因敲除、基因插入、基因表达抑制以及基因表达激活等基因操作时面临的主要挑战是如何设计出同时具有高效率和较低脱靶效应的向导 RNA，这已经成为能否成功进行 CRISPR-Cas 基因编辑实验的关键。由于 CRISPR-SpCas9 基因编辑系统向导 RNA 设计的优化主要在提高打靶效率和提高特异性，即降低脱靶发生概率两个方面，因此，下文将详细介绍这两个部分已有的 CRISPR-SpCas9 基因编辑系统优化工具。

4.3.1 CRISPR-SpCas9 基因编辑系统打靶效率优化工具

影响 CRISPR-SpCas9 基因编辑系统打靶效率的因素有很多，如 PAM 序列的位置、向导 RNA 与目标 DNA 的实际序列以及切割后由 DNA 非同源末端连接途径修复所产生的 DNA 插入或删除情况。另外，目标 DNA 区域的表观遗传学性质，如染色质开放程度，也有可

能影响CRISPR-SpCas9基因编辑系统对目标DNA的切割效率[235, 236]。Xu 等人的研究指出，目标 DNA 序列会影响 Cas9 蛋白与目标 DNA 之间的亲和力[68, 69]。具体而言，在人类以及小鼠细胞中，当目标 DNA 近 PAM 序列侧第一位或第二位为鸟嘌呤 G 时会提高 CRISPR-SpCas9 基因编辑系统对目标 DNA 的切割效率。而相对应地，当胸腺嘧啶 T 出现在目标 DNA 近 PAM 序列侧第四位时会导致 CRISPR-SpCas9 基因编辑系统对目标 DNA 的切割效率下降[68, 69]。另外，目标 DNA 与向导 RNA 对应区域的下游碱基序列也可能对 CRISPR-SpCas9 基因编辑系统的打靶效率产生影响[236, 238]。目前还没有研究显示目标 DNA 与向导 RNA 对应区域的上游碱基序列对 CRISPR-SpCas9 基因编辑系统的打靶效率产生影响。而 Chari 等人对多个物种基因组中多个位点的向导 RNA 打靶效率的研究也表明目标 DNA 位点的染色质开放程度以及向导 RNA 的序列组成是影响 CRISPR-SpCas9 基因编辑系统打靶效率的重要因素[239]。

另一方面，CRISPR-Cas9 基因编辑系统不仅可以用于基因敲除，还可以作为转录抑制元件或转录激活元件。这种元件通常会将 Cas9 蛋白的核酸酶活性结构域失活为 dCas9 蛋白，然后将这种 dCas9 蛋白与相应的作用蛋白结合形成效应器。这种 CRISPR-Cas9 基因编辑系统的变体又被称为 CRISPRi（转录抑制元件）以及 CRISPRa（转录激活元件）。在这些系统中，Cas9 蛋白首先与转录抑制蛋白或转录激活蛋白结合为蛋白复合体，然后通过向导 RNA 引导其与目标基因的相应调控元件如启动子结合，从而沉默或激活[240]目标基因。在

CRISPRi 与 CRISPRa 系统之中，序列背景信息对其抑制或激活效率的影响目前还尚未阐明，其原因可能是因为 CRISPRi 与 CRISPRa 的主要目标是启动子等基因调控元件而非基因的编码区，这也导致了针对 CRISPRi 与 CRISPRa 向导 RNA 优化所进行的特征选择与 CRISPR-SpCas9 基因敲除有着明显得不同。与 CRISPR-SpCas9 基因敲除类似，CRISPRi 与 CRISPRa 系统也倾向于目标 DNA 中出现嘌呤的向导 RNA，但不同之处在于，CRISPRi 与 CRISPRa 系统在近 PAM 侧的第三位没有出现胞嘧啶的富集。另外，CRISPRi 与 CRISPRa 的效应器结构域与 CRISPR-Cas9 基因编辑系统也有着很大差异，而这些直接与目标 DNA 发生作用的结构域可能是影响其打靶效率的决定性因素[236, 241]。目前只有很少一部分工具应用于 CRISPRi 与 CRISPRa 基因调控系统的向导 RNA 设计进行预测和优化，这类工具包括 CRISPR-ERA[242] 和 SSC[236]，这种情况出现的主要原因是目前 CRISPRi 与 CRISPRa 实验数据的积累不足，无法生成稳定的向导 RNA 优化规则，所以针对 CRISPRi 和 CRISPRa 系统向导 RNA 的优化设计还需要进一步的数据积累和更加系统地分析其影响特征。

基于目前多个研究团队已经发表的 CRISPR-SpCas9 基因敲除实验数据，不同研究者对 CRISPR-SpCas9 基因编辑系统中的向导 RNA 打靶效率，优化设计了不同的向导 RNA 设计规则以及优化工具。这些向导 RNA 优化设计工具可以分为三类，其一为序列比对型，即基于 PAM 序列将向导 RNA 与给定的基因组进行简单的序列比对；其二为人工规则型，即人工选择出向导 RNA 的若干特征，如 GC 含量、

外显子位置等，建立打分函数用以预测给定向导 RNA 的打靶效率；其三为数据学习型，即选择向导 RNA 的不同特征基于 CRISPR-SpCas9 系统基因敲除实验数据，通过机器学习方法训练机器学习模型，对给定向导 RNA 的打靶效率进行预测。后两类优化工具常常优于第一类工具，原因就在于后两类模型考虑了更多的特征，包括序列特征以及其他相关特征。另外，不同的向导 RNA 优化工具都具有适用于自身的优化情景，如 CRISPRseek[243]和 Cas-OFFinder[244]适用于基于 PAM 序列模式的向导 RNA 搜索，且这两种工具都支持多种 PAM 序列模式，并不局限于 CRISPR-SpCas9 基因编辑系统使用的 NGG 模式的 PAM 序列。此外，数据学习型工具，如 sgRNA-designer[245], CRISPR MultiTargeter[246], WU-CRISPR[247], sgRNA Scorer[239], SSC[236], CAGE[248]以及 DeepCRISPR[249]均可用于向导 RNA 打靶效率的预测，设计高打靶效率的向导 RNA。其他的优化工具，如 E-CRISPR[250], CRISPR-ERA[242], Protospacer Workbench[251]以及 CRISPR Library Designer 也提供了不同的向导 RNA 优化目标，这其中 E-CRISPR 适用于快速预测向导 RNA 打靶效率，CRISPR-ERA 适用于 CRISPRi 以及 CRISPRa 系统的效率预测，Protospacer Workbench 将向导 RNA 的优化扩展到了人类与小鼠以外绝大多数物种的基因组，CRISPR Library Designer 适用于向导 RNA 库的建立。也有一部分向导 RNA 优化设计工具则适用于特定的物种，如 CRISPR-P[252]适用于植物基因组，flyCRISPR[232]适用于果蝇基因组以及 EuPaGDT[253]适用于病原菌基因组。

4.3.2 CRISPR-SpCas9 基因编辑系统脱靶优化工具

CRISPR-SpCas9 基因编辑系统脱靶优化的一个重要前提是可以对 CRISPR-SpCas9 基因编辑系统所产生的实际脱靶位点进行实验检测。目前，已经出现了若干种实验技术可以对全基因组中由 CRISPR-SpCas9 基因编辑系统造成的脱靶位点进行检测，这些技术的原理各有不同，部分技术是基于对与 DNA 进行结合的 CRISPR-SpCas9 基因编辑系统中的重要蛋白如 SpCas9 蛋白进行捕获，部分技术是对 CRISPR-SpCas9 基因编辑系统所产生的 DNA 双链断裂进行检测[254]，但这些技术均需要进行基因组测序和相应的数据分析从而获得全基因组的脱靶位点信息。依据不同的技术原理，不同的脱靶位点检测技术会产生不同的脱靶位点分布结果。最早被用来检测脱靶位点的技术是染色质免疫共沉淀测序技术(ChIP-seq)，这种技术主要检测了与 DNA 结合的 SpCas9 蛋白，ChIP-seq 实际检测的位点是 DNA 与 SpCas9 蛋白的结合位点而并非 SpCas9 核酸酶所切割的 DNA 双链断裂位点[190, 230, 255-257]，使得 ChIP-seq 实验获得的结果与真实脱靶情况有一定出入。因此，多个研究组针对 DNA 双链断裂开发了相应的检测技术，如 Digenome-seq[258], Guide-seq[259], IDLV 捕获[260]，HTGTS（高通量全基因组易位检测）[261]以及 BLESS（原位断裂标记富集测序检测）[262]。

目前大多数 CRISPR-SpCas9 基因编辑系统脱靶优化工具只是简单地使用错配比对搜索潜在的脱靶位点，并未对这些潜在脱靶位点的脱靶发生率进行预测和讨论，对潜在脱靶位点的脱靶发生情况的进行

预测的工具，相对于对打靶效率进行预测的工具而言比较少见。就当前已经发表的 CRISPR-SpCas9 系统脱靶预测工具而言，这些预测工具在对脱靶发生的预测上依旧没有达到较高的准确度，表现为精确率较低，现有脱靶预测工具的评测工作也指出目前的脱靶预测工具所预测的脱靶位点与实际实验所获得的真实脱靶位点还存在较大出入[254, 259]，不同工具预测的脱靶发生位点亦有较大的差异。这种现象发生的可能原因如下，其一是不同预测工具在进行错配比对时使用了不同的错配数量或比对工具；其二是目前全基因组脱靶分布实验数据还有待积累。

表 4-2[234]列举了常见的 CRISPR-SpCas9 基因编辑系统优化工具。

表 4-2 常见 CRISPR-SpCas9 基因编辑系统优化工具

工具名称	类型	含有脱靶预测
sgRNAcas9	序列比对型	是
CRISPR/Cas9 gRNA finder	序列比对型	否
GT-Scan	序列比对型	是
CRISPRdirect	序列比对型	是
CRISPR RNA Configurator	序列比对型	否
CRISPRseek	序列比对型	是
CCTop	序列比对型	是
Cas-OFFinder	序列比对型	是
SSFinder	序列比对型	否
CRISPR-P	序列比对型	是
CRISPRer	序列比对型	否
CRISPRTarget	序列比对型	否
CRISPRfinder	序列比对型	否
flyCRISPR	序列比对型	是

工具名称	类型	含有脱靶预测
CRISPR gRNA Design tool	序列比对型	否
WGE	序列比对型	是
COD	序列比对型	是
CRISPOR	序列比对型	是
Protospacer Workbench	人工规则型	是
E- CRISP	人工规则型	是
CRISPR	人工规则型	是
CRISPR-ERA	人工规则型	是
CHOPCHOP	人工规则型	是
Cas9 design	人工规则型	否
EuPaGDT	人工规则型	是
CROP-IT	人工规则型	是
Cas-designer	人工规则型	是
sgRNA Designer	数据学习型	否
SSC	数据学习型	否
sgRNA Scorer	数据学习型	是
CRISPR Multitargeter	数据学习型	是
CRISPRscan	数据学习型	是
WU- CRISPR	数据学习型	是
CRISPR Library Designer	数据学习型	是
CAGE	数据学习型	是
DeepCRISPR	数据学习型	是

4.4 基于浅层学习的 CRISPR 打靶效率预测

CRISPR-SpCas9 基因编辑系统是目前应用最为广泛的基因编辑系统，该系统已被大量应用于多种细胞系之中进行基因敲除等操作，在这种基因编辑系统中，存在一个包含了目标DNA序列的向导RNA。

在 PAM 序列的帮助下，这种向导 RNA 可以引导 SpCas9 蛋白切割目标 DNA。当 SpCas9 蛋白将目标 DNA 切割之后，目标 DNA 被切割区域会产生平末端双链断裂[185, 191, 238, 263, 264]，断裂位点通常位于目标 DNA 接近 PAM 序列一侧的第三位与第四位之间，这种平末端的 DNA 双链断裂常常通过一种错误倾向性修复方式修复，即非同源末端连接途径，其结果是在断裂位点产生 DNA 的序列的插入或删除，从而导致基因的移码突变，最终将基因失活，这就是基于 CRISPR-SpCas9 基因编辑系统的基因敲除实验的基本原理[185, 238, 265]。由此可知，CRISPR-SpCas9 基因敲除实验的成功，很大程度上依赖于向导 RNA 与目标 DNA 区域的配对序列以及 SpCas9 核酸酶将目标 DNA 切割后由非同源末端连接修复所产生的 DNA 序列插入或删除[191, 265]。通常情况下针对单个基因设计 CRISPR-SpCas9 基因编辑系统所需的向导 RNA 有多种选择，而不同的向导 RNA 的打靶效率可能存在很大差异，因此能否设计出具有高打靶效率的向导 RNA 就是 CRISPR-SpCas9 基因编辑系统能否对目标 DNA 进行有效切割的重要因素之一。

目前已经有研究团队发表了关于 CRISPR-SpCas9 基因编辑系统打靶效率优化方面的研究。Wang 等人的研究指出向导 RNA 的打靶效率直接依赖于向导 RNA 近 PAM 侧序列中嘌呤与嘧啶的组合构成[237]。Doench 等人报道了目标 DNA 中位于向导 RNA 对应区域下游区域的碱基构成对向导 RNA 打靶效率的影响[238]。Xu 等人系统性的评估了在 HL60 细胞系以及小鼠胚胎干细胞 (mESC) 中进行的 CRISPR-SpCas9 基因敲除实验中向导 RNA 序列特征对其打靶效率的

影响[236]。Chari 等人开发了一种使用于体内实验的“库对库”方法，能够同时对向导 RNA 在多个基因组位点的打靶效率进行评估。此外，Chari 等人还指出了向导 RNA 序列对应的目标 DNA 的表观遗传学参数可能与向导 RNA 打靶效率存在联系[239]。

尽管目前在 CRISPR-SpCas9 基因编辑系统打靶效率优化方面的研究已经取得一定进展，但上述研究工作中所获得的影响 CRISPR-SpCas9 基因编辑系统打靶效率的 DNA 序列特征在不同的向导 RNA 库、细胞系及组织类型中是否具有普适性仍是一个疑问，出现这种情况的主要原因是目前公开发表和积累的 CRISPR-SpCas9 基因敲除实验数据总量较低。此外，已经公开发表和积累的 CRISPR-SpCas9 基因敲除实验数据常常来源于不同的实验平台、实验条件以及不同的细胞系样本。譬如，在 Xu 等人的研究从 HL60 细胞系与小鼠胚胎干细胞 CRISPR-SpCas9 基因敲除实验数据中得到的影响向导 RNA 打靶效率的序列决定因素互不相同，也发现某些细胞系特异性的序列特征倾向性在该研究中会被忽略[236]。Fusi 等人对多个向导 RNA 打靶效率预测模型在两套不同的 CRISPR-SpCas9 基因敲除实验数据集中进行了比较，其中一套使用了流式细胞方法对基因敲除实验的结果进行了检测，而另一套则使用抗性实验进行了检测[266]，结果显示不同模型在前者的预测性能均好于后者，这表明 CRISPR-SpCas9 基因敲除实验的实验结果正如 RNA-seq 等实验一样存在批次效应。考虑到实验条件的异质性，根据不同的实验条件建立定制化的向导 RNA 设计方案在这方面具有一定的优势。除了常见的

CRISPR-SpCas9 基因敲除情况检测方法，二代测序方法也被用于检测由 SpCas9 核酸酶切割所导致的 DNA 序列结构变化，如微同源结构模式、序列插入与删除以及倒置等[267]。多个研究组在其发表的研究中尝试使用这些 CRISPR-SpCas9 基因敲除实验的二代测序数据，分析这些测序数据来改善向导 RNA 的设计策略[263, 267-269]。尽管 CRISPR-SpCas9 基因敲除实验的基因组测序数据对打靶效率优化分析而言比较理想，但当前的计算工具对这种类型数据的分析利用还是有限的。

4.4.1 浅层打靶效率预测系统

4.4.1.1 系统整体概况

基于上述机器学习原理，打靶效率预测系统主要包含如下五个模块：向导 RNA 处理模块、全基因组数据处理模块、移码突变模式分析模块、向导 RNA 特征提取与打靶效率预测模块以及特征可视化模块。

打靶效率预测系统的工作流程主要有两步。第一是预测流程，其目的是对拟进行 CRISPR-SpCas9 基因编辑实验的向导 RNA 列表或目标基因的打靶效率进行预测和排序，最终得到优化的向导 RNA 作为实际 CRISPR-SpCas9 基因编辑实验中使用的向导 RNA。该工作流程示意图见图 4-5。

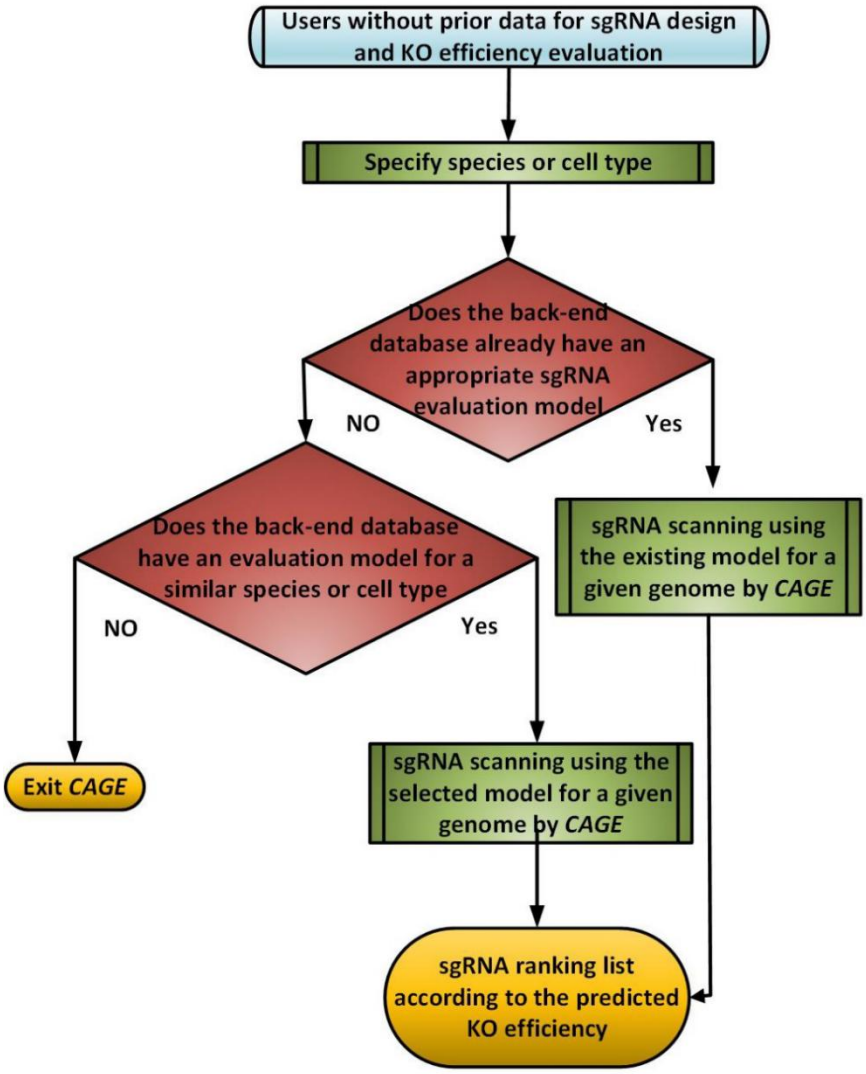

图 4-5 预测流程示意图

在这个工作流程中，打靶效率预测系统会根据拟进行的 CRISPR-SpCas9 基因编辑实验所使用的细胞系来判断预测流程可否进行。由于打靶效率预测系统中的预测模型是基于不同细胞系数据而定制化生成的模型，因此如果拟进行的 CRISPR-SpCas9 基因编辑实验所使用的细胞系无法在打靶效率预测系统中匹配到相应的模型，预测流程就无法进行，反之则可以进行，这时打靶效率预测系统会通过向导 RNA 处理模块对输入的向导 RNA 列表进行预处理，得到预处理结果后打靶效率预测系统调用向导 RNA 特征提取与打靶效率预测

模块，使用根据拟进行的 CRISPR-SpCas9 基因编辑实验所使用细胞系匹配的预测模型对输入的向导 RNA 进行打靶效率预测，最终获得预测结果。

第二个工作流程是模型训练流程，其目的是通过具有向导 RNA 打靶效率标签的 CRISPR-SpCas9 基因敲除实验数据建立 CRISPR-SpCas9 基因编辑系统打靶效率预测模型。该工作流程见图 4-6。

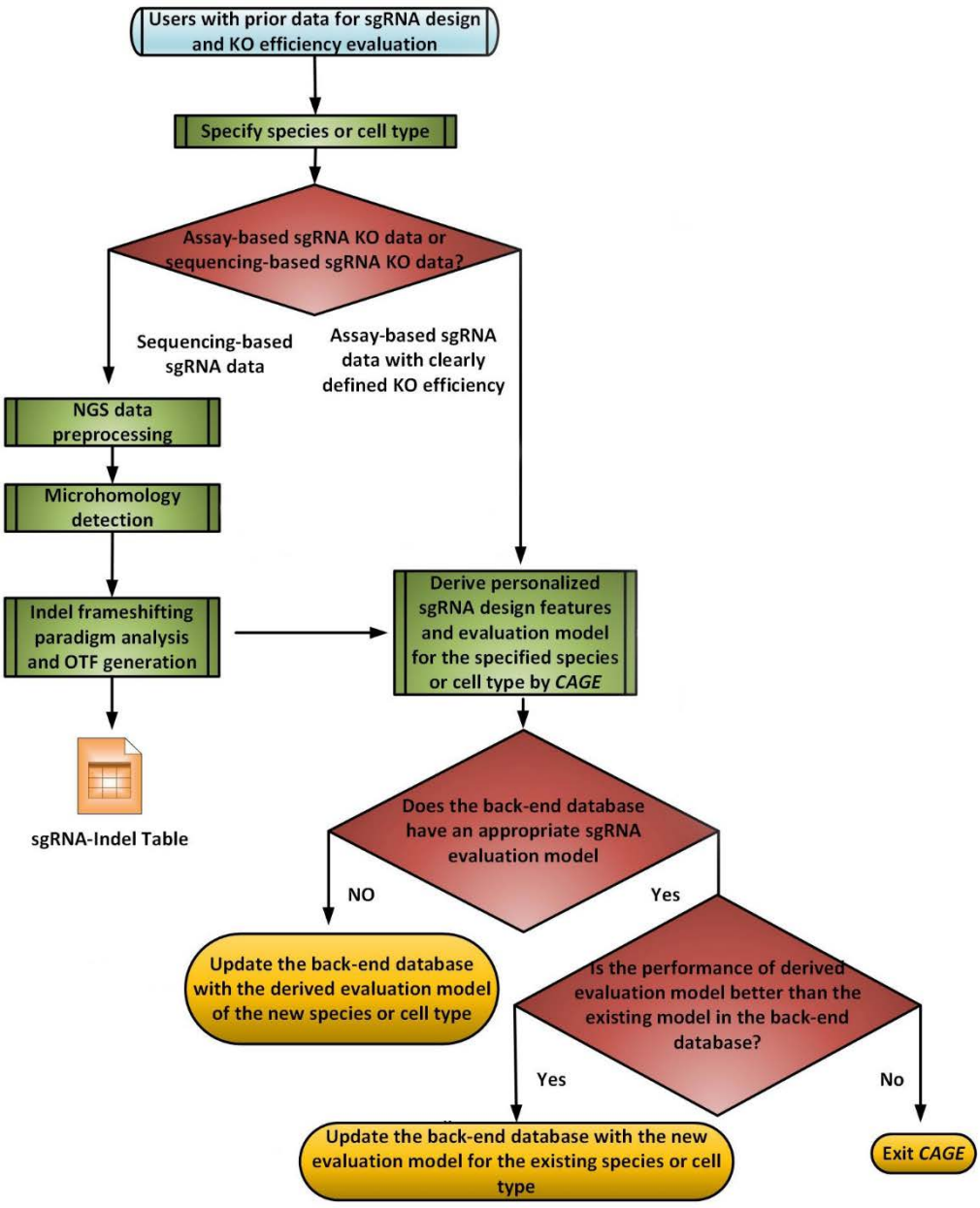

图 4-6 训练及分析流程示意图

这个工作流程同样要对 CRISPR-SpCas9 基因敲除实验中所使用的细胞系进行分支判断，同时还要根据输入数据是否为二代测序数据来选择下面的处理分支。如果输入数据为二代测序数据，该数据还要进入全基因组数据处理模块、微同源序列结构探测模块以及移码突变模式分析模块进行额外处理以获得向导 RNA 的打靶效率标签，输入的向导 RNA 需进入向导 RNA 处理模块进行预处理。当获得这些处理结果后，打靶效率预测系统会基于这些已处理好的数据调用向导 RNA 特征提取与打靶效率预测模块进行特征提取以及模型训练和测试，最终获得给定细胞系的 CRISPR-SpCas9 基因编辑系统打靶效率预测模型以及相应的特征集合，这些特征就是影响该细胞系 CRISPR-SpCas9 基因编辑系统打靶效率的重要因素，且可以使用特征可视化模块进行可视化展示。

4.4.1.2 向导 RNA 处理模块

向导 RNA 模块的输入为向导 RNA 列表，格式为 fastq 格式。向导 RNA 处理模块会使用 BWA (Burrows-Wheeler Aligner)[270] 或 Bowtie2[271] 等序列回帖程序将输入的向导 RNA 回帖到 CRISPR-SpCas9 基因编辑实验所使用的细胞系对应的基因组上，回帖结果根据其回帖质量进行过滤，即将回帖质量较低的结果去除。过滤后的回帖结果会被用来制作向导 RNA 信息表，记录每个向导 RNA 的相关信息，包括目标 DNA 区域所在染色体、坐标、DNA 链方向、目标 DNA 的序列以及切割点坐标。

4.4.1.3 全基因组数据处理模块

全基因组数据处理模块的主要用途是将 CRISPR-SpCas9 基因敲除实验所生成的二代测序数据进行处理并将与对应的向导 RNA 列表进行整合，从而获得每个向导 RNA 对应的目标 DNA 区域测序读段集合，该模块的输入是向导 RNA 处理模块所生成的向导 RNA 信息表以及 CRISPR-SpCas9 基因敲除实验所生成的二代测序数据原始 fastaq 文件。这些二代测序数据会被回帖到 CRISPR-SpCas9 基因敲除实验所用细胞系对应的基因组上，然后使用与向导 RNA 处理模块中相同的方法进行过滤操作，过滤后的结果会和向导 RNA 处理模块所生成的向导 RNA 信息表进行整合，这种整合的依据是测序数据回帖结果中 DNA 插入与删除位点坐标应与向导 RNA 信息表中相应的向导 RNA 的切割点坐标相对应。测序数据回帖结果中无法与向导 RNA 信息表中的向导 RNA 匹配的读段由此被过滤。因此，本模块的输出结果即为向导 RNA 与 CRISPR-SpCas9 基因敲除实验二代测序数据回帖读段的整合结果。

4.4.1.4 移码突变模式分析模块

将全基因组数据处理模块所生成的向导 RNA 与 CRISPR-SpCas9 基因敲除实验二代测序数据回帖读段的整合结果作为输入，移码突变模式分析模块会从中获取 DNA 插入或删除的一系列信息，这些信息包括移码突变的读段数量、整码突变的读段数量以及根据二者所计算出的移码突变比(out-of-frame ratio)，即移码突变的读段数量与该向导 RNA 所对应的所有读段数量之比。由于 CRISPR-SpCas9 基因编辑系

统的打靶效果依赖于因 DNA 双链断裂修复而产生的移码突变，因此移码突变比可以作为一种基于 CRISPR-SpCas9 基因敲除实验二代测序数据的打靶效率衡量手段[265]。移码突变模式分析模块的输出结果为向导 RNA 信息与其移码突变比等信息的整合信息表，其中移码突变比在之后的机器学习模型建立的过程中会作为对应向导 RNA 的打靶效率标签，而该整合信息表将作为向导 RNA 特征提取与打靶效率预测模块的输入被用来建立预测模型。

4.4.1.5 向导 RNA 特征提取与打靶效率预测模块

在向导 RNA 特征提取与打靶效率预测模块中，根据工作流程分支的不同分为了两大部分，其一为根据已有模型对待实验的向导 RNA 列表或基因进行预测，其二为基于已有 CRISPR-SpCas9 基因敲除实验数据进行特征提取并建立预测模型。

1. 预测部分

在这一部分中，本模块可以提供两种策略：一种是对使用者直接输入由向导 RNA 处理模块所获得的向导 RNA 信息表进行预测，另一种是基于使用者提供的基因组范围（如人类基因组 hg38，一号染色体，坐标 1 至 1000000）进行预测。在第二种策略中，本模块首先根据提供的基因组范围通过 PAM 序列模式（如 NGG）扫描获得向导 RNA 列表，并将其输入向导 RNA 处理模块中以获得向导 RNA 信息表，之后将按照与第一种策略相同的方法，选择对应于 CRISPR-SpCas9 基因编辑实验所用的细胞系预测模型对向导 RNA 进行打靶效率预测，从而达到优化设计向导 RNA 的目的。

2. 模型建立部分

向导 RNA 特征提取与打靶效率预测模块的模型建立部分包含三个子模块，分别为特征提取、特征选择以及模型建立。

输入的向导 RNA 序列信息通常使用独热法(one-hot)对向导 RNA 每个位置的碱基进行数字化编码，这种方法是离散型数据常见的编码方法，使用了四比特的二级制编码对每个碱基进行编码，编码表见表 4-3。

表 4-3 向导 RNA 碱基编码

A	C	G	T	N
1000	0100	0010	0001	0000

依据表 4-3，如果一个 DNA 序列为 GCTA，则其编码产物为 0010 0100 0001 1000。

第一步是特征提取，使用者首先须指定需要使用的向导 RNA 目标 DNA 上游以及下游长度。在此之后，本模块会从基因组中将每个向导 RNA 所对应的特征提取出来，包括目标 DNA 上游序列、20nt 长度的目标 DNA、3nt 长度的 PAM 序列以及后方的下游序列。这些序列之后会根据表 4-3 中的编码规则进行数字化编码，其编码后的产物即为向导 RNA 的输入特征集合。向导 RNA 的输入特征集合与其对应的 CRISPR-SpCas9 基因敲除实验打靶效率标签，如 CRISPR-SpCas9 基因敲除实验自定义的打靶效率、由移码突变模式分析模块计算得到的移码突变比以及打靶效率分类，一同构成了下面步骤的前置数据。这些数据会被分为训练集和测试集，其中训练集用于建立特征选择规则和训练预测模型，而测试集用于预测模型的性能测

试。

第二步是特征选择，这一步根据数据标签类型的不同分为两个处理分支。其一为数值型数据，即向导 RNA 的数据标签为 CRISPR-SpCas9 基因敲除实验自定义打靶效率或由移码突变模式分析模块计算得到的移码突变比。在这一处理分支中，本模块会使用 LASSO 方法进行特征选择，将权值不为 0 的特征挑选出来作为后面建立预测模型所使用的输入特征集合。对于另一个处理分支，其向导 RNA 的数据标签类型为分类型，本模块会采用基于 L1 正则化的逻辑斯蒂回归进行特征选择。与上一分支相同，本分支同样将特征选择过程中权值不为 0 的特征挑选出来作为下面建立预测模型所使用的输入特征集合。

最后一步是模型建立，由 LASSO 方法以及基于 L1 正则化的逻辑斯蒂方法生成的特征集合后会被应用于训练集以建立预测模型，这两类模型即为最终的预测模型。

4.4.1.6 特征可视化模块

前文提到了在向导 RNA 特征提取与打靶效率预测模块中可以生成被选择特征列表，而将该表输入特征可视化模块，即可获得选择出的特征集合的示意图，包括了长度设定以及所选择的基因组特征。图 4-7 为其序列特征图。

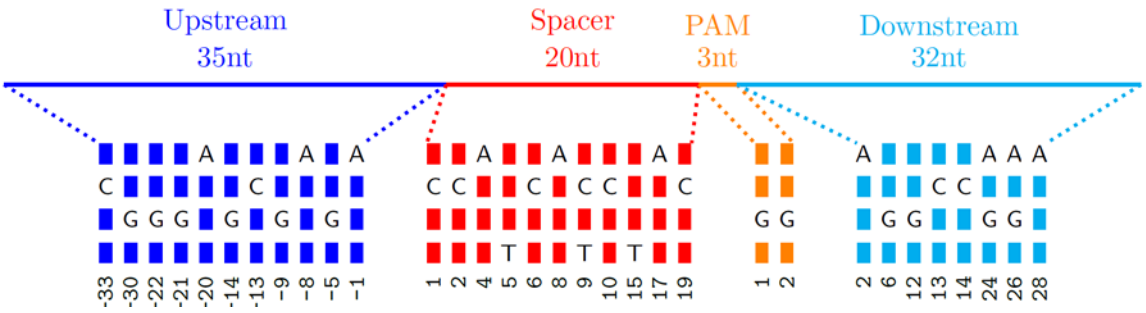

图 4-7 序列特征图举例

4.5 基于深度学习的 CRISPR 打靶效率预测

基于 CRISPR-Cas 基因编辑系统的各种基因编辑实验目前已经得到了广泛地应用，为功能基因组学研究提供了很大的帮助。因此，进一步优化 CRISPR-Cas 基因编辑系统，对推动功能基因组学和合成生物学等生物学领域的研究有着重要意义。

CRISPR-Cas 基因编辑系统在实际应用中主要面临两个问题。一是由向导 RNA 所引导的 CRISPR-Cas 基因编辑可能存在效率低下的问题。二是向导 RNA 引导的 CRISPR-Cas 基因编辑可能存在脱靶的问题。因此，能否有效地提高 CRISPR-Cas 基因编辑系统的效率和特异性已成为这种基因编辑系统能否得到更广泛应用的重要前提。

上一节介绍了基于目前应用最为广泛的 CRISPR-SpCas9 基因编辑系统所建立基于浅层学习的基因编辑打靶效率预测系统的基本结构，这种预测系统使用了浅层机器学习方法，能够学习特定细胞系的 CRISPR-SpCas9 基因敲除实验数据，获得针对该细胞系的打靶效率影响因子，这些影响因子会被用来建立一个定制化的预测模型，对来自同样细胞系的 CRISPR-SpCas9 基因编辑实验中向导 RNA 的打靶效率进行预测，达到优化设计向导 RNA 的目的。但浅层打靶效率预测系

统也存在以下缺陷：在数据利用层面，由于是一套数据或一组相同细胞系数据产生一个预测模型，因而每个预测模型的训练数据量过低，导致每个预测模型的预测性能受到很大的限制。在特征集合构造方面，只考虑了目标 DNA 区域序列层面的信息，如目标 DNA 区域序列、目标 DNA 区域上游序列以及目标 DNA 区域下游序列等信息，没有将表观遗传学信息考虑在内。在方法学层面，由于使用的是较为简单的广义线性模型，即 LASSO 以及基于 L1 正则化的逻辑斯蒂回归，其模型过于简单，导致数据的归纳抽象能力不足，进而使得预测能力降低。

本节将在浅层基因编辑打靶效率预测系统的研究基础之上，参考 Chuai 等人所开发的 DeepCRISPR 系统[249]，探讨基于深度学习的基因编辑打靶效率预测系统。深度打靶效率预测系统在浅层打靶效率预测系统的基础上，优化了系统的设计思路，成功解决了上述三个浅层打靶效率预测系统的缺陷。在数据利用层面，深度打靶效率预测系统将已有的 CRISPR-SpCas9 基因敲除实验数据进行了整合，使用整合数据进行模型开发，大大提升了模型的数据利用率。在特征集合构造方面，深度打靶效率预测系统能够成功将向导 RNA 对应的目标 DNA 区域序列信息与目标 DNA 区域的表观遗传学信息整合，并使用 DNA 序列信息与表观遗传学信息所构成的整合信息作为特征集合在此基础上进行模型训练。在方法学层面，深度打靶效率预测系统使用深度学习技术，获得了具有优秀预测性能的 CRISPR-SpCas9 打靶效率预测模型，打破了细胞系类型及基因组类型对于模型的限制，同时获得

了影响 CRISPR-SpCas9 基因编辑系统打靶效率的重要因素。

4.5.1 向导 RNA 编码模型

深度打靶效率预测系统使用了一种类似图片的编码手段对向导 RNA 进行了编码，其编码的信息包含了向导 RNA 对应的目标 DNA 序列信息、PAM 序列信息以及目标 DNA 区域的表观遗传学信息。通过使用表观遗传学信息，来自不同细胞系的 CRISPR-SpCas9 基因敲除数据便能够得到整合。在该编码模型中，目标 DNA 被抽象为一个单行的多通道图片。传统的彩色图片其每个像素包含 3 个通道值，即红通道、绿通道和蓝通道，而研究定义的“DNA 图片”，其序列信息包含四个通道，即 A 通道、C 通道、G 通道以及 T 通道，表观遗传学信息也包含多个通道，每个表观遗传学信息均有独立的通道表示。见图 4-8[249]。

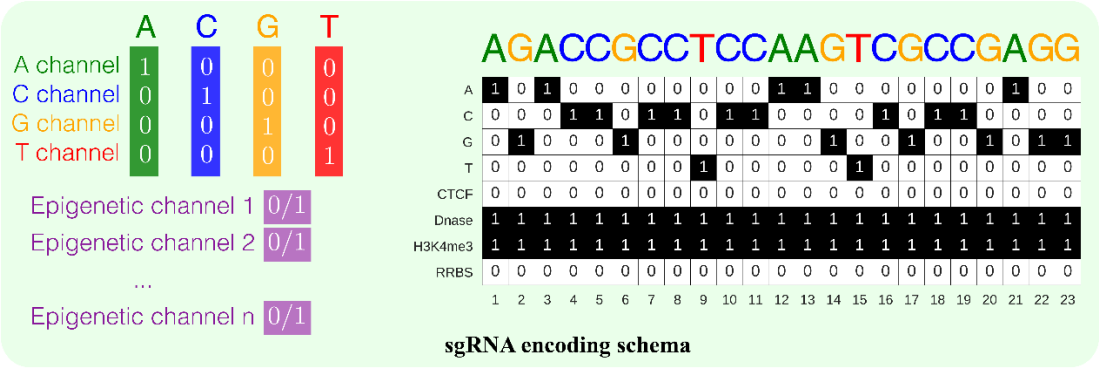

图 4-8　向导 RNA 编码方式

基于这样的向导 RNA 编码方式，考虑到 ENCODE 的表观遗传学数据正在不断增长，深度打靶效率预测系统可以轻松地被扩展到其他细胞系或物种。

4.5.2 深度打靶效率预测系统

4.5.2.1 深度打靶效率预测系统整体概况

目前，使用机器学习方法通过 CRISPR-SpCas9 基因敲除实验数据建立的打靶效率预测模型有以下三个问题。第一，数据异质性问题，即由于 CRISPR-SpCas9 基因敲除实验数据来源于不同的细胞系以及不同的实验平台，因此这些数据需要特别的方式进行整合。第二，数据稀疏性问题，即与基因组潜在的向导 RNA 相比，当前已有的 CRISPR-SpCas9 基因敲除实验数据中向导 RNA 样本数量过少，这导致机器学习模型的建立出现困难。第三，表观遗传学因素不明确问题，即表观遗传学因素对于 CRISPR-SpCas9 基因编辑系统打靶效率的影响目前还未明确[236]。考虑到这些问题，深度打靶效率预测系统采用自动编码器技术，利用全基因组范围内潜在的所有无标签向导 RNA 的相关信息作为训练数据训练得到了一个自动编码器模型以获得向导 RNA 相关信息的抽象表示。这个训练完毕的自动编码器可以使用迁移学习的方法应用于具有 CRISPR-SpCas9 基因敲除实验打靶效率作为标签的向导 RNA 样本的打靶效率模型的训练中。从理论和实践结果可知，应用大量的无标签向导 RNA 信息所训练的自动编码器对 CRISPR-SpCas9 基因编辑系统打靶效率预测模型进行的迁移学习能够有效提高后者的预测性能。

深度打靶效率预测系统具有如下五个特点。第一，考虑到不同细胞系的表观遗传学信息，将来自不同细胞系的 CRISPR-SpCas9 基因敲除实验数据中向导 RNA 对应的目标 DNA 区域整合到了一个统一

的特征空间之中，整合了来自不同细胞系的 CRISPR-SpCas9 基因敲除实验数据。第二，使用了亿级别数量的全基因组潜在的无标签向导 RNA 相关信息训练了一个自动编码器并将其部分作为父网络，由此可以产生向导 RNA 相关信息的高级抽象表示，并将其应用于 CRISPR-SpCas9 基因编辑系统打靶效率预测模型的训练中。第三，使用了特定的数据扩增技术，生成了具有生物学意义标签的衍生向导 RNA 样本，大大增加了 CRISPR-SpCas9 基因编辑系统打靶效率预测模型的训练数据量，从而获得预测性能更强的 CRISPR-SpCas9 基因编辑系统打靶效率预测模型。第四，使用了有 CRISPR-SpCas9 基因敲除实验数据标签的向导 RNA 样本数据，基于自动编码器的编码器部分通过迁移学习建立了 CRISPR-SpCas9 基因编辑系统打靶效率预测模型，从而使用少量有标签的向导 RNA 样本获得了高性能的打靶效率预测模型。第五，能够自动选择出对 CRISPR-SpCas9 基因编辑系统打靶效率产生影响的重要特征，包括 DNA 序列特征以及表观遗传学特征，进而对 CRISPR-SpCas9 基因编辑系统中向导 RNA 的优化设计及 CRISPR-SpCas9 基因编辑系统相关机制的研究提供帮助。

4.5.2.2 全基因组潜在向导 RNA 自动编码器模型

深度打靶效率预测系统可以通过基于卷积层的去噪自动编码器对向导 RNA 相关信息（包括目标 DNA 区域序列、PAM 序列以及相关的表观遗传学信息）表示学习，该自动编码器包含一个编码器以及一个解码器。自动编码器的输入为含有表观遗传学信息的全基因组向导 RNA 样本信息。这些训练样本是通过将人类基因组中符合

CRISPR-SpCas9 基因编辑系统 PAM 序列模式 NGG 的所有序列提取出来，然后根据每个潜在向导 RNA 序列的基因组坐标获得对应的表观遗传学信息而获得的。通过这些样本，可以得到基于卷积层的去噪自动编码器，从而获得向导 RNA 样本的抽象表示。

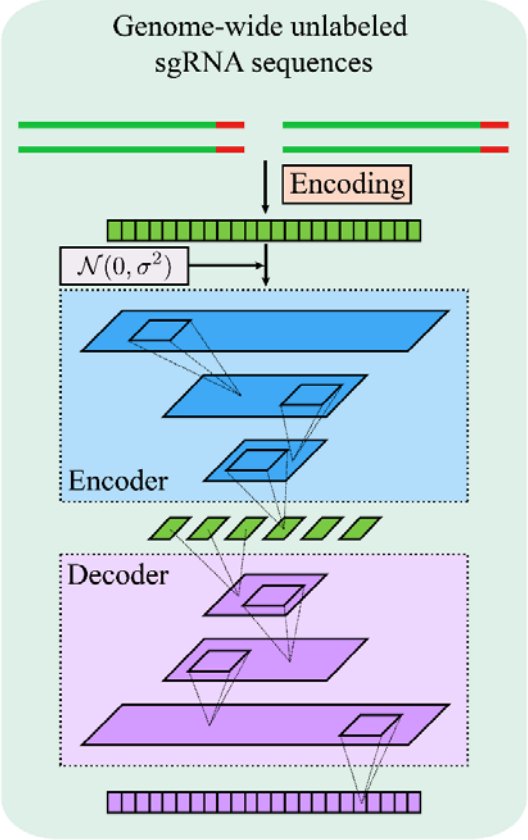

图 4-9 自动编码器结构示意图

自动编码器的模型结构如图 4-9[249]所示，使用了去噪的方法在输入层上加入了基于正态分布的噪声数据。相对于一般自动编码器，这种去噪自动编码器能够稳定地处理巨大的样本量所产生的大量噪音。其用途在于通过其中的编码器部分获得向导 RNA 样本的抽象表示，这种训练得到的特征表示也将应用于之后的打靶效率预测模型的训练中。

4.5.2.3 打靶效率预测模型

基于卷积层的去噪自动编码器训练完成后，可以使用采用全卷积结构的神经网络以进行 CRISPR-SpCas9 基因编辑系统打靶效率预测。该网络的结构与训练过程如图 4-10[249]所示，打靶效率预测模型的训练采用迁移学习方法，将已训练好的去噪自动编码器中的编码器权值迁移到模型的前端特征提取层，且不将迁移的神经网络层权值固定，使用具有 CRISPR-SpCas9 基因敲除实验打靶效率标签的向导 RNA 样本对整个模型的权值进行训练调整。这其中由大量无标签向导 RNA 样本学习得到的向导 RNA 信息流型可以提升打靶效率预测模型的预测性能。

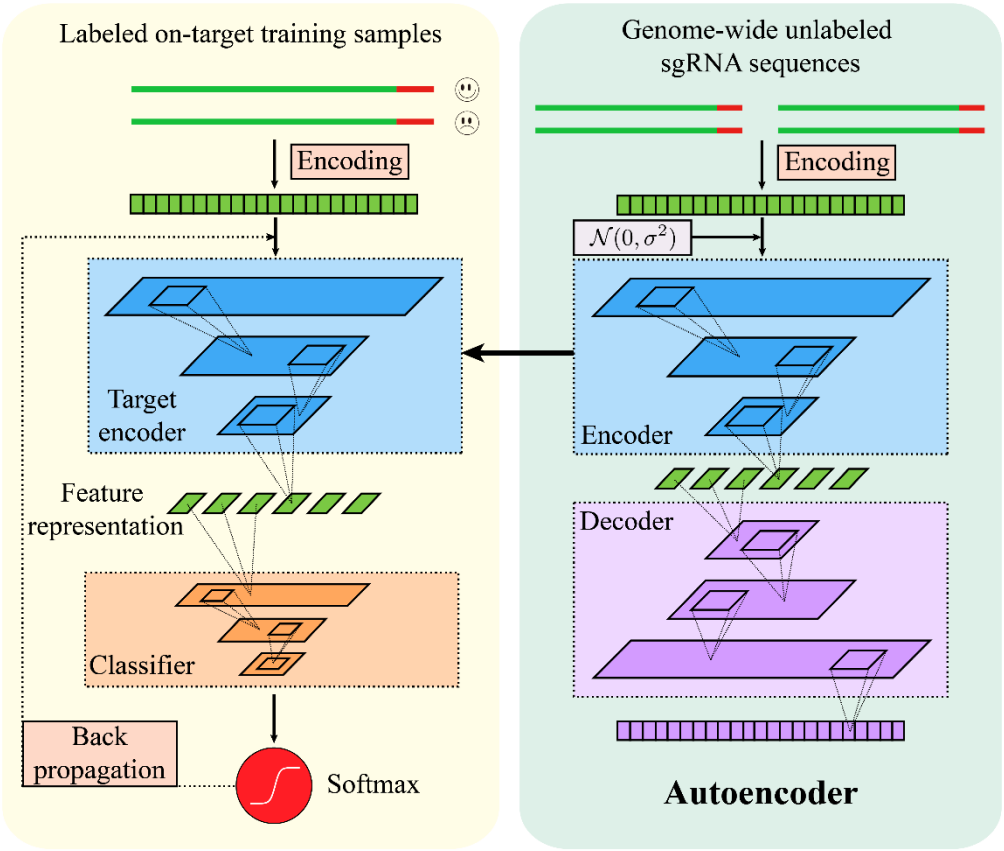

图 4-10 打靶效率预测模型训练过程

4.5.2.4 打靶效率预测模型特征提取及特征意义

已经训练完毕的打靶效率分类预测模型进行影响CRISPR-SpCas9基因编辑系统打靶效率的相关特征提取，其方法为显著图法[272]，这种方法利用已经训练完毕的分类预测模型以及一个指定的类别，运用数值方法计算出一个假想向导RNA目标DNA区域信息，此向导RNA目标DNA区域信息代表了所指定的类别在模型中的模板表示。其数学表达式为：

$$x_g = argmax_g S_c(g)$$

其中c为指定的类，g为一个向导RNA目标DNA区域的信息，Sc(g)为打靶效率分类预测模型基于向导RNA信息g得到的第c类的值。

显著图法的目的就是为了找到一个g使得Sc(g)取得极大值。通常情况下，将模型的输入作为变量，使用梯度上升法迭代，即可得到这样的假想向导RNA目标DNA区域信息。将c设为1，则计算得到的假想向导RNA目标DNA的区域信息即为高打靶效率向导RNA目标DNA区域信息的模板，其特征选择即为影响打靶效率的重要因素。

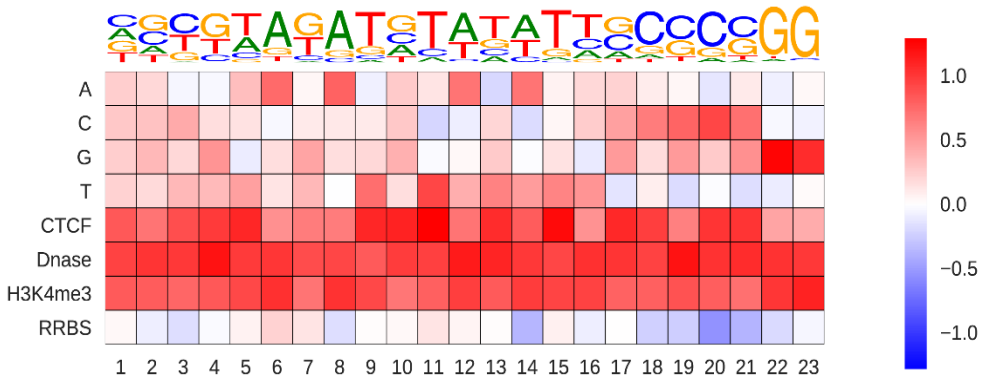

图 4-11 打靶效率预测模型特征显著图

深度打靶效率预测模型所获得的显著图如图 4-11[249]所示，从中可以归纳出以下结论。第一，在高打靶效率的向导 RNA 目标 DNA 区域信息中，PAM 序列模式 NGG 中 N 的选择通常为胞嘧啶 C 或鸟嘌呤 G，这个结果已经被若干体内和体外 CRISPR-SpCas9 基因敲除实验所证实[235,238]。第二，胸腺嘧啶 T 出现在距离 PAM 序列较近的位置（距离小于等于 4）会降低打靶效率，该结果也与实验结果相吻合。CRISPR-SpCas9 基因敲除实验发现，当向导 RNA 的近 PAM 端含有多个尿嘧啶 U 时，其打靶效率会显著降低[238]。第三，高打靶效率向导 RNA 目标 DNA 区域中的第 18 位对胞嘧啶 C 具有明显的倾向性，该结论同样得到了实验证实[236, 237]。第四，染色质的开放对于打靶效率的提升有着重要的影响。第五，DNA 甲基化不利于 CRISPR-SpCas9 基因编辑系统对目标区域 DNA 的切割，这一点也得到了实验的证实[273]。

4.6 基于深度学习的 CRISPR 脱靶分布预测

当前功能基因组学的研究中，CRISPR-Cas 基因编辑系统已经成为一种非常重要的技术手段，利用 CRISPR-Cas 基因编辑系统可以对目标基因进行敲除、编辑或调控相关基因上下游功能性元件，从而探究相关功能基因组学问题。

如前所述，CRISPR-Cas 基因编辑系统在基因编辑实验中主要面临两个问题。第一个问题是向导 RNA 所引导的 CRISPR-Cas 基因编辑可能存在效率低下的问题，第二个问题是向导 RNA 所引导的 CRISPR-Cas 基因编辑可能存在脱靶的问题。脱靶效应已在多个使用

CRISPR-SpCas9 基因编辑系统进行相关研究的实验中被发现[235, 254, 274]，并对基因编辑实验的结果产生了极大干扰，因此如何准确定量地对脱靶位点进行搜索和预测是一个重要的问题[275]。

在 CRISPR-Cas 基因编辑系统的实际应用中，CRISPR-Cas 基因编辑系统的脱靶问题更是限制了该系统有效并安全应用的重大问题。此外，由于每次 CRISPR-Cas 基因编辑实验中实验组细胞在脱靶的数量和位置方面都具有很大的不确定性，CRISPR-Cas 基因编辑实验产生的基因编辑细胞在基因组层面会呈现出异质性，也会被脱靶编辑基因所产生的下游噪音影响，阻碍目标基因的研究。就基因敲除实验而言，基因敲除的脱靶会导致非目标基因被破坏，而被敲除的非目标基因很可能事先无法确定，最终会导致细胞产生错误的表型和系统性错误的产生，使得实验失败。因此针对 CRISPR-Cas 基因编辑系统出现的脱靶问题，设计特异性更强的向导 RNA 可以有效的提高 CRISPR-Cas 基因编辑实验的成功率及安全性，这将大大提高 CRISPR-Cas 基因编辑系统的灵敏度和准确性，对功能基因组学等领域的研究有着重要意义。

目前许多已有的工具采用了简单的 DNA 序列错配比对，根据设定的错配碱基数量以及 PAM 序列模式，使用相关的 DNA 序列比对软件列举出基因组中所有与目标 DNA 相似的 DNA 位点作为脱靶位点，但也有少数工具能够对与目标 DNA 相似的潜在脱靶位点进行脱靶发生预测，这类软件以 CFD[235]和 MIT[274]为代表，根据实验数据总结或猜想得到的经验性规则对潜在脱靶位点的脱靶发生进行预测。而

相关的脱靶实验数据通常来源于全基因组 DNA 双链断裂检测等实验方法，如 GUIDE-seq[259]、Digenome-seq[258, 276, 277]、HTGTS[261]，BLESS[278]和 IDLV[260]等，但这种基于人工规则的脱靶发生预测工具有很大的局限性，如启发式人工规则的制定存在较大的限制、CRISPR-SpCas9 基因编辑实验数据的利用率低等。

实践证明浅层学习难以处理脱靶发生的预测问题，因此下文将重点描述基于深度学习的全基因组脱靶分布预测系统。这里主要参考了 Chuai 等人所开发的 DeepCRISPR 系统[249]，通过深度学习方法，对给定向导 RNA 目标 DNA 序列相似的潜在脱靶位点进行脱靶发生预测，并使用深度学习中的特征选择技术，得到与 CRISPR-SpCas9 基因编辑实验脱靶发生相关的重要特征。

4.6.1 数据编码

就输入数据的编码而言，本节使用与上一节中 CRISPR-SpCas9 基因编辑系统打靶预测研究中相同的编码方式，即将向导 RNA 对应的目标 DNA 区域抽象为一个单行的多通道图片。其序列信息包含四个通道，即 A 通道、C 通道、G 通道以及 T 通道，表观遗传学信息包含相应的独立通道。但与 CRISPR-SpCas9 基因编辑系统打靶预测研究不同的是，脱靶问题中除了需要对目标 DNA 区域进行编码外，还需要对基因组中与目标 DNA 序列相似潜在的脱靶位点进行编码，而编码方式与上述方法相同，即将潜在的脱靶位点区域看作“DNA 图片”，将这个潜在的脱靶位点的序列信息和表观遗传学信息组合为多通道的图片，这种“DNA 图片”的具体表示在上一节有所介绍，详情

见图 4-12。

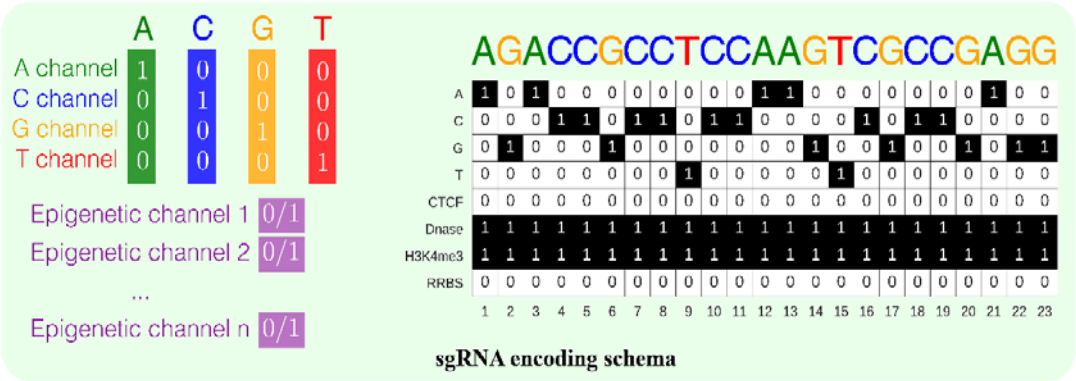

图 4-12 向导 RNA 编码方式

4.6.2 深度全基因组脱靶分布预测系统

4.6.2.1 深度全基因组脱靶分布预测系统整体概况

当前，使用基于数据的机器学习等方法建立的通过 CRISPR-SpCas9 基因敲除实验脱靶位点数据建立 CRISPR-SpCas9 基因编辑系统向导 RNA 脱靶发生预测模型有以下四个问题，其中的三个问题与 CRISPR-SpCas9 基因编辑系统打靶效率预测中出现的问题相同，包括数据异质性问题，即由于 CRISPR-SpCas9 基因敲除实验脱靶位点数据来源于不同的细胞系和不同的实验平台，导致这些数据需要特别设计的整合方式进行合并；数据稀疏性问题，即与基因组潜在的向导 RNA 数量相比，当前已有的 CRISPR-SpCas9 基因敲除实验脱靶位点数据中的向导 RNA 数量极其稀疏，如本研究中向导 RNA 数量为 30，这导致机器学习模型的建立出现极大的困难；表观遗传学作用不明确问题，即目标 DNA 区域或潜在脱靶位点区域的表观遗传学信息对脱靶发生的影响目前还不明确。而有一个问题是 CRISPR-SpCas9 基因编辑系统脱靶发生预测所特有的，即数据不平衡问题，该问题的表现形式为各种脱靶检测实验所证实的真实脱靶位点

的数量远远少于通过 DNA 错配比对所得到的全基因组范围的潜在脱靶位点数量。为此，全基因组脱靶分布预测系统同样使用前述的自动编码器技术。该自动编码器使用了全基因组范围内所有的潜在无标签向导 RNA 目标 DNA 区域信息进行训练，并能够自动学习向导 RNA 相关的目标 DNA 区域信息的抽象表示。由于全基因组范围的潜在脱靶位点实际上也是潜在的向导 RNA 目标 DNA 区域，因此这个训练完毕的自动编码器不仅适用于打靶效率预测系统，同样也能应用于全基因组脱靶分布预测系统，之后可以使用迁移学习的方法将这个去噪自动编码器的编码器部分应用于使用具有 CRISPR-SpCas9 基因敲除实验脱靶位点标签的数据集进行训练的脱靶发生预测模型的训练中。

深度全基因组脱靶分布预测系统具有如下几个特点：第一，运用细胞系的表观遗传学信息将来自不同细胞系及不同实验平台的 CRISPR-SpCas9 基因敲除实验脱靶位点的数据进行了有效整合。第二，使用了由上亿数量的全基因组潜在的无标签向导 RNA 目标的 DNA 区域信息训练的基于卷积层的去噪自动编码器，用于生成向导 RNA 目标 DNA 区域以及相应的潜在脱靶位点区域的抽象特征，并将其应用于 CRISPR-SpCas9 基因编辑系统脱靶发生预测模型的训练中。第三，使用了具有 CRISPR-SpCas9 基因敲除实验脱靶位点标签的数据集，基于上述自动编码器使用迁移学习建立了 CRISPR-SpCas9 基因编辑系统脱靶发生预测模型，利用了少量有标签的样本获得了较高性能的脱靶发生预测模型。第四，在训练的过程中使用了一种有效的平衡抽样方法，极大地缓解了由于数据不平衡所造成的脱靶发生预测性能较差的问题。第五，能够自动选择出对 CRISPR-SpCas9 基因编辑系统脱靶发生产生影响的重要特征，这些特征包括 DNA 序列特征

以及表观遗传学特征，通过分析这些特征，为 CRISPR-SpCas9 基因编辑系统向导 RNA 的优化设计及脱靶机制研究提供帮助。

在实际预测中，全基因组脱靶分布预测系统首先使用 Bowtie2[271] 和 Cas-OFFinder[244]等基因组比对工具获得输入的待预测向导 RNA 在基因组中的目标 DNA 位点以及全部的潜在脱靶位点，通过 ENCODE 等表观遗传学数据库，获得这些位点的表观遗传学数据，再将这些数据输入脱靶发生预测模型，最终得到给定向导 RNA 的全基因组脱靶分布情况。对于全基因组脱靶分布情况的描述如图 4-13[249]所示。图中的外圈为按坐标排列的人类 24 条染色体，每一个黑点为一个潜在的脱靶位点。若黑点落在红色区域，则该位点为由脱靶发生预测模型计算认为具有高脱靶风险的位点，同样地，黄色区域为中等脱靶风险区域，而绿色区域为低脱靶风险区域。

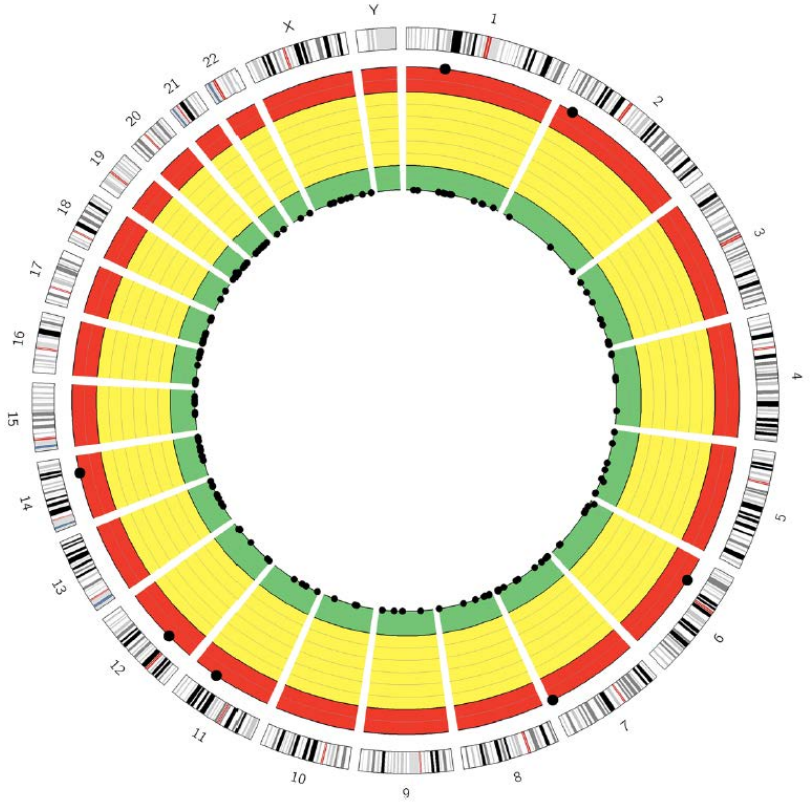

图 4-13 全基因组脱靶分布示例

4.6.2.2 基于拔靴法的样本平衡抽样算法

CRISPR-SpCas9 基因编辑系统脱靶发生预测所使用的训练样本存在着数据不平衡的问题。在本研究所使用的 CRISPR-SpCas9 基因敲除实验脱靶位点数据，脱靶检测实验验证的真实脱靶位点的数量与其他理论上为潜在脱靶位点却未发生实际脱靶情况的位点的数量之比通常高于 1:100。由于神经网络的主要训练方式为随机梯度下降法(Stochastic Gradient Descent, SGD)，因此当输入的训练样本长时间保持为单一类型样本时，梯度的更新会出现极大的偏差，即模型更新的方向会指向长时间出现的类型，而偶尔出现的其他类型样本将很难改变模型更新的方向。这种情况的持续会导致模型失去了对其他较少出现类型的预测能力，预测结果将保持为训练中长时间出现的类型。就 CRISPR-SpCas9 基因编辑系统脱靶发生预测而言，如果忽视训练数据不平衡问题，最终训练的模型会将输入的所有潜在的脱靶位点判断为不会发生脱靶。而在脱靶发生的预测中，很明显脱靶位点的重要性大大高于未发生脱靶的潜在脱靶位点，因此解决这个问题的关键在于，训练出的模型需要具有鉴别出现的脱靶位点的能力，而不能将所有潜在脱靶位点都判断为不会发生脱靶。

为了解决这个问题，一种可行的方法是使用基于拔靴法(bootstrapping)的样本平衡抽样算法来构建模型训练时输入的迷你批次数据。这种算法的基本思路是在构建迷你批次数据时首先使用无放回抽样法从样本数量较多的类型中抽取样本，再使用有放回的抽样方法从样本数量较少的类型中抽取样本，保证每次进入模型的迷你批次

数据中的每种类型样本的数量之比为 1:1，即实现了平衡抽样。通过这种抽样方式，模型训练时输入的迷你批次(mini-batch)数据具有相同数量的正样本与负样本，由此可以有效避免梯度更新时对具有大数量样本类型的偏差，提升了模型对于小样本量类型的敏感度。

4.6.2.3 脱靶发生预测模型

基于卷积层的去噪自动编码器训练完成后，采用全卷积结构建立用于进行脱靶发生预测的神经网络模型。与深度打靶效率预测模型不同的是，脱靶发生预测模型有两个与去噪自动编码器编码器部分相同结构的编码器，一个融合层以及之后的卷积处理和分类层，这个脱靶发生预测模型有两个输入，这两个部分被分别输入到两个编码器之中，两个编码器的输出结果在融合层进行合并，合并的方向为通道方向。合并后的结果之后会进入后端的卷积处理和分类层，最终输出结果。

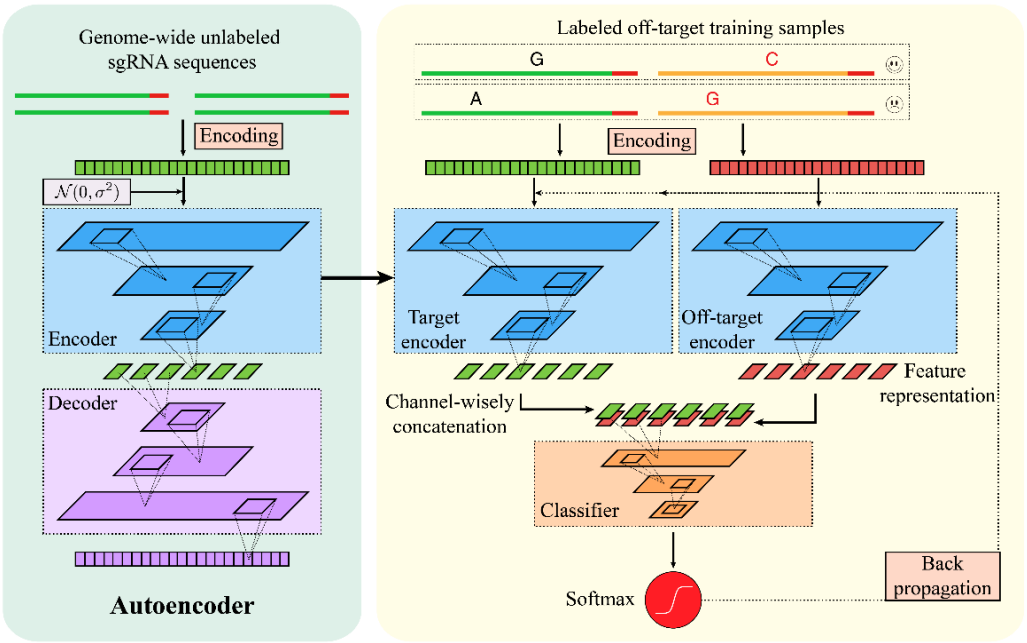

图 4-14 脱靶发生预测模型训练过程

该网络的结构与训练过程如图 4-14[249]所示，将处理好的给定向导 RNA 目标 DNA 区域信息与其对应的一个潜在的脱靶位点的区域信息作为一个样本对，即脱靶发生预测模型的训练样本。每个样本对包含两个部分，其一为向导 RNA 对应的目标 DNA 的区域信息，包括 DNA 序列信息以及表观遗传学信息；其二为与目标 DNA 序列相似的潜在脱靶位点的区域信息，同样包括该区域的 DNA 序列信息以及相应的表观遗传学信息。这样的两部分编码可以准确地将向导 RNA 作用的目标 DNA 区域信息和与目标 DNA 序列相似的潜在脱靶位点信息合并为一个整体表示。训练的过程中，训练样本的每部分都会进入一个与去噪自动编码器编码器部分相同结构的编码器以获得每部分的抽象表示，而这其中提到的两个编码器在数据进入前均已经通过迁移学习得到了去噪自动编码器中编码器部分的权值。在此之后，从编码器中得到的上述两部分数据的抽象表示会在融合层中沿通道方向进行合并，合并的结果中既包含了目标 DNA 区域的抽象表示，又包含了潜在脱靶位点区域的抽象表示。将该结果输入到下面的卷积结构的分类器中，即可得到最终的预测结果。

脱靶发生预测模型的训练使用了从多个 CRISPR-SpCas9 基因敲除实验脱靶位点数据来源的样本。训练的过程中为了克服模型训练时的样本不平衡问题，需要采用了上一小节介绍的基于拔靴法的平衡抽样算法以对模型训练使用的迷你批次数据进行平衡化构造，缓解样本不平衡问题。此外，与深度打靶效率预测模型相同，去噪自动编码器的编码器部分迁移到脱靶发生预测模型的两个编码器中的权值会随

着模型的训练而进行调整，最终实现对 CRISPR-SpCas9 基因编辑系统特异性的优化。

4.6.2.4 脱靶发生预测模型特征提取及特征意义

对已经训练完毕的脱靶发生分类预测模型进行脱靶位点的相关特征提取，采取的特征提取方法与深度打靶效率预测模型所使用的特征提取方法相同为显著图法，使用这种方法可以构建包含全部目标 DNA 区域 20 个位置的 16 种碱基错配的显著图。由于向导 RNA 样本量通常较少，所以需要使用费舍尔精确性检验对显著图中的每个点进行了统计显著性检验并去除缺乏统计显著性的数据点，得到的显著图如图 4-15[249]所示。从图中可以得出以下结论，且一部分结论已得到实验证实，如发生在第 16 位的鸟嘌呤 G 到胞嘧啶 C 的错配以及鸟嘌呤 G 到胸腺嘧啶 T 的错配会显著降低潜在脱靶位点发生脱靶的风险[235, 279]，而其他位置的碱基错配仍需要实验验证。

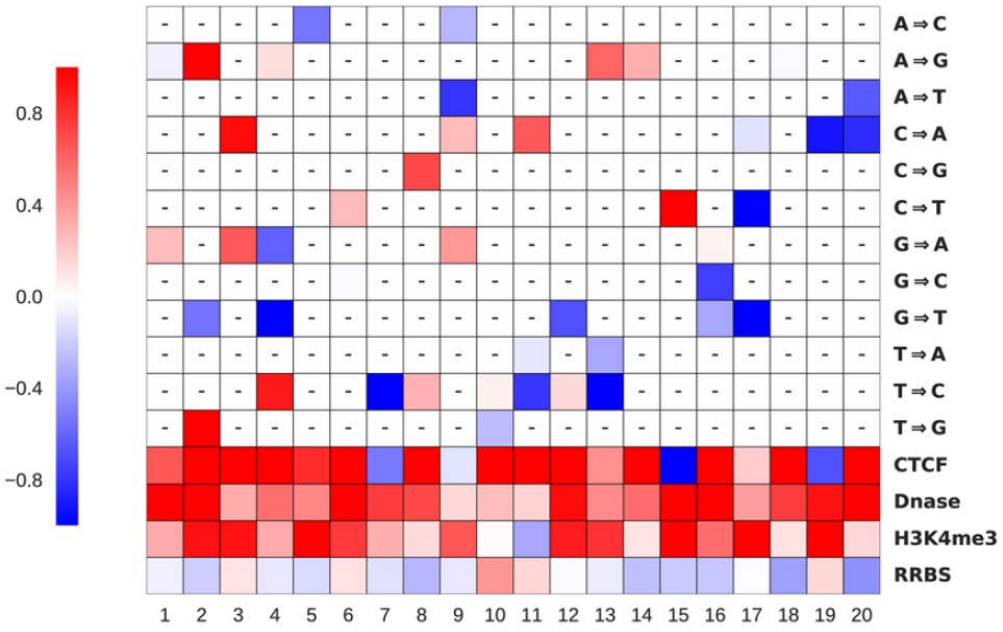

图 4-15　脱靶发生预测模型特征显著图

4.7 人工智能在基因编辑中的发展前景

随着数据科学的不断发展，机器学习已成为目前计算机科学的一个热点。将机器学习理论和向导 RNA 设计优化相结合，通过收集基于 CRISPR-Cas9 的大规模全基因组级功能性基因筛选实验数据建立机器学习模型，利用这些数据对机器学习模型进行训练，进而得到最终的模型。通过这些模型能对候选的向导 RNA 进行效率或特异性的预测，最终获得优化的向导 RNA。这种数据驱动方法的创新性是能够直接从已有的数据中通过算法自动归纳影响结果的向导 RNA 设计的重要因素，而且随着发布的实验数据的增加，模型预测的准确度能够不断提高。这些从数据归纳的 CRISPR-Cas9 基因编辑系统的影响因素将在理论上为该系统的进一步优化提供参考，预测模型将为实验中向导 RNA 的设计提供合理优化方案，并使得一些需要很高特异性的基因编辑实验的实现成为可能。近期已有研究者对这一方法做出了尝试，但根据这些模型已发表的评估结果显示，这些模型的预测性能与随机猜测相比并无显著优势，其主要原因在于无法完全利用基于多种细胞系产生的实验数据而导致数据量无法有效增多，以及机器学习算法的使用存在缺陷。因此，更好的利用基于多个细胞系产生的实验数据并选择合适的特定的机器学习算法对实现向导 RNA 设计优化有着决定性的作用，这种优化的 CRISPR-Cas9 基因编辑系统对于进一步推动功能基因组学和合成生物学等生物学领域的研究有着重要意义。

人工智能在基因编辑中的发展重点与建议如下：

(1) 目前研究中所使用的深度学习结构框架相对简单，而在今后的研究中，可以尝试使用更加复杂的模型结构，如神经图灵机等。

(2) 人工定义的用于预测 CRISPR-SpCas9 基因编辑系统效率及特异性的特征可以与向导 RNA 的原始特征可以一同整合到深度学习模型，从而形成一种层次更加丰富的表示学习，进一步提高模型的预测能力。

(3) 高通量 CRISPR-SpCas9 基因敲除实验数据量有待提升。

(4) 目前已有的 CRISPR-SpCas9 基因敲除实验数据中有一定量噪音的存在，对模型的构建会产生一定的影响。可通过特别设计的 CRISPR-SpCas9 基因编辑系统实验获得标记更加准确的训练数据，提高模型的准确性。

(5) 目前所建立的模型依赖于表观遗传学信息，而目前已有的表观遗传学信息数据库的物种覆盖范围有限，这种限制将随着表观遗传学信息数据库内容的不断丰富而慢慢消除。

(6) 对于深度学习模型构建的过程，本章中提到的主要是进行人工直接构建，即人工定义神经网络每一层的各个参数。未来可以尝试基于强化学习的自动化深度学习模型构建方法，不仅能够大大提升模型构建和训练的效率，也能够打破人工构建所带来的预测性能瓶颈。

第五章 人工智能与疾病智能诊断

5.1 人工智能与疾病智能诊断概述

1956 年，约翰麦卡锡在达特茅斯会议上提出了人工智能一词，并指出人工智能是“表现出看似聪明行为的硬件或软件”[280]。人工智能自诞生以来已在众多领域（如机器人、图像识别、自然语言理解、石油化工、医疗诊断、专家系统、军事等）获得了广泛应用。二十世纪四十年代，研究人员尝试研发能作为医学顾问的计算机程序，用以辅助临床医师的诊疗工作，这是人工智能在医学领域的最早报道[281]。随着计算机科技的发展，人工智能展现了强大的数据处理能力，并且适用于医学图像的识别和复杂临床数据的分析。

在早期的计算机辅助程序中，主要使用流程图、贝叶斯算法、模式匹配、布尔代数和统计学决策分析等方法来进行工作。这些单纯的数学程序与医学专业知识相脱离，因此只能解决一些非常简单的问题且这种程序的实用价值并不高。

专家系统(Expert System, ES)是某个领域内具有专家水平的智能推理系统，是人工智能技术在医疗诊断领域中的最富有代表性和最重要的应用，分别由知识库、知识获取机构、综合数据库、推理机、人-机接口、解释器五部分组成，其组成结构图如图 5-1。而医学诊断专家系统(medical expert system, MES)是运用专家系统的设计原理与方法，吸收了大量的知识和经验，模拟医学专家诊断的思维活动及推理

判断，得出与人类专家一样的结论的方法。它可以帮助医生解决复杂的医学问题，也可以作为医生诊断的辅助工具。在二十世纪七十年代，研究人员通过模拟人类专家解决医学问题的方法，研发了多种类型的专家系统，例如 PII、MYCIN、CASNET 和 INTERNIST，这些专家系统在特定的领域中已经被证实可达到人类专家的水平[282]，但在临床实践中的作用不大。

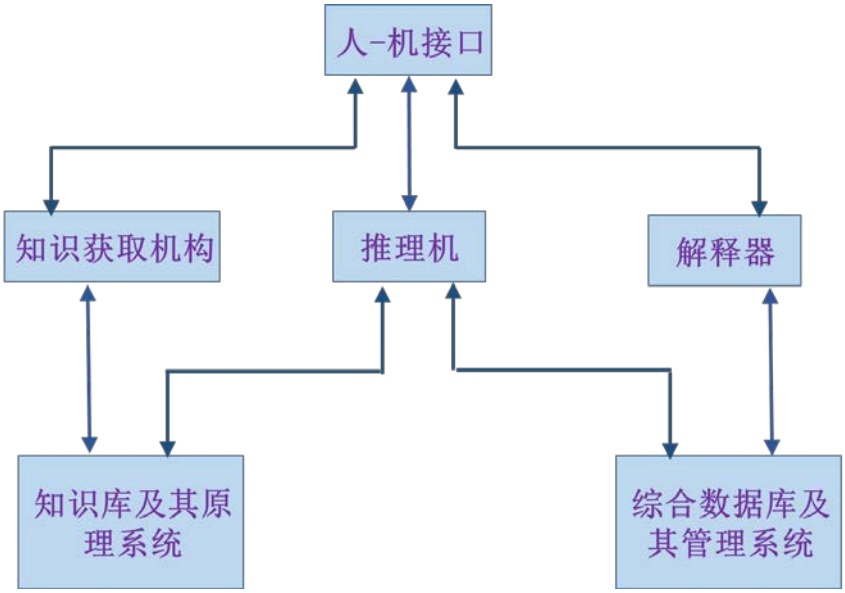

图 5-1 专家系统的组成结构

人工神经网络(Artificial Neural Networks)的结构在很大程度上模仿了大脑，功能上模拟神经元的感知器组成。神经网络和模糊系统都模仿大脑，但仍可以明显的区分这两个系统。神经网络操纵明确定义受限数据，而模糊系统可以处理和决定未定义且不确定边界的数据集，并且决策依赖于不同类别对象的差异程度，因为它们不是由粗略边界定义，这与布尔真和假二元决定性逻辑不同。与传统计算机回归分析的单层结构不同，人工神经网络是一个复杂的多层感知模型，其基本结构包括输入层、模拟神经元层和输出层三个部分。人工神经网

络建立了一个决策单元，通过感知器单元的相互连接可以实现非线性分析（图 5-2）。人工神经网络有两种类型，一种最常用的 ANN 格式是多层感知器(MLP)，这代表一个前馈网络，其中一层输入感知器连接到许多隐藏的感知层，然后是输出层。在 1988 年，Szolovits 等人[283]提出将深度神经网络运用于疾病的诊治。随后，随着人工智能的发展，MLP 模型用于多项临床研究，来预测卒中患者的死亡风险和缺血性卒中出血的风险[284, 285]和心血管疾病的诊断[286, 287]。MLP 已被证明是预测某些已建立的癌症生物标志物潜力的有效载体[288]。

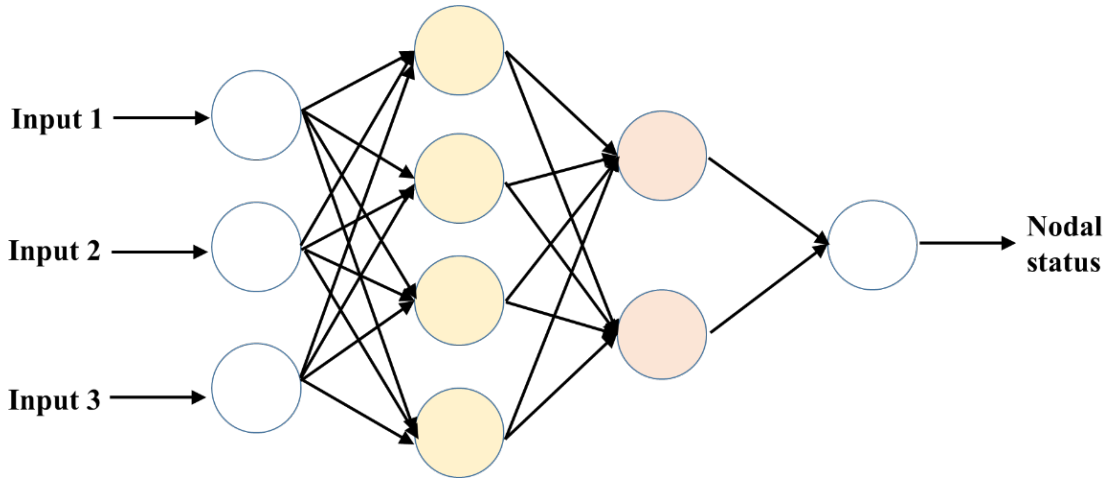

图 5-2 人工神经网络模型

5.2 智能诊治的应用实例

5.2.1 智能诊治在消化系统疾病中的应用

下文主要描述了三种人工智能技术在消化系统疾病诊治中的应用。

第一、肝脏疾病的诊治。2014 年，Streba 等人[289]使用深度神经网络系统结合人口统计学和临床数据以及成像数据对 486 例肝局灶性病变进行鉴别诊断，诊断的准确率达到 92.7%，使用神经网络综合

分析患者各项数据的分类准确性较单纯影像更高。人工智能技术除了能鉴别诊断肝肿瘤，还能应用于肝肿瘤 CT 图像的三维分割。2017 年，Vorontsov 等人[290]提出了一个基于多层感知器的可变模型用于立体像素分类。借助纹理分析，该模型可以区分肿瘤组织和非肿瘤组织并自动拟合肿瘤边界，重建肿瘤三维模型，计算肿瘤体积。人工智能技术在组织学诊断上同样具有独特的优势。同年，Li 等人[291]研究发现，与病理学家依据经验的定性判断不同，分级模型可对 HE 图像上每个细胞核的类别进行统计，对病理切片作定量分析，分类准确性可达 0.811 ± 0.029。在区分正常细胞与高分化癌细胞时，系统提供的定量信息可为医师提供有价值的参考。在无创性肝纤维化诊断领域，整合多方面的数据结合人工智能技术对肝纤维化进行分级，同时全面评估患者出血[292]、肝癌[293]等并发症风险才是未来更具前景的方向。在影像学领域，2014 年，Gao 等人[294]运用了灰度梯度共生矩阵和灰度共生矩阵提取超声图像纹理特征供反向传播神经网络学习，诊断肝纤维化分期(S0～S4)的准确率分别为 100%、90%、70%、90% 和 100%。2017 年，Chen 等人[295]比较了 K 近邻算法、多元回归分析、随机森林、朴素贝叶斯算法和支持向量机算法结合弹性成像诊断肝纤维化的能力，结果表明这四种机器算法明显的优于统计学多元回归分析，其中随机森林算法的准确性最高。在血清学领域中，Lemoine 等人[296]在 2017 年进行的研究结果显示 γ-谷氨酰转肽酶/血小板比率(GPR)诊断肝纤维化的价值和传统谷草转氨酶/血小板指数(APRI)和 Fib-4 比较有升高，三者诊断显著性肝纤维化的 AUC 分别为 0.73、0.62 和 0.57，

诊断弥漫性肝纤维化的 AUC 分别为 0.93、0.89 和 0.71。Lu 等人[297]在 2018 年的研究中指出虽然 GPR 具有一定的诊断价值，但是仅仅通过血清学检查难以实现肝纤维化的诊断。

第二、食管疾病的诊断。2011 年，Johanson 等人[298]利用人工智能技术分析食管刷片检查结果，有效地提升了 Barrett 食管(BE)的诊断率和筛查效率。Chan 等人[299, 300]在 2016 年和 2017 年提出了一种非侵入性 BE 智能诊断技术，通过“e-nose”与患者呼出的挥发性有机物进行相互作用，测量其电子电导率分布，再使用人工神经网络对检测的数据进行分析，以确定 BE 患者的电子电导率分布特征。“e-nose”技术虽然不能从食管黏膜组织学层面作出明确的诊断，但是在 BE 大规模筛查和监测方面比是使用内镜检查更具成本效益。此外，“e-nose”诊断肠道艰难梭菌感染的敏感性和特异性均较高。上述研究的研究对象并非是真正意义上的筛查人群，因此相关结论有待进一步完善。

第三、胰腺疾病的诊断。以鉴别胰腺癌和假肿瘤性胰腺炎为例，目前在临床上常用的防范技术是超声引导下细针穿刺活检，但是阴性价值较低。对部分怀疑胰腺癌，但是穿刺结果为阴性的患者来说，并不能排除胰腺癌的可能性。Săftoiu 等人[301]在 2012 年进行的研究中发现以深度神经网络为基础的超声内镜弹性成像系统对 32 例胰腺癌和 11 例假肿瘤性胰腺炎的鉴别诊断价值较高，训练准确率高达 97%，测试准确率高达 90%。因此，未来人工智能系统所提供的诊断信息将为医师的决策提供更多的参考。

5.2.2 智能诊治在呼吸系统疾病中的应用

呼吸系统疾病是影响国民群众健康的重要疾病之一，人工智能在该领域的应用也取得迅速的发展。近年来，研发者在呼吸系统疾病的多个方面开发出了专家系统，在呼吸系统疾病的诊断、治疗和预后方面起到了较好的临床指导作用。以下主要从支气管哮喘诊断、肺癌诊疗、肺结核诊疗来了解相关的应用。

第一，支气管哮喘疾病的诊断。Burge 等人[302]开发了通过测量峰流速变化来诊断职业性哮喘的辅助诊断软件 Oasys 并解释了职业性峰流速最好的统计学方法是通过 Coté 等描述，该方法与 Oasys 相比，Coté 的灵敏度、特异度和精确度分别是 100%、50%、60%，而 Oasys 的灵敏度、特异度和精确度分别是 69%、94%、86%。由此可知，Oasys 具有更高的精确度。为了评价哮喘的严重性，“哮喘专家”建立了“AI 评分”制度，并与三种评分标准进行比较。这项研究选择了 100 例在门诊部第一次咨询的患者为研究对象，采用不同的评分标准研究哮喘严重水平的分布，AI 评分和三种评分标准之间的可信度通过 Kappa 和 MacNemar 进行检验测量。结果表明，按照 AI 评分测得哮喘的严重水平评分值要比其他标准高，MacNemar 检验差异均有显著性，因此该评分标准能准确判断哮喘患者的严重程度[303]。

第二，肺癌的诊断。肺癌主要对良、恶性肺结节进行鉴别诊断。2002 年，Matsuki 等人[304]开发了一种用于鉴别良、恶性肺结节的专家诊疗系统，研究人员使用人工神经网络去鉴别高分辨率 CT 下发现的肺结节的良、恶性，并通过接收器工作特性（曲线）分析方法评价人

工神经网络对放射科医生的影响。该项研究选择了 155 个肺结节直径小于 3cm 的患者（良性 56 例，恶性 99 例）为研究对象，对比通过 12 位放射学医师判断和通过人工神经网络判断高分辨率 CT 片子的结果，结果表明运用人工神经网络比人工判断的准确性高。MuCulloch 等人[305]开发了一种基于模式的计算机辅助诊断系统，运用系统在薄层 CT 扫描中自动识别实囊性肺结节,将大大提高肺癌诊断的准确率。

第三，肺结核诊断治疗。“Artificial nose”智能诊断系统[306]用于快速发现幽门螺杆菌、结核菌等。初步实验结果表明，利用这种技术可以提高结核菌的检出率。目前有团队研制了肺结核诊断治疗专家系统，该系统具有三种功能：诊断治疗功能，提供患者的诊断结果及治疗方案；鉴别诊断功能，鉴别其他常见的肺部病变；病员管理功能，具有查询、输入、删除、打印等功能。因此，在计算机专家和医学专家们的共同努力下，专家系统将成为临床医生的好助手，在呼吸系统疾病的预防、诊断和治疗发挥重要的作用。

5.2.3 智能诊治在骨质疏松症中的应用

骨质疏松症是一种骨代谢疾病，其特征是骨矿物质密度低(BMD)和骨组织微结构恶化，导致骨脆性增加，从而导致骨折风险增加[307]。骨质疏松症常发生在骨量的减少比身体更换能力更快时，致使骨强度大幅下降[308]。从九十年代开始，有研究者就开始对人工智能技术与骨质疏松疾病的诊断进行了研究。1997 年，Ongphiphadhanakul 等人[309]的研究表明使用综合危险因素与人工神经网络(ANN)技术将有助于确定骨质疏松性骨折的高风险子集。该技术作为需要 BMD 评估，

是一种有助于提高骨折预防效果的筛查工具。在这项研究中，验证了 ANN 性能在诊断低骨矿物质量(BMD)的常规统计方法中没有表现出优异的结果。

随着人工智能技术的发展，不同的技术应用于骨质疏松疾病诊断。因此，针对近几年的人工智能与骨质疏松症诊断的研究做了分析。2014 年，Tafraouti 等人[310]提出了基于分数布朗运动模型的特征提取与支持向量机(SVM)算法结合在骨质疏松症鉴定中的应用。该研究收集了法国奥尔良医院 77 名患者的骨骼 X 射线图像，原始图像被分为子图像并且以角度 θ 施加旋转，从而得到的信号是由分数布朗运动建模的。通过连接所有子图像的特征获得了每个图像的纹理特征，将这些参数输入四个核函数（多项式，RBF，二次型，线性）后，线性和多项式核函数在精度方面表现最佳，其结果分别为 95%和 93%。

Iliou 等人[311]在 2015 年从希腊人群中提取了一组 589 条的记录，进行了骨骼和实验室密度测定检查，应用多层感知器分类器和十倍交叉验证方法进行研究。考虑到预测骨质疏松症风险的 3-5 个诊断因素，将这些记录分为正常，骨质减少和骨质疏松症三类。同年，Liu 等人[312]在其研究中使用了 725 例包含两性的控制病例样本，其中包含了台湾大学国立医院的 228 例患者，第一例低创伤髋关节肿块的患者，215 例未发生髋部骨折的患者和 282 例随机选择的居民。这项研究中的预测模型仅涉及 60 岁以上的老年人，并侧重于预防髋部骨折的风险。对所有患者进行了同样的问卷调查，将收集的数据输入数据库，确定最重要的变量应用于敏感性分析和连接权重。使用这种方法，根

据重要性将输入变量分类，确定了 10 个主要变量分别为总 BMD、家庭跌倒、身高、BMI、高血压、大便失禁和教育，这些变量显示了预测髋部骨折的大部分贡献。在这项研究中，三层向后传播的 ANN 模型分别应用于男性和女性患者，男性由于模型复杂程度较低，表现优于女性。2016 年，Yu 等人[313]以 119 例住院患者为样本，使用 X 射线成像椎体渗透性增加、椎骨水平骨小梁消失、垂直纵向骨小梁增大、脊柱变形和椎骨骨折特征，这些特征都具有背部或身体疼痛、吸烟史、使用糖皮质激素的特点。17 个参数中 6 个是成像特征，6 个是由放射科医师和骨科医生临床提取的参数，剩下的 5 个是患者的主诉。结果显示敏感性分别为 94.5%和 63.6%，特异性分别为 96.9%和 87.5%。

以上是使用人工智能技术诊断骨质疏松证的一些研究。我们比较了这些研究中所使用的人工智能技术的类型、风险因素的数量、患者的数量和性别，在表 5-1 中列出。

表 5-1　使用人工智能作为骨质疏松症的诊断辅助汇总表

Year	AI	VAR	PAC	Gender
2014	SVM	16	77	M/F
2015	MLP	35	589	M/F
2015	MLP	10	725	M/F
2016	ANN	17	119	M/F

SVM: support vector machines, MLP: Multilayer Perceptron, ANN: artificial neural networks, AI: artificial intelligence, VAR: amount of variables, PAC: number of patients.

5.3 人工智能在疾病诊治中的发展前景

随着计算机技术、人工神经网络技术等一些非线性技术、人工智能技术、遗传算法的发展和成熟，各种疾病诊断的智能化和精确化成为可能。人工智能不仅在消化系统疾病、呼吸系统疾病和骨质疏松疾病诊断中的应用取得了较大进展，也在其他一些疾病的诊断中取得了较大突破，但是其在疾病诊断中的精确度还有待提高。如何利用人工智能技术处理多模态的医疗数据，充分利用各种数据、文本和影像等综合信息进行疾病诊断建模，提高诊断模型的精确度和泛化能力还需要进一步的研究，也是一项技术挑战。此外，由于医学领域的特殊性，其对诊断模型的可解释性也有较高要求。

根据目前人工智能技术辅助诊断疾病的发展状况，未来应该从以下方面推动人工智能技术疾病诊断的发展。从计算机技术方面而言，针对现有模型泛化能力不强，效率不高的情况，应该研究更好的算法，建立更加稳定的模型，提高疾病诊断的准确率。此外，针对准确率低的情况，应从疾病信息数据的提取方面入手，寻找更好的标注疾病信息的特征。

第六章 人工智能与药物开发

6.1 人工智能与药物开发概述

药物研发是一个系统工程，通常包括药物靶标的发现和验证、先导化合物的发现和优化，候选化合物的筛选和开发，以及最后进入临床研究等过程。该过程不仅研发费用高、研发周期长、而且研发成功率低[314, 315]。根据塔夫茨药物开发研究中心(Tufts Center for The Study of Drug Development)统计，一种新药开发的平均成本为 25.58 亿美元，上市将花费大约 10 年的时间，然而只有 10%~14%的候选药物能通过第一阶段的测试走向市场[316]。毫无疑问，要想提高药物的上市率，就需要新的思维、新的药物发现过程理念、高质量药物的创新方法以及以较低的生产成本，即新药研发亟需一场新的变革。目前，我国医药工业生产的药品中仿制外国的药品较多，自主创新的新药占市场份额少，医药产业总体经济效益低下，与发达国家的水平还有一定差距。缺乏具有自主知识产权的创新药物是造成这种状况的关键因素之一。随着国际知识产权保护的相关法律法规在我国的逐步实行，新药研制的自主创新性地位已日益重要。

人工智能(AI)在药物开放中的应用是指应用分析、学习和解释与药物相关的大数据来开发新药物的算法，将机器学习的发展以一种更加集成和自动化的方式结合起来。由于机器学习方法的发展以及化学和药理学数据的积累，人工智能技术作为一种数据驱动的计算方法，在药物设计领域已发展成熟。基于机器学习的方法，作为人工智能的

一个分支，与传统方法相比并不依赖于复杂的物理和化学具体原理的理论进步，而是更加注重将生物医学大数据转化为新的洞察能力和可重用的知识。机器学习的常用算法有逻辑回归(LR)、朴素贝叶斯分类(NBC)、K 最近邻(KNN)、多重线性回归(MLR)、支持向量机(SVM)、概率神经网络(PNN)、二元核鉴别(BKD)、线性判别分析(LDA)、随机森林(RF)、人工神经网络(ANN)、偏最小二乘(PLS)、主成分分析(PCA)等[317]。近年来，人工智能技术，特别是深度学习模型由于其强大的通用性和特征提取能力，在药物开发中显示出广阔的前景。传统的机器学习方法采用人工设计的特征，而深度学习方法可以从输入数据中自动学习特征，通过多层特征提取，将简单特征转化为复杂特征。此外，与传统的机器学习方法相比，深度学习方法的生成误差较小，使得其在一些基准测试或竞争测试中获得更令人满意的结果。因此，深度学习方法作为一种数据挖掘方法，在药物设计领域显示出巨大的前景。深度学习方法主要包括深度神经网络(DNN)、卷积神经网络(CNN)、递归神经网络(RNN)、自动编码器、受限玻尔兹曼机(RBN)[317]。

在过去的几年中，得益于图形处理单元(GPU)的广泛使用，计算机数据处理的能力得到了迅速提升，加上无限可扩展存储的使用，大量不同类型生物数据（如图像、文本信息、可穿戴设备、化验信息和高维多组学数据）的获得以及先进算法的开发，很大程度上推动了机器学习的发展[318]。人工智能本身及其在各类领域的应用已经从大量的理论研究转向生产实践。随着研究的深入，新药研发领域产生了大

量数据，这也让人工智能有了用武之地。人工智能在新药开发领域的主要应用包括根据序列预测蛋白质折叠、蛋白质-蛋白质相互作用的预测、药物重定位、虚拟筛选、活动评分、qsar、生成模型、理化性质分析、药物吸收-分布-代谢和排泄分析、毒性与 ADME/T 多任务神经网络研究等。来自 Tech Emergence 的报告研究表明，人工智能将新药研发速率从 12%提高到 14%，为生物制药行业节省数十亿美元，大量的国内外 AI+新药研发公司获得市场融资。这表明 AI+新药研发已经进入快速成长期。

6.2 药物开发智能分析

尽管大部分服务行业早已应用人工智能领域的相关新方法提升其工作效率，但人工智能方法在制药行业的应用比较滞后，直到近几年才有所改善。众所周知，药物开发的成功率（从第一阶段临床试验到药物批准）在所有治疗领域和全球制药行业都非常低。在人工智能应用中，统计和计算技术被应用于学习复杂的关系和构建模型，制药行业中业务需求的使用驱动人工智能技术的使用，以降低总体损耗和成本。因此，人工智能将在药物发现中发挥越来越重要的作用。在药物发现和开发的所有阶段，包括临床试验，已着手开发和利用人工智能算法和软件来识别新靶点、为目标疾病关联提供更有力的证据、改进小分子化合物的设计和优化、深入了解理解疾病机制，加深对疾病和非疾病表型的理解，为预后、进展和药物疗效开发新的生物标志物。

6.2.1 药物靶标识别

药物的潜在用途通常表现为它与特定药物靶标的结合和调节

[318]。确定药物目标可以缩短新药物发现的时间并提高新药物研发的成功率。虽然某些蛋白质的结构表明它们可以成为潜在的药物，但其中许多蛋白质的结合位点并不产生治疗作用[319]。对疾病治疗和疾病预防而言，了解药物如何结合和调节其靶蛋白的功能至关重要。药物发现中最重要的就是开发药物（如：小分子，肽，抗体等），通过它们调节分子靶标的活性来改变疾病状态，其基本假设为靶标的调节将导致疾病状态的改变，根据现有证据选择该靶标称为靶标识别和优先化。现代生物学的的一个主要特点就是生物数据越来越丰富，包括大规模人群中的人类遗传信息、健康个体以及特定疾病的转录组学，蛋白质组学和代谢组学数据。获得这些大型数据集并通过公共数据库重用这些数据为早期靶标识别和验证提供了新的机遇[335]。药物靶标的识别在药物发现和疾病机制研究中具有重要作用，药物靶标识别方法的发展已成为该领域研究的热点问题。一个潜在靶标是否有成药性需要考虑几个方面：(1) 靶标的 RNA 表达是否与其蛋白表达和疾病假设相关。(2) 是否参与某种代谢通路或分子相互作用网络。(3) 是否具有成药性。(4) 其基因定位是否与遗传标记相关等。

药物靶标的发现是一个复杂的过程，最初科学家利用生物实验来鉴定药物目标。生物实验识别靶标的核心在于微生物基因组，微生物基因组学研究包括微生物，如病毒、细菌、小真菌以及单细胞动植物。表 6-1 列举了目前可识别靶标的生物实验方法。

表 6-1 目前可识别靶标的生物实验方法

生物技术	内容描述
免疫印迹分析	根据抗原抗体识别的特异性检测和蛋白质鉴定
微生物基因组学	微生物检查
趋化性研究	在某些化学物质的作用下，生物体进行定向运动
克隆	生物体通过体细胞进行无性繁殖以及后代的无性繁殖
酶联免疫吸附试验	定量测定
Matrigel 侵犯分析	反映细胞的浸润能力
细胞芯片	细胞特异性结合配体的鉴定
基因转染	核酸进入细胞的转运
RT-PCR 分析	实时监控整个 PCR 过程
差异蛋白质组学	发现不同因素引起的蛋白质表达的差异
核磁共振	生物膜结构研究
基因敲除动物模型	先改变生物体的遗传组成， 再研究改变基因的生物学功能

基因转染实验将核酸转运到细胞中，使其在细胞内维持其生物学功能。基因敲除实验改变生物体内一个基因的遗传组成，然后检查该基因的特定功能的变化，使其不受生物学功能的影响。酶联免疫吸附试验依据酶的特异性抗原抗体反应设计，酶的特异性着色是由酶与底物的结合产生的反应，从而可以进行定量测定。RT-PCR 分析将荧光基团添加到PCR反应系统中，并允许使用荧光信号实时监测整个PCR过程。

然而，随着人类基因组学和信息技术的快速发展，最近越来越多的科学家使用机器学习计算方法来确定药物目标，许多具有重要功能的生物大分子（如蛋白质、核酸、酶等）的三维结构被解析，成为药

物研究的新靶标。蛋白质药物靶标的识别可以看作是一个二元分类问题，药物靶标可分为蛋白质和非蛋白质药物靶标，机器学习算法广泛用于区分这两类靶标。机器学习算法的性能可以通过良好的特征选择提高。在机器学习方法中，序列特性，包括氨基酸和二肽含量（频率），可用于预测蛋白质药物目标[320-326]。结构氨基酸和蛋白质特征可根据其序列进行计算，通常用于准确预测蛋白质功能[327-329]。此外，已有研究表明，已知的药物靶标具有共同的功能、结构、物理化学和定位特性[330]，因此其网络拓扑特征、组织表达特征和亚细胞数据也是训练机器学习算法中常用的特征[331]。基于随机 SVMS[332]、逻辑回归、决策树、综合分类器、径向基函数和贝叶斯网络[333]的机器学习算法已广泛应用于药物目标识别工作。通常机器学习方法预测蛋白质药物靶标有三个步骤：首先选择数据集和特征，其次选择学习算法，最后评估模型的预测性能。

近年来，生物实验和机器计算方法在蛋白质药物靶点鉴定中的应用越来越广泛。生物实验方法需要对靶标疾病生物学的分子水平有深入的了解，虽然目前相关的理论知识研究更加深入，实验方法有所改进，但仍存在许多不确定性。因此，计算模型方法成为主流的药物靶标鉴定方法。随着高质量数据集的整合和高精度计算方法的出现，机器学习算法成为机器计算方法的主要组成部分。通常，选择一组最佳特征来预测具有相似特性的新药靶点。目前广泛使用的特性包括序列特性、网络拓扑特性、结构特性和亚细胞位置。由于机器学习方法种类繁多，因此其性能改进需要结合更好的特性子集，并为所涉及的各

种数据集选择适当的模型。虽然机器学习方法准确度高，成本低，大大增加了研究中确定的药物靶标数量，但其仍需要生物实验进行辅助预测，为此研究人员应将生物学实验和机器学习方法结合起来，以更加准确的确定药物靶标。通过机器学习方法初步预测潜在的药物目标，再进行生物实验验证这些预测的准确性，这样就能达到事半功倍的效果。然而由于数据集不平衡、不完整或特征选择方法不完善，生物和计算方法仍有许多局限性，为此更好的解决方案仍待进一步研究。下面具体介绍几种上述两大药物靶标识别类型中常见的识别方法。

6.2.1.1 生物化学方法

长期以来，具有生物活性的小分子化合物被用来调控生物过程，同时众多生物化学技术手段被用来检测活性化合物的结合靶标，其中搜寻目标小分子作用靶标蛋白最为直接有效的方法是亲和纯化，该方法是利用生物分子间特异性结合的作用原理进行生物物质的分离纯化(Target identification in chemical genetics: the often missing link)。亲和纯化方法已被广泛应用于天然产物与合成化合物的蛋白靶标识别，但是该方法倾向于识别高亲和力配体与高丰度靶蛋白的结合，同时由于严格的洗脱条件会造成蛋白识别偏差。此外，研究人员还通过化学和紫外线诱导的交联方法来对靶蛋白进行共价修饰，提高应用亲和纯化技术识别底丰度蛋白的能力。亲和层析(Affinity chromatography)与质谱(Mass Spectrum)等新技术的结合使用为敏感且无偏的靶蛋白识别方法提供新的启发。随着组学数据的不断累积，实验方法已经不能

满足高通量大规模数据分析的需求，计算化学生物学方法的出现和发展为药物新靶发现提供了技术支撑，并为靶标识别提供新的思路。

6.2.1.2 蛋白可药性评价方法

通常具有治疗效应的靶标需要满足“疾病修饰”和“成药”两个性质，即靶标蛋白在疾病表型产生和发展过程中发挥重要作用，且可以被药物分子调节[336]。一方面，随着人类基因组计划的完成，完整的基因组序列被成功测定，人们获得了前所未有的大量潜在的靶标，人类基因组中仅与疾病相关的潜在的药物靶标就有 3000 多个。此外，还有更加庞大的来自微生物和寄生生物的蛋白数量，都可以作为潜在治疗的药物靶标。然而在这些潜在靶标中仅仅有数百个是可成药且能被疾病修饰的药物靶标，因而识别这些潜在的蛋白质药物靶标成为药物研发所面临的一个重要挑战。Hopkins 等[337]人通过分析疗效靶标的基因组序列同源性评估了蛋白可药性，该方法利用 InterPro 域映射策略预测了包含成药基因的新的 3000 个治疗靶标(The druggable genome)。此外，研究者还开发了仅依赖于靶蛋白三维结构，无需特定蛋白家族的计算方法来识别可药蛋白。预测可药蛋白的第二步是评估结合位点是否能够高亲和、特异地与药物分子结合，其最直接的方法是对化合物集合进行生物化学手段筛选，进而识别命中小分子的数目和类型，但该方法需要确保搜索的化合物空间足够大。因而，基于核磁共振(Nuclear magnetic resonance)的方法筛选片段库来评估靶蛋白潜在成药性成为了主要方法。

6.2.1.3 药物信息学方法

药物信息学预测靶标的方法可以分为化学信息学预测方法和生物信息学预测方法。生物信息学预测方法与传统的细胞、分子生物学、药理学等技术紧密结合，并广泛应用于发现和验证新型药物靶标。化学信息学预测方法主要根据“相似特征原理”，即化学结构相似的化合物具有相似的理化特征和生物活性，将化合物的潜在靶标注释信息通过简单的二维相似性与已知药理学特征的药物关联。此类方法通常受制于二维分子指纹表征和相似拓扑结构的有偏性，最有名的算法是相似性集合方法 SEA(Similaity Ensemble Approach)。Keiser 等[338]通过该方法证实了包括 GPCR、离子通道以及 HIV 逆转录酶在内的多个靶标之间存在预测的多重药理关联，同时揭示了与医疗领域相关的药物可能存在更多的潜在靶标。Mestres 等[339]人研发了用于探索药物分子药效空间的一体化的工具 iPHACE，该软件通过产生药物和靶标的特异识别信息来进行计算。

6.2.1.4 基于网络的方法

将系统生物学扩展到药物-靶标和配体-靶标网络的基于网络的方法，通常也称之为“系统化学生物学”、“网络药理学”或“系统药理学”。在过去数十年，药物发现方法着眼于“一个药物，一个靶标”的思路，并认为具有高度选择性的药物会更加安全和有效，然而近年来的研究显示被认为是具有选择性的有效药物事实上作用于多个靶点，许多表型是由化合物对多个靶标的影响而引起的。因此，在一个统一的“药理空间”探讨药物化学结构、靶蛋白质序列和药物靶网络拓扑之间的关系，可以预测新的化合物和蛋白质的配体-靶标相互作用，这种方

案为传统的药物研发提供了新的策略[340]。目前用于药物目标识别的网络可访问数据库如表 6-2 所示。

表 6-2 目前用于药物目标识别的网络可访问数据库

可访问数据库	网址
药物基因组学数据	www.pharmgkb.org
潜在药物目标数据库	http://www.dddc.ac.cn/pdtd
目标毒素数据库	www.t3db.org
蛋白质表达信息数据库	www.proteintalas.org
治疗目标数据库	http://bidd.nus.edu.sg/group/cjttd/
药物目标数据库	www.drugbank.ca
蛋白质数据库	www.pdb.org
毒性比较基因组数据库	http://ctdbase.org
PDBSite	http://wwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/
LigBase	http://guitar.rockefeller.edu/ligbase
SitesBase	http://www.bioinformatics.leeds.ac.uk/sb
MSDsite	http://www.ebi.ac.uk/msd-srv/msdsite
AffinDB	http://www.agklebe.de/affinity

(1) 基于网络推理的方法

推荐算法就是根据用户喜好数据，为用户推荐其可能喜欢的事物。在复杂网络领域，从网络的角度而言，该问题就是链路预测问题。为用户推荐可能喜爱的对象这一行为，可以抽象为在用户-对象关系网络中预测可能存在的边。因此，可以将药物看作用户、靶标看作对象，把用户-对象关系预测的算法，应用到药物发现领域，为药物推荐潜在的靶标。Cheng 等[341]人将一种推荐算法成功应用到药物发现领域，为药物-靶标相互作用预测服务，该算法称为基于网络推理

(Network-based Inference)又称为概率传播。实验结果表明该方法是一种具有一定应用价值的药物-靶标预测方法。在此基础上，又出现了边加权的网络推理方法(EWNBI)和节点加权(NWNBI)的网络推理方法。

(2) 基于相似性的推理方法

协同过滤算法也被应用到药物发现领域来预测潜在的药物-靶标相互作用，该方法同时具有 NBI 方法和相似性搜索方法的特性，计算过程中不仅需要已知的药物-靶标相互作用网络，还需要一些相似性信息，其中代表性的方法有基于药物相似性推理方法(Drug-based similarity inference, DBSI)和基于靶标相似性推理(Target-based similarity inference, TBSI)的药物-靶标相互作用预测方法。DBSI 方法的基本假设是具有相似化学结构的药物倾向于与相似的靶标产生相互作用。TBSI 方法的假设是具有相似序列的靶标蛋白倾向于被相似的药物靶向。

(3) 基于随机游走的方法

基于随机游走的方法不仅可以用于推荐系统，还可以用于基因-疾病关联、药物-靶标相互作用的预测。Chen 等[341]人提出了一种基于随机游走的药物-靶标相互作用预测方法-异质网络上的可重启随机游走(NRWRH)，该方法整合了药物-靶标相互作用网络、药物化学结构相似性、靶标序列相似性等信息，构建了异质网络，然后通过在该网络上进行带重启机制的随机游走来预测药物-靶标相互作用。此后的研究中，通过引入根据扩展连通性指纹、二维药效团指纹以及 ROCS 程序计算的几种新型药物相似性，使得 NRWRH 方法得到进一步改进。

(4) 网络结合机器学习的方法

靶标识别需要建立靶标与疾病之间的因果关联。因果关系的建立需要证明靶标的调节会对疾病产生影响，比如来自自然发生（遗传）变异或精心设计的实验干预。机器学习的方法可以用来分析具有靶标功能信息的大数据集，以预测潜在的因果关系和驱动关系。已经以这种方式在靶标识别领域的若干方面应用了机器学习的方法。Costa 等[342]建立了一个基于决策树的元分类器，通过它训练蛋白质-蛋白质、代谢和转录相互作用的网络拓扑结构，以及组织表达和亚细胞定位，以预测与发病率相关的基因，这些基因也具有可药物性。通过检查决策树，其以多种转录因子(TF)、代谢通路的中心性和细胞外位置作为关键参数进行鉴定。

6.2.1.5 基于文本挖掘的方法

文献是靶标与疾病相关的知识的主要来源，文献自动处理能解锁非结构化文本中的信息，否则这些信息将无法访问。自然语言处理(NLP)，一种应用于文本挖掘的 ML 方法，能有效挖掘并识别相关论文。BeFree 等人通过 NLP 核方法来识别 Medline 摘要中的药物-疾病，基因-疾病和靶-药物关联。这种监督学习方法依赖于手动注释的欧盟不良药物反应(EU-ADR)数据库关系语料库和基于遗传关联数据库的半自动注释语料库[343]。

6.2.1.6 小分子设计和优化方法

发现可以阻断或激活靶标蛋白的候选药物通常需要通过从化合物数据库中大范围的虚拟和实验高通量的方式进行筛选，然后进一步

修改和完善候选结构以提高靶标的特异性和选择性，以及优化药效、药代动力学和毒理学特性。但重要的是，由于新化学缺乏足够高质量的数据，如蛋白水解-靶向嵌合体(PROTACs)和大环化合物，数据的缺乏限制了机器学习计算该化学反应的影响。目前已经有大量的工作应用机器学习的方法，如多任务神经网络进行基于配体的虚拟筛选。给定一个先导化合物，可以通过计算方法识别具有相似化学结构的化合物，通常采用经典的统计方法进行识别，但多任务深度神经网络(DNN)被证明识别效果更佳，尤其是推断小分子的性质和活性时，DNN 可以显着提高预测能力。一次性学习技术能够大大减少在新实验装置中对分子读数进行有意义预测时所需的数据量。

6.2.1.7 基于定量构效关系的方法

定量够效关系的方法主要借助分子的结构或理化性质，通过数学手段定量研究有机小分子和生物大分子之间的相互作用。目前已经从使用相对简单的回归方法定量研究小系列同源化合物演变为使用各种统计和机器学习技术研究包含数千种分子结构的大规模数据集的方法[344, 345]。

6.2.2 药物重定位

药物重定位(drug repositioning)又称“老药新用”、“药物再利用”、“重审旧药”，参见图 6-1，通常是指已批准上市或处于临床研究阶段的药物被发现其具有超出原来医学使用范围的新用途。药物重定位包括对药物进行重定位(reposition)、重定用途(repurpose)、重评价(reprofile)、重新定位治疗方向(redirecting)等[346-353]。药物重定位的基

础首先是许多药物具有多个靶蛋白[354]，因此多靶药物可用于多个治疗目的，其次不同的疾病可共享遗传因素、分子途径和症状[355]，因此作用于这些重叠因素的药物可能有益于多种疾病。与开发全新药物相比，药物重定位具有以下优势：(1) 重定位失败的风险较低。因为已经发现重新利用的药物在临床前模型和人类中已经足够安全，如果早期试验已经完成，则在随后的功效试验中至少从安全的角度来看不太可能失败。(2) 能够减少药物开发时间。因为大多数临床前测试，安全性评估以及在某些情况下配方开发已经完成。(3) 需要的投资较少。尽管不同的药物之间有很大差异，它将取决于重新利用候选药物的发展阶段和过程而定。对于改变用途的药物和相同适应症的新药，监管和 III 期费用可能大致相同，但在临床前和 I 期和 II 期中仍然可以节省大量费用。因此药物重定位是目前已知的药物研发策略中风险/效益比最好的策略之一。

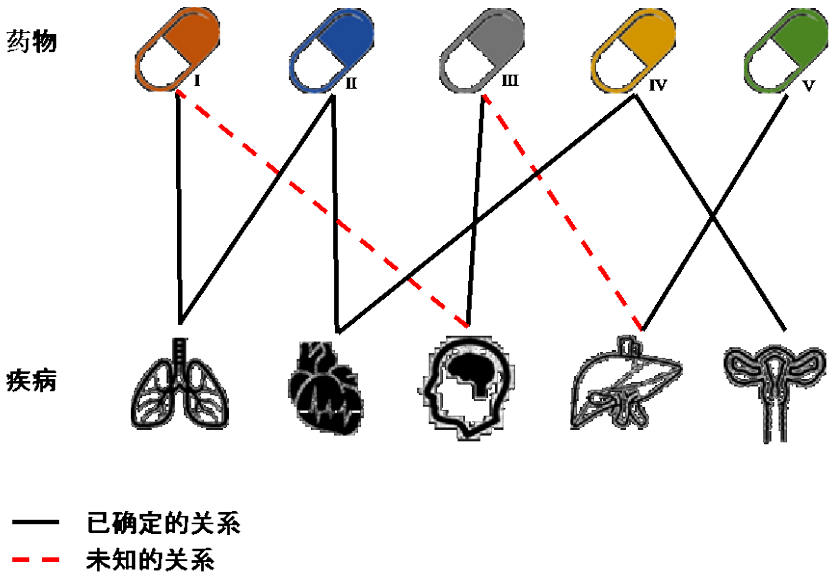

图 6-1 药物重定位[356]

通常，药物重定位策略包括三个步骤，首先识别用于给定适应症

的候选分子（假设产生），其次进行临床前模型中药物效应的机制评估，最后评估 II 期临床试验的疗效。在这三个步骤中，步骤一中确定正确的药物候选分子是至关重要的。目前进行该步骤的方法主要分为实验方法和计算方法。实验的方法主要是以高通量筛选技术为主的筛选方法，主要通过专门的仪器设备、特定的试剂盒进行测试并开展专门的数据分析和挖掘工作。该方法主要的缺陷是只能对较少的药物进行筛选，结果受药物化学性质、稳定性等因素影响。计算的方法是以大量数据驱动的，它们涉及到生物数据分析的方方面面，如基因表达、化学结构、表型、蛋白质数据或电子健康记录等，它们都可以用于形成药物重定位假设[346]。计算的方法主要是以计算机的虚拟筛选和生物计算为主。目前较高效的策略是先以计算机虚拟筛选获得候选药物，再利用高通量进一步筛选。从高层次上讲，药物重定位的方法可分为以下几种：基于蛋白质-靶点相互作用网络预测现有药物新用途的方法[357-359]、通过分析各种药物作用后的基因表达激活来预测药物的新用途的方法、基于药物副作用进行预测的方法[360, 361]等。考虑各种疾病相似性和药物相似性测量的方法，以下阐述几个具体的较为常见的药物重定位的方法。

6.2.2.1 基于化合物结构的药物重定位

化合物的分子结构决定了化合物的性质和功能，也决定作为药物的药效。基于化合物结构的药物重定位方法主要就是通过比较一个给定的药物与其它药物化学结构特征，从而发现化学上相似的是否意味着相同的生物活性。通常，合适的分子相似性度量有三个组成：表征

分子或者化学特征相关的特征呈现、特征呈现的权重表示、包含特征信息和权重参数相似性函数或者相似性系数[354]。Keiser 等人通过对每个药物选择化合物特征集并基于统计的化学方法预测了 878 个小分子药物的靶标和 2787 个化合物[338]。

6.2.2.2 基于药物副作用的药物重定位

药物意料之外的副作用是世界范围内致残、致死的主要原因，同时也严重阻碍了新药开发的进程。有关药物副作用的数据对药物重定位有重要意义。Tatonetti 等人基于非典型偏性的统计校正构建了药物作用数据库和药物-药物相互作用的副作用数据库，并利用该资源识别了药物靶标，预测了药物适应症[346]。

6.2.2.3 基于分子对接的药物重定位

分子对接是基于结构计算策略来预测配体和目标之间的结合位点，其包含一些列基于直接的物理相互作用的仿真和建模，用来发现新的药物与靶标关系。分子对接包括向前对接(forward docking)和向后对接(inverse docking)。其中，向前对接是指许多化合物与一个靶标的分析，向后对接是指一个化合物与许多靶标的分析[346]。通过使用高通量计算对接，Dakshanamurthy 等人在 FDA 证实的 3671 个药物、2335 个人类蛋白质结构上进行分子拟合计算，并发现了抗寄生虫药物甲苯咪唑具有抑制血管内皮生长因子受体 2(VEGFR2)的结构潜力，后者是血管生成的介质，之后实验证实了这一结果[347]。

6.2.2.4 基于多源数据整合的药物重定位

目前大多数方法仅使用单一数据源，而生物数据具有噪声或某类

数据较为稀缺都会对单一数据源的分析产生影响，因此通过多源数据整合进行药物重定位是药物发现的重要方向之一[348]。多源数据的药物重定位涉及到数据的整合，在该阶段中数据通常有两种类型：原初始数据(primary data)和衍生数据(derived data)。原初始数据通常指操作或工作数据库中的数据。衍生数据是指经过提炼、概况的数据，如原初始数据经过聚合、可视化、统计表征等形成的数据。因此，数据整合涉及到整合什么类型的数据，什么阶段整合以及用什么方法整合数据等问题。Li 等人基于蛋白质互作网络和文本挖掘整合基因、蛋白或药物连接的信息构建疾病特异的药物-蛋白连接图，并通过一个计算框架预测了候选药物[349]。Iwata 等人使用有监督的网络推理分类方法进行系统的药物重定位，该方法利用了每个药物-疾病对的特征，包括药物表型特征（治疗作用和副作用）和来自疾病国际分类 ICD-10 的疾病各种分子特征（致病基因、诊断标志物、疾病关联的通路和环境因素等）[350]。

6.2.3 药物靶向的相互作用预测

药物通过与靶蛋白结合并影响其下游活性而对人体产生影响，因此，药物靶向相互作用的识别对于药物的关键特性（包括药物副作用、治疗机制和医学适应证）的不确定性而言非常重要。近年来，许多研究集中在利用机器学习进行药物靶向相互作用预测上。药物靶向相互作用的原理是相似的药物倾向于共享相似的目标利益，反之亦然。利用这一原理，可以将预测表述为一个二元分类任务，其目的是预测是否存在药物-靶相互作用。这种分类方法将已知的药物靶相互作用视

为阳性标签，并使用药物的化学结构和靶蛋白的 DNA 序列作为输入特征（或核）[362]。此外，许多方法将副作用信息整合到分类模型中，如药物副作用[363]、基因表达谱[364]、药物疾病协会[365]和基因功能信息[366]，这些数据为药物靶向相互作用预测提供了多视图学习设置[367]。例如，使用 kernelized matrix factorization 并结合多种类型的数据（即视图），将每种数据类型视为不同的内核，以获得比单个内核方案更好的预测性能[368]。另一种常见的方法是将多种类型的数据表示为异构网络，并随机游动预测目标蛋白。这种方法通过扩散分布来计算网络中每个节点（蛋白质）的得分，从而使得分反映特定药物针对蛋白质的概率[369]。除了随机游动预测之外，还可以使用元路径从异构网络中提取药物和蛋白质特征向量，然后将它们输入分类器的方法进行预测[370]。

上述几种方法需要在特性工程中大量作业以及专业知识支撑，因此可以防止其获得的数据流失至大数据集，可以采用矩阵分解算法来学习异构网络向潜在特征空间的最优投影。学习的潜在特征空间用于通过矩阵运算的序列推断药物目标网络，产生的药物目标网络用于预测药物目标相互作用。经典矩阵分解法潜在的的局限之一是它将一个同构网络作为输入源。因此需要将一个异类网络折叠成一个同构网络，则会丢弃某些可能有用的信息，为此多可通过多视角、集体和张量因子分解打破此局限性，以预测药物靶相互作用[368, 371]。除了使用矩阵分解（浅特征学习算法）之外，还可以使用深特征学习算法，例如深自动编码器来集成与药物相关的信息。深特征学习算法为数据集

中的每种药物和蛋白质生成一个特征向量，根据药物和蛋白质特征，找到从药物空间到蛋白质空间的最佳投影。根据蛋白质在投影空间中与药物载体的几何接近程度对蛋白质进行排序，从而预先确定特定药物的目标蛋白质。该预测被用来最小化药物靶向相互作用训练数据集的预测误差[372]。

6.2.4 药物相互作用与药物组合预测

药物组合的使用是一种常见的治疗方法，许多患者同时服用多种药物来治疗复杂疾病或共存疾病[373]。由于药物组合中的药物可以调节不同蛋白质的活性，因此药物组合可以通过克服不脱轨生物过程中的冗余来提高疗效[374]。虽然使用多种药物可能是治疗多种疾病的一种良好做法，但对患者而言，药物组合的主要威胁是由于药物与药物相互作用而产生副作用的风险更高[375]。这种副作用具有出现的可能性，因为如果一种药物与另一种药物同时服用，其中一种药物或两种药物的活性可能会改变，这意味着联合用药会导致患者产生过度反应，而这种反应超出了我们在没有进行药物相互作用预测的情况下预期的效果。因此，药物相互作用是药物研究中的一个重要问题。一种给定的药物组合的副作用可以通过临床表现出来，而且每种组合仅对特定的患者子集有效。然而，药物组合方式多样，几乎不可能检测所有可能的药物对，并且在相对较小的临床试验中观察其副作用[376]。鉴于药物的数量众多，药物配对组合的实验筛选在成本和时间上成为了一个巨大的挑战。

为了解决该问题，开发了相关的计算方法来识别可能相互作用的

药物对。药物-药物相互作用通过协同作用和拮抗作用的概念来定义[377]，并通过测量剂量效应曲线[378, 379]或细胞活力[376, 380]进行生物学定量。计算方法使用生物学定量数据来识别可能相互作用的药物组合，通常是成对的药物，通过估计代表一对药物相互作用整体强度的分数来预测药物-药物相互作用。现有的计算方法主要基于分类或相似性原理。基于分类原理的 AP-Prophes 认为药物相互作用预测是一个二元分类问题[376, 380]，其使用已知的相互作用的药物对作为阳性例子，其他药物对作为阴性样本。首先获得每对药物的特征表示，然后将个别药物的特征向量聚合，以获得药物对的综合特征向量，最后设计二元分类器，如逻辑回归类分类器、支持向量机或神经网络对药物对进行特征表示。相比之下，基于相似性的方法假定相似药物具有相似的相互作用模式[381, 382]，此方法结合了药物化学亚结构、结构相互作用指纹、药物副作用、靶外副作用和分子靶点连接上定义的不同类型的药物相似性，通过聚类或标记聚合相似性指标，预测新药相互作用[383, 384]。

除了预测药物与药物相互作用的可能性外，最近有方法测定了给定药物对在患者群体中的临床表现[385, 386]。通过分子、药物和患者数据来预测与成对药物相关的副作用，例如 Decagon[385]构建了蛋白质-蛋白质相互作用、药物-蛋白质相互作用和药物-药物相互作用的多模式图，将每种类型的副作用表示为不同的边类型，并利用该模式图开发了一种图卷积神经网络，用于预测药物对的副作用。

6.3 人工智能在药物开发中的发展前景

药物的挖掘和筛选一直是医疗行业的重要领域之一，换言之，药物研发的水平和规模在某种程度上决定了医疗行业的发展形态。从历史上看，药物挖掘经历了随机筛选药物、组合化学库筛选和虚拟药物筛选三个阶段。最初，随机筛选药物的典型做法是通过细菌培养法从自然资源中筛选抗菌素，这种做法是初级低效的。随着组合化学的出现，人们可以迅速合成大量化合物，并在此基础上运用高通量筛选的技术完成化合物的筛选，但是这种做法的缺点主要就在于研发成本较高。到了目前的虚拟药物筛选阶段，人们可以在计算机上模拟药物筛选的过程，预测化合物可能的活性，从而进行更具有针对性的实体筛选，这样可以大大减少药物开发成本[388]。人工智能正成为现代生物医学研究的一部分，目前已经出现了许多可以整合不同生物医学数据集的方法，这些方法旨在提高大量数据的生成能力、加深研究者们对生物医学系统的理解，从而反映生物学的三元复杂性。尽管可能没有任何一种单一的方法能最好地解决所有问题，但人工智能的方法论发展和新出现的应用为生物医学数据集成提供了帮助。因此，需要根据不同类型领域的特定模型、特定类型的数据和不同类型的生物医学结果来选择合适的处理方法。随着信息化时代的到来，系统生物学和系统医学很可能成为一个新的交叉学科，以形成生物学和医学的新知识。此外，生物医学的海量数据引发了制药行业对人工智能的兴趣，计算能力的不断增加和大型数据集的产生，使得大量生物信息学的算法使得人工智能应用于在药物研发领域。

人工智能技术从经验和大量数据中学习，这些数据涉及遗传、基因组、化学数据库、化合物、副作用数据库、生物通路和疾病等，从而进行生物标识物的发现、靶标的选择、靶标的优先化，并显著减少临床实验中的失败率。药物开发的速度不仅取决于人工智能技术的先进性和应用程度，还取决于所涉及的学科（如遗传学、生物学、生物化学、药理学）的发展，AI 也能加快收集这些学科知识[388]。人工智能在药物开发领域已经应用了几十年，传统的机器学习建模已经发展成多种新模式，如 Combi-qsar 和 Hybrid-qsar，并且仍然是研究各种药物相关问题的主要方法。尽管在建模研究中使用机器学习方法（如 qsar）具有普适性和推广性，但近年来，深度学习正逐渐取代机器智能成为药物发现的新兴技术。深度学习方法的发展是由大量生物医学数据的积累和 GPU 强大的并行计算能力驱动的。更重要的是，深度学习方法无需人工输入就可以处理基于大型、异构和高维数据集的复杂任务。这些方法已被证明在许多生活和商业活动中得到了广泛应用，包括药物发现研究活动。

在制药行业，计算机软件在药品设计中的商业潜力是显而易见的。许多研究人员愿意将其项目，如 Deepchem、DeltaVina、SCScore 等，在 Github 或其他开源平台上共享他们的程序，将人工智能与药物开发方法结合起来，例如基于人工智能的药物设计模型的开源平台[317]。这些开源项目将促进人工智能技术在该领域的广泛应用。目前与人工智能药物研发相关的市场主要有 AI 技术公司、药物研究机构和大型药企。AI 技术公司的业务涉及到药物研发的各个环节，它们

并不生产药物，而是向药物研究机构和大型药企提供服务。药物研究机构和大型药企具有较高的研发水平，利用其拥有的海量数据不断促进企业彼此间的合作与投资，这已成为当前药物市场发展的趋势[389]。例如，2019 年美国最大的合同研究组织之一 Charles River Labs 与加拿大初创公司 Atomwise 合作，将 AI 驱动的方法应用于基于结构的药物发现(SBDD)。罗氏与 Exscientia 合作，利用其基于 AI 的药物发现平台 Centaur Chemist™设计临床前候选药物。丹麦 Lundbeck 制药公司宣布与 AI 驱动的药物发现公司 Numerate 达成协议，利用其专业知识和平台识别治疗中枢神经系统(CNS)疾病（包括抑郁症）的有希望的临床候选药物[390]。

人工智能技术，尤其是深度学习方法，可以用来从大量的药物数据中学习药物知识（例如，qsar 和化学结构）。然后将所学知识应用于发现和设计具有所需性质的分子，优化分子性质，提高分子的临床成功率。人工智能技术具有强大的数据挖掘能力，因此为计算机辅助药物设计注入了新的活力。但是，一些问题也会随之出现。一方面，由于神经网络的成功训练高度依赖于大量数据，因此，作为一种数据挖掘技术，可使用数据量的多少直接影响到相关的深度学习模型的性能，转移学习技术的发展可能是解决这一问题的一种潜在途径。另一方面，目前试图揭示深度学习模式机制的研究仍处于早期阶段。此外，神经网络模型的训练涉及多个参数的调整，其实际的指导思想较少，对这些模型进行优化的完整理论体系尚未建立，建立相关完整理论体系仍有很长的路要前进。在不久的将来，人工智能技术将涵盖新药发

现、开发的所有方面，为此致力于建立一个能整合所有理论计算结果（例如分子对接、分子动力学模拟和量子化学计算）、组学数据、化学数据和生物医学数据的自动化的药物开发人工智能平台，见证新药物发现的革命。

第七章 人工智能与基因组分析

7.1 人工智能与基因组分析概述

7.1.1 基因组的定义

基因组是指细胞内所有遗传信息，这种遗传信息以核苷酸序列的形式存储。在细胞或生物体中，一套完整的单倍体的遗传物质的总和称为基因组。

基因是生命遗传的基本单位，由 30 多亿个碱基对组成的人类基因组，蕴藏着生命的奥秘。现代遗传学家认为，基因是 DNA（脱氧核糖核酸）分子上具有遗传效应的特定核苷酸序列的总称，是具有遗传效应的 DNA 分子片段。基因位于染色体上，并在染色体上呈线性排列。基因不仅可以通过复制把遗传信息传递给下一代，还可以使遗传信息得到表达。人类许多表型如肤色、身高、头发颜色等的不同，均由基因组的差异所致。

7.1.2 测序技术的发展历史

DNA 测序可用于确定任何生物的单个基因的序列：较大的遗传区域（即基因簇或操纵子的簇）、完整的染色体或整个基因组。DNA 测序也适用于 RNA 测序。目前，DNA 测序已成为生物学、医学、法医学、人类学等学科的关键分析技术。

根据技术的不同，DNA 测序技术可以分为三个阶段 。

第一代测序技术，也称 Sanger 测序法。由 Frederick Sanger 发明的 Sanger 法是根据核苷酸在某一固定的点开始，随机在某个特定的碱基处终止，并且通过对碱基进行荧光标记，产生以 A、T、C、G

结束的四组不同长度的一系列核苷酸，然后在尿素变性的 PAGE 胶上电泳进行检测，从而获得可见 DNA 碱基序列的一种方法。Sanger 测序精度高，是金标准方法。

第二代测序技术，也称高通量测序技术，以能一次并行对几十万到几百万条 DNA 分子进行序列测定和读长较短为主要标志。大量文献中称其为下一代测序技术(next generation sequencing, NGS)，足见其划时代的改变。同时高通量测序使对一个物种的转录组和基因组进行细致全貌的分析成为可能。高通量测序平台的代表是罗氏公司(Roche)的 454 测序仪(Roch GS FLX sequencer)、Illumina 公司的 Solexa 基因组分析仪(Illumina Genome Analyzer)和 ABI 的 SOLiD 测序仪(ABI SOLiD se-quencer)。

第三代测序技术是指单分子测序技术，也叫从头测序技术，即单分子实时 DNA 测序。DNA 测序时不需要经过 PCR 扩增而实现对每一条 DNA 分子的单独测序。基因测序技术逐渐成为临床分子诊断中重要技术手段，第三代测序技术是未来主要发展方向。第三代测序技术按照技术原理主要分为单分子荧光测序和纳米孔测序[391]。单分子荧光测序代表性的技术为美国螺旋生物(Helicos)的 SMS 技术和美国太平洋生物(Pacific Bioscience)的 SMRT[392]技术。用荧光标记脱氧核苷酸，通过显微镜实时记录荧光强度的变化。当荧光标记的脱氧核苷酸被掺入 DNA 链的时候，它的荧光就同时能在 DNA 链上探测到。当它与 DNA 链形成化学键的时候，它的荧光基团就被 DNA 聚合酶切除，荧光消失。这种荧光标记的脱氧核苷酸不会影响 DNA 聚合酶

的活性，并且在荧光被切除之后，合成的 DNA 链与天然的 DNA 链完全一致。纳米孔测序的代表公司为英国牛津纳米孔公司。新型纳米孔测序法(nanopore sequencing)[393]是采用电泳技术，借助电泳驱动单个分子逐一通过纳米孔来实现测序。由于纳米孔的直径非常细小，仅允许单个核酸聚合物通过，而 ATCG 单个碱基的带电性质不一样，因此通过电信号的差异就能检测出通过的碱基类别，从而实现测序。

在分子生物学中，DNA 测序可被用于研究基因组及其编码的蛋白质。利用测序获得的信息，研究人员能够识别基因的变化、基因与疾病和表型的关联并确定潜在药物靶点。由于 DNA 是携带有遗传信息的大分子，在进化生物学中，DNA 测序被用于研究不同生物体之间的相关性以及它们是如何进化的。

宏基因组学是一门直接取得环境中所有遗传物质的研究。环境包括但不限于水体、污水、污垢、以及从空气中过滤出的碎片或者从生物体采集的样本。了解在特定环境中存在哪些生物体对于生态学、流行病学、微生物学和其他领域的研究至关重要。DNA 测序使研究人员能够确定微生物群中可能存在哪些类型的微生物。

医疗人员可通过对患者基因测序结果确定该患者是否有携带遗传性疾病的风险。需要注意的是，该方法属于基因检测，但有些基因检测不会用到 DNA 测序技术。

7.1.3 主要研究问题与领域

7.1.3.1 基因组组装

由于现有的测序技术无法获得完整的基因组序列，而是获得大量

基因组序列片段，因此需要计算机科学技术提供算法，将这些大量片段逐步组装为完整的基因组。

基因组组装一般分为三个步骤：contig 构建、scaffolding 和 gap 填充.。contig 表示从大规模测序得到的序列片段，即读数(reads)，从序列片段中找到的更长的一致性序列片段[394], 395]。组装的第一步就是从利用读数文库形成更长的 contigs。Scaffolding 是基于不同类型的读数文库（双端读数，长读数，Hi-C 读数等），利用读数文库和 contig 之间的比对信息，确定 contig 在基因组上的方向和先后顺序，最终形成一些 scaffolds(supercontigs 或 meatacontigs)。每条 scaffold 是一组确定了方向和顺序的 contig 组成的序列，其中相邻的两个 contigs 之间的 gap 区域用 N 来填充。最后对上一步得到的 scaffolds 间的 gap 区域进行填充，形成更加完整的序列。

7.1.3.2 碱基识别

碱基识别是测序的信号的识别过程。第二代测序的数据分析分为图像分析、碱基识别、序列组装、突变识别和功能分析五个环节，碱基识别是其中最核心的一步，是此后各种数据加工和分析的基础。其涉及到荧光信号交叉污染的校正、化学反应相位差异的校正、信号纯净度计算、碱基识别、碱基质量过滤和碱基质量评分等方面。

二代测序中的碱基识别[396]也就是从荧光信号的产生到碱基序列的识别这一过程，主要包括图象校正（空间校正）、簇的识别、荧光校正（光学校正）、phasing/prephasing（化学校正）、碱基识别、PF、质量评估等 7 个步骤。碱基识别(basecalling)主要是信号强度校正、交

叉污染矩阵(cross-talkmatrix)、化学校正(phasing correction)和碱基识别。在信号收集过程中，每个 tile 需要在 A、G、C、T 这 4 种波长处各拍摄一张黑白图像，分别纪录这 4 种碱基（簇）在 tile 里的分布信息。在图像处理过程中，需要将这 4 张图叠加在一起，生成一张虚拟的彩色图，以便同时包含这 4 个波长的信息。在序列读取过程中，还需要把每个 tile 的虚拟彩色图像按时间顺序排列，然后读取每个簇在每个测序循环的信号颜色，从而进行碱基识别。后两种运算都需要图像严格对齐，使碱基的坐标保持稳定不变。

7.1.3.3 变异识别

变异检测是基因组数据分析的重要任务之一，因为变异不仅影响个体表型，还与疾病特别是癌症紧密关联。一般而言，变异可通过以下几条途径影响个体的表型和疾病易感性。一是直接影响基因的拷贝数变化，进而导致基因表达量的改变。二是通过改变基因位置或者基因调控序列影响基因表达并诱发疾病。三是引起基因重排，诱发基因结构改变（基因断裂、基因融合），进而产生致病基因。此外，不同类型的变异相互作用也会对个体的表型和疾病易感性产生影响。

在人类基因组上的变异类型，大致可以分为 SNP、INDEL 和 SV 三个大类。SNP 主要是指在基因组水平上由单个核苷酸的变异所引起的 DNA 序列多态性，是目前最常见也最简单的一种基因组变异形式。INDEL 是指较短的 Insertion 和 Deletion，其长度通常在 50bp 以下，更多时候甚至不超过 10bp。SV 是指基因组大片段碱基序列结构变异。

单核苷酸多态性(single nucleotide polymorphism，SNP)主要是指

在基因组水平上由单个核苷酸的变异所引起的 DNA 序列多态性。SNP 是人类可遗传的变异中最常见的一种类型，占所有已知多态性的90%以上。SNP 在人类基因组中广泛存在，平均每 500~1000 个碱基对中就有 1 个，估计其总数可达 300 万个甚至更多。SNP 所表现的多态性只涉及到单个碱基的变异，这种变异可由单个碱基的转换或颠换引起，也可由碱基的插入或缺失所致。通常所说的 SNP 并不包括后两种情况。理论上讲，SNP[397]既可能是 2 个等位多态性，也可能是 3 个或 4 个等位多态性，但实际上，后两者非常少见，几乎可以忽略。因此，SNP 通常指 2 个等位多态性。这种变异可能是转换(C T，在其互补链上则为 G A)，也可能是颠换(C A，G T，C G，A T)。转换的发生率总是明显高于其它几种变异，具有转换型变异的 SNP 约占 2/3。转换的几率之所以高，可能是因为 CpG 二核苷酸上的胞嘧啶残基是人类基因组中最易发生突变的位点，其中大多数是甲基化的，可自发地脱去氨基而形成胸腺嘧啶。在基因组 DNA 中，任何碱基均有可能发生变异，因此 SNP 既有可能在基因序列内，也有可能在基因间区域。总的来说，位于编码区内的 SNP(coding SNP, cSNP)比较少，因为在外显子内，其变异率仅及周围序列的 1/5。但它在遗传性疾病研究中却具有重要意义，因此 cSNP 的研究更受关注。

7.1.3.4 甲基化识别

甲基化是指从活性甲基化合物（如 S-腺苷基甲硫氨酸）上将甲基催化转移到其他化合物的过程，可形成各种甲基化合物或是对某些蛋白质或核酸等进行化学修饰形成甲基化产物。在生物系统内，甲基化

是经酶催化的，这种甲基化涉及重金属修饰、基因表达的调控、蛋白质功能的调节以及核糖核酸(RNA)加工。甲基化是蛋白质和核酸的一种重要的修饰，调节基因的表达和关闭。其与癌症、衰老、老年痴呆等许多疾病密切相关，是表观遗传学的重要研究内容之一。

甲基化包括 DNA 甲基化或蛋白质甲基化。脊椎动物的 DNA 甲基化一般发生在 CpG 位点（胞嘧啶-磷酸-鸟嘌呤位点，即 DNA 序列中胞嘧啶后紧连鸟嘌呤的位点）。经 DNA 甲基转移酶催化胞嘧啶转化为 5-甲基胞嘧啶。人类基因中约 80%-90%的 CpG 位点已被甲基化，但是在某些特定区域，如富含胞嘧啶和鸟嘌呤的 CpG 岛则未被甲基化，这与包含所有广泛表达基因在内的 56%的哺乳动物基因中的启动子有关。1%-2%的人类基因组是 CpG 群，并且 CpG 甲基化与转录活性成反比。蛋白质甲基化一般指精氨酸或赖氨酸在蛋白质序列中的甲基化。精氨酸可以被甲基化一次（称为一甲基精氨酸）或两次（精氨酸甲基转移酶，PRMTs），将两个甲基同时转移到精氨酸多肽末端的同一个氮原子上成为非对称性甲基精氨酸或者在每个氮端各加一个甲基成为对称性二甲基精氨酸。赖氨酸经赖氨酸转移酶的催化可以甲基化一次、两次或三次。在组蛋白中，蛋白质甲基化是被研究最多的一类。在组蛋白转移酶的催化下，S-腺苷甲硫氨酸的甲基转移到组蛋白。某些组蛋白残基通过甲基化可以抑制或激活基因表达，从而形成为表观遗传。蛋白质甲基化是翻译后修饰的一种形式。

7.1.3.5 基因功能与可变剪接分析

可变剪接是指 mRNA 前体通过不同的剪接方式产生不同的

mRNA 剪接异构体，从而使同一个基因产生多个不同的 mRNA 转录本，进而能够翻译成多种不同的蛋白。可变剪接是调节基因表达和产生蛋白质多样性的重要原因，是真核生物转录组复杂性和多样性的重要原因。

可变剪接是真核生物基因中非常普遍的一种现象。在人和植物中，分别有约 95%和 60%的多外显子基因会发生可变剪接，因此可变剪接在转录调控、基因功能方面具有重要作用。可变剪接在动物的生长发育、细胞分化、细胞功能等方面具有重要作用。可变剪接在肿瘤中经常发生，与肿瘤发生发展密切相关。研究发现可变剪接可通过影响在肿瘤中经常发生突变的蛋白基因家族，进而改变肿瘤相关信号通路中的蛋白-蛋白相互作用，说明可变剪接也是驱动肿瘤发生的一种重要原因。

7.1.3.6 调控基因组学

基因调控是指生物体内控制基因表达的机制，表达的主要过程是基因的转录和信使核糖核酸(mRNA)的翻译。基因调控主要发生在三个水平上，即 DNA 水平上的转录调控和翻译控制；微生物通过基因调控可以改变代谢方式以适应环境变化，这类基因调控一般是短暂的和可逆的；多细胞生物的基因调控是细胞分化、形态发生和个体发育的基础，这类调控一般是长期的，而且往往是不可逆的。

基因调控的研究具有广泛的生物学意义，是发生遗传学和分子遗传学的重要研究领域。调控基因组学的主要的研究内容包括蛋白质与基因组的相互作用位点及预测、染色质开放区识别、转录调控网络构

建分析等方面。通过研究不同发育时期开放染色质区域的差异，分析调控染色质开放性的调控元件，预测这些调控元件在发育过程中何时何地处于活性状态，从而调控细胞发育类型和决定细胞命运。

7.1.3.7 疾病基因预测

检测特定物种中不同个体间的全部或大部分基因，从而了解不同个体间的基因变化有多大的方法称为全基因组关联研究(Genome Wide Association Studies, GWAS)。GWAS 在全基因组范围内对患者与对照样本的 SNP 位点进行比较，找出所有的变异等位基因，从而避免了像候选基因策略一样需要预先假设致病基因的问题。同时，GWAS 研究找到了许多从前未曾发现的基因以及染色体区域，为复杂疾病的发病机制提供了更多的线索。该方法通过表型选择、多重假设校正等方法对与疾病相关的基因进行预测，为人们打开了一扇研究复杂疾病机制的大门，对于精准医疗具有重大意义。

7.1.4 人工智能在基因组中的应用

随着人类基因组计划的实施和测序技术（第二、第三代测序技术）的快速发展，产生了大量的基因组数据。Hi-C 数据则使三维基因组序列组装成为可能。随着各种生物体基因组序列组装完成，基因组的结构变异预测和功能区域发现成为基因组数据分析的核心内容。长期以来，人工智能技术应用于基因组研究中，以机器学习为代表的人工智能技术被广泛使用在基因组组装、碱基识别、甲基化识别等基因组研究中。近年来，以深度学习为代表的人工智能技术逐渐成为公众的焦点。受到深度学习在不同领域成功案例的启发，深度学习技术开始

被探索及应用于基因组研究，如 DeepVariant 是 Google 基于深度卷积神经网络开发的一款突变检测工具，通过学习海量已标记基因组比对数据快照图像，实现从高通量测序数据中寻找基因变异进而完成基因分型的功能。DeepCpG 是一个用来预测多个细胞中 CpG 位点甲基化程度的深度神经网络，它可以精确地归纳不完整的 DNA 甲基化谱，以发现具有预测意义的序列改变，同时还可以量化序列变异带来的影响。DeepGO 是一个基于深度卷积神经网络开发的从序列来预测基因和蛋白质功能的工具，通过使用 GO 结构中的信息，并利用 GO 类之间的依赖关系作为背景信息来构建深度学习模型和学习特征。DeepGS 是基于深度卷积神经网络开发的用来预测基因型表型的工具，通过使用隐藏变量来联合表达基因型标记中的特征，提高了预测性能，此外对缺少异常个体和基因型标记子集有着一定的稳健性。

以上列出的仅仅是近年来人工智能在基因组领域所取得的成果的冰山一角。从中可以看到，人工智能已经逐渐演变成在基因组领域中可广泛使用的工具。通过使用人工智能技术，研究基因组组装、结构变异预测与功能区域发现等问题，可以使人们更加精准地透过生物数据认识生命规律，具有非常重要现实意义。

7.2 基因组组装

7.2.1 基因组组装概述与挑战

脱氧核苷酸(DNA)序列可组成遗传指令，引导生物发育与生命机能运作。因此，获取完整和准确的 DNA 序列是理解生命活动内在组织和过程的基础。基因组组装是将测序得到的大量 DNA 序列片段（读

数/read）还原成一个完整的 DNA 序列的过程，de novo 组装的主要流程如图 7-1 所示。

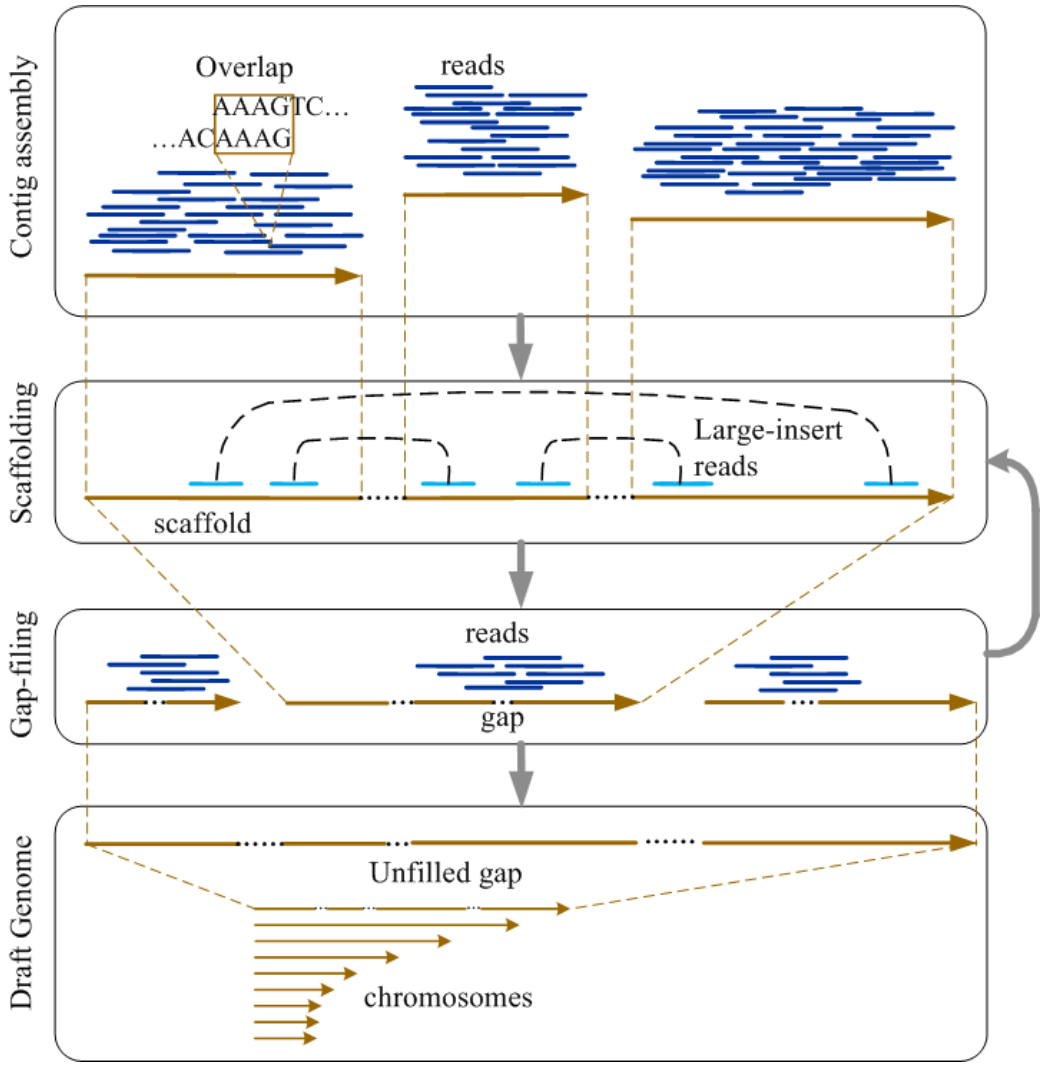

图 7-1　de novo 组装的主要流程[398]

基因组组装的主要目标是利用现有的算法技术在保证组装序列的连续性、完整性和准确性的同时，设计出耗时短、内存消耗小的组装算法。基因组组装的难点与挑战和测序技术密切相关。在第二代测序技术的背景下，基因组组装的难点和挑战主要表现在测序错误、测序不均衡、重复区和巨大的计算资源消耗这四个方面。而对第三代测序技术背景下的基因组组装而言，由于其测序片段平均长度可以达到

10kbp 左右，所以能克服重复区带来的问题，但是其测序错误率较高，达到 15%左右，因此其难点和挑战主要表现在测序错误率偏高、序列比对和序列纠错难度较大且需要耗费较大的计算资源。

第二代测序技术产生的测序片段准确度较高、长度较短、数目较多，目前适用于第二代测序片段组装的方法主要有两种，第一种是 OLC(Overlap-Layout-Consencus)方法，该方法将所有测序片段进行两两比对，如果任意两条片段之间的重叠区域长度超过一定的阈值，则在两条片段之间建立一条连边，以此来构建重叠图(Overlap graph)，然后在重叠图上选择一条能够尽可能贯穿所有节点的路径来生成组装结果。第二种是 De Bruijn graph 方法，该方法不以测序片段为基本的组装单位，而是将其切分成更加细小的单位，称之为 k-mer，其中 k 是细小片段的尺寸。De Bruijn graph 方法以 k-mer 为节点，当任意两条 k-mer 的后缀与前缀之间有 k-1 碱基重合时，则根据其先后顺序在两条 k-mer 之间建立一条有向边，并以此来构建 De Bruijn graph，然后在该图上寻找一条 Hamilton 路径来生成组装结果。De Bruijn graph 方法适用于海量、高精度的短序列片段组装，不适用于第三代测序技术产生的高错误率、长片段。目前用于三代测序片段组装的算法都是基于 OLC 方法而设计。

随着人工智能和深度学习技术的快速发展，利用智能化方法探索人类基因组密码已成为未来生物信息学发展的必然趋势。深度学习适用于大量、高维的数据集，其通过训练多层复杂的网络来捕获数据内部的结构。随着第三代测序技术的快速发展，基于 DL 技术的序列纠

错方法可以极大的提高三代读数的质量。第三代测序读数平均长度能够达到10kb以上并且可以部分解决传统DNA序列组装过程中遇到的若干挑战（测序不均衡和重复区问题），基于OLC graph和DL技术实现的组装方法可显著提高组装的精度和完整性。

7.2.2 基因组组装中的人工智能算法

随着DNA测序技术的进步，研究人员对生物体的遗传组成进行研究并取得了一些进展，特别是在健康科学方面。然而测序项目的复杂性产生了许多仍未得到解决的问题，其中之一是组装来自先前未测序生物体的DNA。这个问题被归为NP难（非确定多项式时间）问题，对不存在高效计算的解决方案，目前使用了几种近似算法的解决方案极大促进了这一领域的发展。与其他NP难问题一样，人工智能算法已成为近年来用于帮助解决DNA组装问题的工具之一。

宏基因组组装中，测序仪同时测得多个基因组的序列片段，如果能够根据序列的相似性将宏基因组序列数据进行聚类分组，将不同组的序列单独组装，这样就能够降低组装过程的复杂性。基于这个目的，Angeleri[399]等人利用递归神经网络学习序列之间的相似性，预测宏基因组中序列数据的所属类别然后再分开组装，在序列分组的基础上，还可以在不同分组内进行序列变异检测。Krachunov[400]等开发了一种序列变异检测算法，根据同一分组中所有序列在每个位置的碱基频率，利用机器学习方法预测低于某个固定频率阈值的候选变异位点是真实变异位点还是测序错误。

基因组组装过程中，根据测序数据两两读数之间的重叠信息连接

成更长的序列片段。Palmer[401]等人通过机器学习算法评价序列重叠准确性，从而剔除错误重叠，能够降低组装过程中的错误连接。除了过滤重叠区，研究人员还将组装时图的分支选择过程与机器学习方法相结合，这些分支通常源于重复区和测序错误。Zhu[402]等人利用 SVM 模型，预测 contig 遇到分支时是断开还是继续扩展连接，预测准确率达到 99.7%，使得组装得到更长更准确的 contig。Wang[403]等人将组装过程的 de Bruijin 图和隐马尔可夫模型将结合，组装结果生成了更多的 contig, 并具有更高的基因覆盖率。

在组装结果评价过程中，Lanc[404]等人利用无监督聚类方法，将组装结果的各个部分进行分组，其目的是将组装的错误区域与正确区域分离开来。基于从头组装通常不具有确定性验证结果的前提，Kuhring[405]等人提出了一项基于机器学习 contig 分类方法，对 contig 进行排序评估，从而更可靠地分析重建后基因组中包含的基因组信息。

7.2.3 碱基识别的人工智能算法

碱基识别是指根据基因测序仪器产生的信号序列分析出碱基序列的过程。不同设备采用的技术不同，会产生不同类型、质量和长度的信号序列。如第二代测序仪 Illumina 设备产生质量较高但较短的荧光信号序列，第三代测序仪 PacBio 设备产生质量较差但较长的荧光信号序列，第三代测序仪 Nanopore 设备则产生质量较差但较长的电流信号序列。不同碱基的信号有不同程度的重叠，信号质量越差越难识别。碱基识别的任务就是设计出有效的算法从信号序列识别出最可

能的碱基序列。碱基识别是基因组装、变异识别等后续操作和分析的基础。碱基识别算法的主要衡量指标是准确率。当前主流算法识别第二代测序仪 Illumina 设备产生的信号的准确率约为 98-99%，识别第三代测序仪 PacBio 设备和 Nanopore 设备产生信号的准确率约为 70%-90%。

为了提高碱基识别的准确率，在过去十年中出现了一批碱基识别算法。图 7-2 对现有识别 illnumina 设备信号序列的算法进行了分类，这些算法既有基于统计模型的方法，也有基于机器学习的方法。其中 Bustard 是 Illumina 官方所采用的识别方法，也是广泛使用的方法。它是基于非参数模型的方法，首先将荧光信号转换成表示质量分数的实际序列数据，然后标准化碱基浓度，使用马尔可夫模型确定转移矩阵，最后使用转移矩阵和观察到的每个碱基浓度确定碱基。

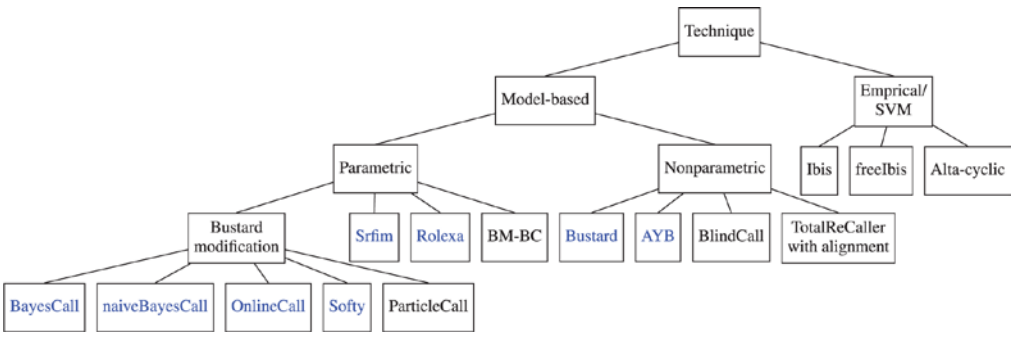

图 7-2 Illumina 信号序列的碱基识别算法分类[406]

识别 Nanopore 设备的信号序列的方法主要分成两类，基于马尔科夫模型的方法和基于深度学习的方法。基于马尔可夫模型的方法有 Metrichor 和 Nanocall。这两种算法都采用隐式马尔科夫模型，隐藏状态对应碱基序列。通过 Baum–Welch 算法计算转移矩阵，再根据观察的信号确定碱基序列。基于深度学习的方法有 albacore、guppy、

Chiron、DeepNano 等。它们的特点是从数据中学习模型特点进而识别碱基。这类方法逐渐成为识别 Nanopore 设备的信号序列的主流方法。

碱基识别本质上是分类问题。人工智能算法具有很好的分类效果，目前已有许多研究将机器学习和深度学习的算法移植到碱基识别算法上以提高算法准确率。第二代测序仪 Illumina 信号质量较高，基于模型的方法已经取得不错的识别效果，机器学习方法的引入可以减少模型假定，使算法具有更好的适应性。Alta-cyclic 学习模型运行特定的噪声模式，并使用 SVM 找出减少噪声源影响的优化解决方案。Ibis 不对每个可能的错误来源进行建模，而是直接对未经处理的强度信号使用多分类的 SVM 算法。Nanopore 设备的信号序列含有更多的错误和不确定性。利用深度学习能够自动从数据学习模型的特点，可以有效减少错误和不确定性的影响。因此基于深度学习碱基识别算法已成为识别 Nanopore 设备信号序列主要的算法。目前 Nanopore 官方的碱基识别工具 albacore 和 guppy 都采用的深度学习中的 RGRGR 模型。这是一个 5 层的反向 GRU 和正向 GRU 交替的结构，如图 7-3 所示。

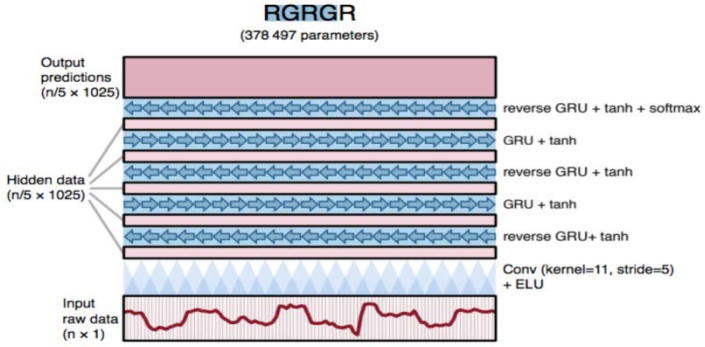

图 7-3 RGRGR 模型结构

首先通过一个卷积层对原始信号进行下采样，将采样后的特征传入 5 个堆叠的循环神经网络层提取时间序列的相关特征，最后通过输出层预测碱基序列。除此之外，由第三方研究者开发的 Chiron，采用的是一组卷积层加上一组循环层的结构，如图 7-4 所示。卷积层用来提取原始信号中的局部模式，循环层将局部模式变换成碱基的概率，模型顶层是 CTC 解码器，根据概率产生最终的碱基序列。

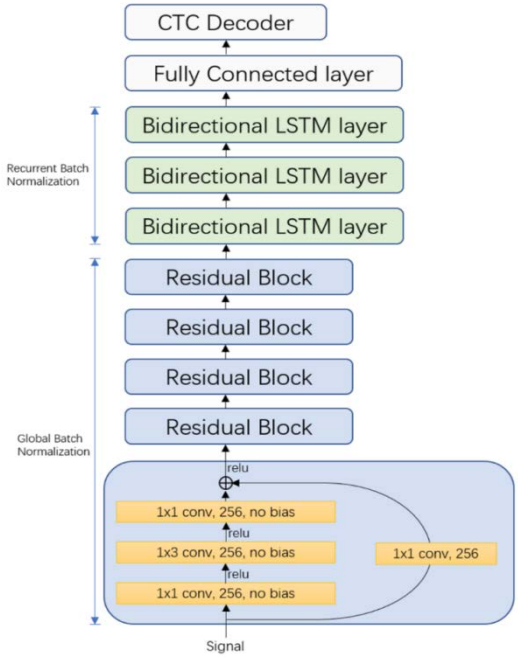

图 7-4 Chiron 算法结构[407]

7.3 变异识别

7.3.1 变异识别概述

变异是一个相对的概念，主要是指基因组之间的序列差异。目前关于人类基因组变异的讨论，大部分都是以“人类基因组计划”中所组装出来的人类基因组作为参照物，部分也有采用正常人群的基因组作为参照物。在人类基因组上的变异类型，大致可以分为如下三个大类，如表 7-1 所示。

表 7-1　人类基因组变异类型

变异类型	解释	定义
SNP	单核苷酸多态性	在基因组水平上由单个核苷酸的变异所引起的 DNA 序列多态性
INDEL	Insertion 和 Deletion	在基因组某个位置上发生的较短长度的线性片段插入或者删除的现象
SV	结构变异	在基因组上发生的较长长度的序列变化和位置关系变化

在研究中发现：SVs 对基因组的影响比 SNP 更大，一旦发生变异往往会给生命体带来很大的影响，比如出生缺陷、癌症等[408]，而且相比于 SNP，基因组上的 SVs 更能代表人类群体的多样性特征。稀有的一些结构性变异往往与疾病（包括癌症）的发生相互关联，甚至还有可能是直接的致病诱因。比如，在《我不是药神》电影中提到的慢粒白血病，它和基因组的结构性变异直接相关。它是由于细胞中的 22 号染色体长臂与 9 号染色体长臂相互易位，导致 ABL 基因和 BCR 基因融合，形成了一个会导致 ABL 异常表达的小型染色体（费城染色体）而发生的。这就是一个典型的结构性变异致癌的例子。

准确识别 DNA 序列变异是基因组学的一项重要而又具有挑战性的工作。基因组学的一个基本问题是找出单个基因组中相对于参考序列的核苷酸差异，即变异识别。因此准确和有效地识别变异是至关重要的，以便能够正确地检测表型差异和疾病的基因组变异。然而，由于测序错误率较高，单核苷酸和 INDEL 变异识别仍然具有挑战性[408]。

7.3.2 变异识别的主要算法

目前变异检测算法和工具已有很多，例如：GATK、FreeBayes、

SAMtools、16GT、Strelka、Sprites、Delly、Lumpy、Pindel、SVseq2、TIGRA 等。GATK 使用了逻辑回归、隐马尔可夫模型、朴素贝叶斯分类等技术，这些技术使得 GATK 在 Illumina 测序平台上高的准确性较高。16GT 是第一个使用 16 基因型概率模型进行变异识别的算法，与基于局部组装的变异识别相比，16GT 在 SNP 识别和 INDEL 识别中具有更好的灵敏度，但目前 16GT 只能应用于种系变异识别。变异检测算法和工具总结起来主要有 4 种基于 NGS 数据的变异检测算法的策略和方法，分别为：(1) Read pair（也称为 Pair-end Mapping，简称 PEM）；(2) Split read（简称 SR）；(3) Read Depth（简称 RD）；(4) 基于 de novo 组装的方法。

长读数可以跨越大部分的结构变异，通过比对工具 BLASR、Minimap2、NGMLR 等可以将长读数比对到基因组参考序列上，更容易识别结构变异，因此长读数可以显着提高检测结构变异的可靠性和分辨率的潜能。常见的方法包括 Sniffles、PBHoney、SMRT-SV、SVIM 和 NextSV 等。

7.3.3 变异识别的人工智能算法

单分子测序(Single Molecule Sequencing, SMS)技术的出现极大的加快了分子生物学的发展，且已用于重要的研究领域。DNA 序列变体的准确鉴定是基因组学中一项重要且具有挑战性的任务。对单细胞测序来说尤其困难，其核苷酸错误率约为 5%-15%。传统的算法无法处理如此高的测序错误率，尤其是存在大量 Indel 类型的错误。人工神经网络(ANN)由于其在许多领域中的速度和适用性的进步而在各

种分类和分析任务中变得越来越重要，使用神经网络模型可以预测这些变异，大大提高基因组应用的性能。

体细胞结构变异检测(SV)是基因组学中的重要研究课题。目前已经开发出许多方法来检测高通量测序数据中的体细胞变异。利用深度神经网络在真实数据集中植入体细胞突变数据，这样就能有大量的训练实例来训练模型，使模型更好的预测变异。

当前，深度学习结合第二代测序数据也被运用到结构变异检测中。基于比对结果，GARFIELD-NGS 针对不同类型的结构变异抽取了包含多种不同特征的一维数组，然后训练得到每种结构变异类型的深度学习模型。DeepVariant 和 CNNdel 都是基于卷积神经网络训练比对结果转化过来的图像进行分类，最终得到检测结果。CNNdel 为了解决输入尺度不一致的问题，采用了两种图像压缩方法，使所有长度的结构变异能压缩到相同的大小。最近 Google 开发出的 DeepVariant，通过将序列比对结果转化成图片，进而采用深度卷积神经网络实现对 SNP 和小的 INDEL 的准确检测。

7.4 甲基化识别

7.4.1 基因组甲基化与分析方法

表观遗传学研究的是在不改变 DNA 序列的前提下，通过某些机制引起可遗传的基因表达或细胞表现型的变化。DNA 甲基化在染色质结构的表观遗传调控中发挥重要作用，5mC 和 6mA 是最普遍和最广泛研究的两种甲基化类型。5mC 已被证明在胚胎发育、动脉硬化、老化以及其他疾病中具有重要作用。6mA 被证明在癌症发展和神经

发育中具有重要作用。

重亚硫酸盐测序技术(Bisulfite Sequencing)广泛地应用于 CpG 甲基化位点的检测。亚硫酸盐将非甲基化的胞嘧啶变成尿嘧啶，进而在全基因组扩增中变成胸腺嘧啶，甲基化的胞嘧啶保持不变。结合重亚硫酸盐技术，微阵列芯片技术和第二代测序技术都被用来进行 CpG 甲基化位点的检测。其主要技术有：(1) 微阵列芯片技术（Bisulfite Sequencing Microarrays），微阵列中的探针和待测数据都经过了亚硫酸盐处理，待测数据与探针结合后，用带抗原的单碱基扩展。荧光对抗原染色后扫描芯片后就可以得到甲基化位点信息。亚硫酸盐芯片技术不需要 PCR 扩增，同时可以与其他芯片技术结合，但是芯片的探针必须是设计好的，所以芯片技术只能测部分已知基因，而且荧光探测时存在强度偏差。(2) 全基因组亚硫酸盐测序技术(Whole Genome Bisulfite Sequencing, WGBS)，经过亚硫酸盐处理后的 DNA 片段通过 PCR 扩增，通过二代测序仪进行测序。测序后的 reads 和参考基因组序列进行比对，得到全基因组层面单碱基水平的甲基化信息。全基因组亚硫酸盐技术是目前检测 5mC 的金标准。然而在测序过程中，由于转换过程不完全，DNA 容易降解等原因，会导致结果中存在一定的假阳性位点。(3) 甲基化 DNA 免疫沉淀技术(Methylated DNA immunoprecipitation, MeDIP)也被用来检测 DNA 甲基化位点。该技术采用抗体或甲基化 DNA 结合蛋白来捕获富集甲基化 DNA 片段，结合微阵列芯片或二代测序技术，该技术可以检测全基因组的高度甲基化区域，但不能进行单碱基水平的分析。

第三代测序技术的发展为表观遗传学的研究带来了更多的机会。PacBio 的单分子实时测序和 Oxford Nanopore 测序均可在测序天然 DNA(Native DNA)时直接鉴定甲基化位点。经统计验证，利用 PacBio 测序仪器测序时，通过分析聚合酶沿模板 DNA 掺入核苷酸的速率-脉冲间持续时间(InterPulse Duration，IPD)来检测甲基化位点。如果当前的核苷酸被修饰，会使聚合酶在结合下一个核苷酸之前暂时停顿，从而导致 IPD 的可检测的变化。也就是说，正常位点和甲基化位点之间的脉冲间距是不同的，并且每种甲基化类型都有一个“特征”脉冲宽度。研究发现当 DNA 链穿过 Nanopore 测序仪器的纳米测序孔时，其产生的电信号对碱基的表观遗传变化敏感。

7.4.2 甲基化位点主要检测算法

基于二代测序的全基因组亚硫酸盐测序技术是检测 5mC 的黄金标准。目前，Bismark 是应用最广泛的针对全基因组亚硫酸盐测序技术的甲基化检测计算方法[410]。Bismark 首先利用比对工具将读数比对到转化后的参考基因组上，保留具有唯一比对的结果。然后根据比对结果判定位点是否甲基化，并确定位点所在的 motif。Bismark 的准确性主要依赖于比对的准确性，比对参数设置的不同也会导致结果的不同。

目前利用三代测序数据进行碱基修饰位点预测的方法通常都包含读数比对参考基因组和利用信号变化预测碱基修饰位点两个步骤。根据采取策略的不同，这些方法可以分为基于统计的方法和基于模型的方法两大类。

基于统计的方法一般都需要含有甲基化位点的 Case 样本和完全没有修饰位点的对照两组样本。PacBio 的 BaseMods[411]软件首先将样本测序产生的读数比对到参考基因组上，然后通过 Welch's t-test 判断每个位点在两组样本中是否有显著性差异。针对 Oxford Nanopore 数据的甲基化，也有两种基于统计学的方法，nanoraw[412]和 NanoMod[413]被提出。nanoraw 采用 Mann-Whitney U-test 来判断某个位点是否被修饰。实验发现利用 nanoraw 可以很好地发现大肠杆菌数据中的 4mC，5mC 和 6mA。与 nanoraw 不同，NanoMod 采用了 Kolmogorov-Smirnov test 计算 Control 样本和 Case 样本中相应位点的显著性差异。基于统计的方法不需要训练数据，并且可以检测任意类型的甲基化位点。但是，基于统计的方法同时需要含有甲基化位点的 Case 样本和完全没有甲基化的对照样本。而且此类方法一般对测序数据的覆盖度有较高的要求。例如，PacBio 建议使用超过 500X 覆盖度的测序数据来检测常见的 5mC 甲基化。

最近，为了利用 Oxford Nanopore 测序数据检测甲基化位点，研究人员提出了系列基于模型的方法。Nanopolish[414]和 SignalAlign[415]是两种基于隐马尔可夫模型(HMM)的预测方法。两种方法都首先收集为 HMM 模型训练数据，然后用训练得出的模型预测新数据中的碱基修饰位点。实验表明，Nanopolish 能够准确的预测出人类数据中的 5mC 甲基化位点。SignalAlign 将层次狄利克雷过程(hierarchical Dirichlet process，HDP)技术与 HMM 结合，可以同时区分同一位点的多种甲基化类型，如预测三种胞嘧啶(C，5mC 和 5hmC)。区别于

Nanopolish 和 SignalAlign，McIntyre 等人提出的 mCaller[416]利用多种机器学习模型(neural network，random forest，naïve Bayes，logistic regression)对大肠杆菌数据中的 m6A 位点进行了预测。与基于统计的方法相比，基于模型的方法在训练好模型之后只需要检测的测序样本，但是基于模型的方法对于不同的甲基化类型通常都需要训练不同的模型。因此，对于此类方法来说，如何收集准确的训练数据以扩展模型来研究多种类型的甲基化位点是未来研究中的一个挑战。

7.4.3 甲基化识别的人工智能算法

人工智能算法已经越来越多的被运用到甲基化位点检测当中。Angermueller 等人[417]提出了一种基于深度神经网络的计算方法 DeepCpG，用来预测单细胞中的甲基化位点状态。DeepCpG 利用了 DNA 序列模式和甲基化状态之间以及相邻 CpG 位点之间的关联。DeepCpG 由 DNA 模块，CpG 模块和连接模块三个模块组成。DNA 模块取以目标位点为中心的 1001bp DNA 序列作为输入，利用卷积神经网络(CNN)提取 DNA 序列特征。CpG 模块以目标 CpG 位点在细胞内相邻的 25 个 CpG 位点的状态作为输入，利用双向循环神经网络(BRNN)提取其特征。连接模块结合 DNA 模块和 CpG 模块的输出，用 sigmoid 激活函数确定目标 CpG 位点在不同细胞里的甲基化状态，通过在其他测序方案生成的五种细胞类型的单细胞甲基化数据的评价得出结果。DeepCpG 比已有的方法有更高的准确性。Fu 等人[418]提出了一种融合多种生物信息的深度学习模型 MHCpG，用来预测 DNA 位点的甲基化状态。和 DeepCpG 相同，MHCpG 首先取 1001bp

长度的 DNA 序列作为 CNN 的输入。其次，MHCpG 认为只有序列信息，而没有具体的细胞类型的信息，可能会造成准确性不足。因此 MHCpG 融合了具有组蛋白的甲基化信息和 MeDIP-seq 的信息，与序列信息一起作为 CNN 的输入。经过 CNN 提取其融合后的特征，MHCpG 利用全连接网络和 sigmoid 激活函数进行目标 CpG 位点的甲基化状态预测。实验结果表明，通过结合多种信息，MHCpG 能比其他已有方法获得更高的准确度。基于 Hi-C 数据和 DNA 甲基化二代数据，Wang 等人[419]提出了结合基因组拓扑信息和 DNA 序列模式的 DNA 甲基化预测方法 DeepMethyl。DeepMethyl 采用了降噪自编码器模型，包括两个阶段。第一阶段是无监督预训练阶段，利用训练数据学习构建函数。第二阶段是调参阶段，该阶段利用训练数据进行参数优化。实验结果表明，通过利用深度学习模型，DeepMethyl 要优于基于机器学习的方法。利用 Nanopore 三代测序中电流信号对甲基化碱基的敏感性，Ni 等人[420]提出了基于双向循环神经网络（BRNN）和 CNN 的深度学习方法 DeepSignal 来预测 DNA 甲基化位点。DeepSignal 取包含目标位点的 17 个碱基的序列及其电流信号作为特征，在预测人类 CpG 的甲基化水平上优于已有方法。Liu 等人[421]也提出了基于 BRNN 的方法 DeepMod。DeepMod 在 6mA 的检测上取得了较高的准确度。

除了预测单个位点的甲基化状态之外，深度学习也被用来预测特定区域中的甲基化模式。Peng 等人[422]研究了基因启动子中的甲基化模式与不同类型的癌细胞中的基因表达之间的相关性，并开发了一种

新的深度学习框架 E2M。E2M 基于 CNN 和全连接神经网络，通过捕获整个基因表达的统计特性来预测几种生物标记基因的启动子区域中不同甲基化位点的状态。在 3671 个癌症样本的验证中，E2M 的准确度为 82%。

7.5 基因功能与可变剪接分析

7.5.1 基因功能注释与可变剪接预测

基因的功能主要由蛋白质体现，蛋白质是生命体中最基本、用途最广泛的大分子，是细胞的重要组成部分，在细胞重量中所占的比例仅次于水。蛋白质负责生物体中一些最重要的功能，生物分子中最重要的一类是蛋白质。实际上每一种蛋白质都参与新陈代谢、身体运动和结构支持。了解蛋白质及其编码基因的功能，在新药开发、新作物开发、甚至生物燃料等合成生物化学品的开发等生物技术和医药方面发挥着重要作用。

基因功能可以从表型、生理等不同程度进行描述。为了获得所有不同程度的特征，基因本体论提供了细胞组成、生物过程和分子功能三种不同的功能分类。细胞成分决定了基因激活的结构成分的位置。生物过程获得蛋白质功能的功能定义，并允许指定细胞内基因的过程。分子功能描述了涉及细胞的基因产物。传统的生物实验方法，如基因敲除、靶向突变和抑制基因表达都需要耗费大量的人力和物力。随着基因、蛋白质相关数据的大量增加，人工智能技术将进一步提高基因功能预测的精准度。

可变剪接，又称选择性剪接(Alternative splicing)是基因表达的方

式。真核细胞的基因序列中，包含了内含子(intron)与外显子(exon)，两者交互穿插。其中内含子在基因转录成 mRNA 前体后一般会被 RNA 剪接体移除，剩下的外显子才是能够存在于成熟 mRNA 的基因片段。一条未经剪接的 RNA，经过不同类型的可变剪接，便可将同一基因中的外显子以不同的方式组合表现出来，使一个基因在不同时间、不同环境中能够制造出不同的蛋白质，参与不同的生物体功能。可变剪接的基本类型如图 7-5 所示。其基本类型为(1) Exon Skipping，称为外显子跳跃。(2) Mutually Exclusive Exons，称为互斥外显子，是只有部分外显子出现在成熟 mRNA 中的情况。(3) Alternative 5’ donor site，简称 A5SS，是指 5’剪接供体位点的变化导致外显子长度变化。(4) Alternative 3’ acceptor site，简称 A3SS，是指 3’剪接受体位点的变化导致 exon 长度发生了变化。(5) Intron Retention，称为内含子保留。

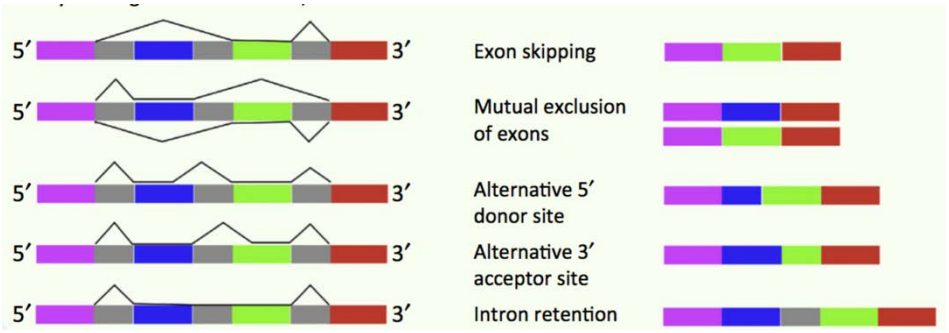

图 7-5　常见的五种可变剪接模式[423]

可变剪接的分析方法主要是基于高通量 RNA-seq 数据，以及一些基因特征和使用数学方法分析数据特征，识别各种类型的可变剪接。迄今为止已研发出的主要工具包括 MISO，rMATS，CASH，leafCutter，IRFinder，iREAD 等。

7.5.2 基因功能预测的人工智能算法

随着高通量生物技术的快速发展，已经产生了大量的生物学数据，这些数据为预测基因功能提供了有价值的信息。目前主要使用基因本体(Gene Ontology)来标注基因或蛋白质的功能。基因本体论包含了细胞组成、生物过程和分子功能。基因本体共有超过 40000 个功能类别。过去十年中一些基于机器学习的计算方法被广泛用于基因功能预测，其中 Barutcuoglu[423]开发了一个贝叶斯框架，用于根据功能分类约束组合多个分类器。Vinayagam[425]开发了一个大规模的注释系统，并通过应用多个 SVM 来分类正确和错误预测，通过 Gene Ontology(GO)术语提供注释。Li[426]提出了多实例层次聚类(MIHC)，该方法主要基于多标签学习和基因本体层次结构信息进行功能预测。Troyanskaya[427]设计了一种基于贝叶斯的方法(MAGAG)来整合异构类型的高通量生物数据进行基因功能预测。Guan[428]利用 GO 层次结构的上下文信息提出了一种基于 SVM 的集成分类器方法预测基因功能。

基因可以转录并翻译成各种蛋白质，执行许多不同的功能。目前基因本体被广泛应用于蛋白质的功能预测，即用基因功能来标注蛋白质功能，基因功能预测与蛋白质功能预测有着相似和密切关系。一些研究学者也尝试使用深度学习方法预测蛋白质功能。Kulmanov[429]提出了一个带注意力的深度分类器进行蛋白质功能预测。Gligorijevic[430]利用深度学习融合多个网络信息进行蛋白质功能预测。Cao[431]使用人工智能技术将功能预测转化为语言翻译问题，即蛋白质信息到基因本

体的信息的转换。对传统的机器学习方法而言，从大量复杂的数据中提取特征是一个艰难的挑战，不仅是因为一个基因或者蛋白质包含多个功能，而且复杂的生物过程和功能而且生物学过程不是由单个基因或者蛋白质独立完成，而是由多个基因或蛋白质协同作用。因此基因或蛋白质功能预测是一个多标签、多分类问题。这对传统的机器学习来说仍然是一个巨大的挑战。近年来深度学习得到了快速的发展，在处理大数据和提取特征具有显著效果。深度学习方法的分类性能随着数据量的增加结果而提高，而传统的机器学习方法则随着数据量的增加到一定程度并不会进一步提高性能。与传统的机器学习方法相比，深度学习可以提取更加抽象、非线性的复杂数据特征，进一步提高功能预测的精确度。综上所述，深度学习方法将成为进一步提高基因或蛋白质功能预测的一个关键技术。

7.5.3 可变剪接预测的人工智能算法

近年来用于可变剪接预测的人工智能算法有 spliceAI、DARTS 以及 Hui Y. Xiong 等人 2015 年在 Science 发表的一个机器学习计算模型。其中，spliceAI 在临床测序中由于难以解释而在很大程度上忽略了的非编码基因组的变异的识别，因此 Kishore Jaganathan 等人引入了一个深度学习网络，可以准确地预测主核苷酸序列的剪接，进一步识别破坏正常外显子和内含子模式和对生成蛋白质造成严重影响的非编码突变。DARTS 则解决了利用 RNA-Seq 进行可变剪接分析的一个主要限制——依赖于高测序覆盖率。Zijun Zhang 等人设计了一个集成基于深度学习和贝叶斯假设检验的计算框架 DARTS，利用公开

的大量 RNA-Seq 数据，提供剪接调控的知识库，帮助研究人员更好地利用 RNA-Seq 数据集描述可变剪接。Hui Y. Xiong 等人提出的机器学习计算模型对全基因组进行分析，可用于评估基因变异对 RNA 剪接的影响程度，其研究进一步了解了脊髓性肌萎缩症、遗传性非息肉病性结直肠癌和自闭症谱系障碍的遗传学基础提供了深入的了解。

7.6 调控基因组学

7.6.1 调控基因组概述

基因调控是指生物体内控制基因表达的机制，基本内容是调控蛋白与其靶 DNA 或 RNA 分子之间的相互作用，其主要发生于 DNA 水平上的调控、转录控制和翻译控制，微生物水平调控和多细胞生物的基因调控三种水平。微生物通过基因调控改变其代谢的方式以适应周围环境的变化，这类调控是短暂和可逆的。多细胞生物的基因调控是细胞分化、形态变化、个体发育的基础，这种调控一般是长期、不可逆的。1961 年 Jacob 和 Monod 提出著名的操纵子学说，基因调控的研究逐渐成为分子遗传学的重要内容。

调控基因组学研究具有重要的生物学意义。基因调控可以调节微生物的氨基酸、核苷酸之类物质的合成，受调控的菌种应用于发酵工业，可使产量大幅增加，例如植物苯丙氨酸解氨酶(PLA)应用于农作物生产。随着生产、生活环境的需要，植物抗病育种、生物活性成分成为现代生物技术焦点之一，能够提高植物抗性，减少经济损失和农药对人类健康及生态环境的影响，另外，其次生代谢产物富集在疾病的治疗和医疗方面具有不可替代的作用。除此之外，基因表达调控在

环境保护、食品工艺、养殖等很多行业中也有重要的应用。

7.6.2 基序检测的人工智能算法

识别基因组中转录因子结合位点或基序是破译基因调控机制的关键之一。基序是基因组中重复出现的序列模式，是转录因子的结合位点，对调节细胞内蛋白质的产生至关重要，基序分析对促进医学治疗和理解细胞过程具有重要意义[432]。转录因子结合位点可以通过几种高通量检测方法来测定，包括 PBM、SELEX、ChIP-和 CLIP-seq 技术。近年来，基序识别算法主要分为基于统计策略和基于人工智能学习两大类。基于统计的模型仍有较大的局限性，特别是对于识别序列组成和二级（或三级）结构组成的 RNA 结合蛋白。深度学习非常适用于基因组学，多个学习层可以在神经元内捕获多个级别的处理和抽象信息，DeepBind 和 DeepFinder 是其中两种应用较广的基序识别算法。

Alipanahi[433]等人将来自深度学习的最先进技术融入到 DeepBind 的开发中。该方法通过两个步骤预测蛋白质与 DNA 或 RNA 序列的结合亲和力，包括应用卷积模块进行表征学习和应用特征组合预测模块。Deepbind 从原始序列中捕获结合特异性，通过发现新的序列 motif 并将他们组合起来预测绑定分数。图形处理单元(GPU)用于自动训练高质量模型，不依赖于专家调优。DeepBind 的优点是能够自动选择模型参数和复杂度，深度学习在训练方式上的进步使得 DeepBind 更加实用。DeepBind 对挖掘更大的数据集非常有用，如 ENCODE 和 Roadmap Epigenomics。

Lee[434]等人提出的 DeepFinder 使用具有与结合位点相关特征的深度学习神经网络来构建基序识别模型，采用一种改进的三阶段 DNA 基序预测方法，这种方法具有两个新特征：第一，采用一组基序发现工具，用于从输入序列子集中初步预测候选结合位点。第二，提取与最具潜力的候选结合位点相关的特征，用于深度神经网络学习。DeepFinder 计算框架有三个连续的步骤：(1) 数据集划分为五个非重叠子集。(2) 四种新的基序发现工具应用于其中一个分区子集，以预测假定的基序和各自的结合位点，使用聚类算法对每个工具返回的前三个基元进行合并和分割，提取并集中与候选绑定位点相关的 76 个特征，并将其用于堆叠式自编码神经网络学习。(3) 学习神经网络用于预测相关的结合位点。

7.6.3 基因调控网络构建的人工智能算法

基因调控网络(GRNs)自基因表达数据产生以来，一直是生物信息学研究的热点。利用计算方法揭示基因表达调控的复杂结构是近几十年来出现的一项具有挑战性的任务。识别基因的相互作用有助于理解 GRNs 的拓扑结构和每个基因的作用，这对研究细胞在环境中行为背后的复杂机制至关重要[435]。

近年来提出了几种构建基因调控网络的无监督、半监督和监督学习方法。无监督模型主要分为布尔模型，贝叶斯模型，微分方程模型和信息理论模型四种类别。微分方程是最广泛使用的一类动力学模型，Jiguo 等人[436]提出的微分方差模型考虑了代谢物浓度随时间的变化。半监督学习有仅从样本中学习和从正样本和未标记数据中学习两

种方法。Patel 等人[437]提出了一种基于随机森林(RF)和 SVM 的迭代方法，通过自训练来预测每个转录因子的调节。Jisha 等人[438]使用聚类方法从未标记的数据中提取可靠的反例，提出的监督学习主要有 SVM 和深度神经网络。监督学习需要基因表达数据和已知的基因间调控，但比无监督和半监督方法更准确。Sirene[439]将基因调控网络推断的问题分解为大量的二元分类问题，每个子问题都与转录因子相关联，而 SVM 用于预测 GRN。compareSVM[440]可以用来比较线性，高斯，Sigmoid 和多项式内核四种 SVM 内核函数，其包括优化，比较和预测三个步骤。深度神经网络是一个受大脑神经网络启发的强大模型，其性能在广泛的应用中已经大幅度的提升。Mandal 等人[441]利用杂交杜鹃搜索—花传粉算法(FPA)训练一个递归神经网络来选择基因的最佳组合。

7.7 疾病基因预测

7.7.1 基因变异与复杂疾病

在人体中，基因在调控、突变和表观遗传等过程中的变化都可能引起疾病。识别疾病和基因之间的关系可以帮助我们理解疾病的机理，对疾病的诊断、药物的研发都有着重大的意义。根据引起疾病的基因的个数，已知基因疾病可以分为由单基因引起的孟德尔疾病和由多个基因协同引起的复杂疾病。孟德尔疾病因为其单基因的特性，识别其致病基因的关键通常是找到合适的研究对象，并通过连锁分析(linkage analysis)确定疾病对应的基因。复杂疾病则相反，他们的致病基因数量较多，同一种疾病在不同病人身上的 DNA 变异也不尽相

同。通常只需要一部分疾病基因的相互作用，就可能引发疾病[442]。因此，研究复杂疾病所对应的疾病基因是生物化学领域的一个重点。

7.7.2 疾病基因预测的主要方法

目前，复杂疾病所对应的疾病基因主要通过全基因组关联研究(Genome-wide association study, GWAS)来进行预测。GWAS 通过统计学的方法，研究已知的单核苷酸多态性(single nucleotide polymorphisms, SNP)在病人和正常人中发生的频率，找出疾病相关的 SNP。这些 SNP 所对应的基因则被当作可能的疾病基因，通过进一步的实验来验证其是否与疾病相关。一次 GWAS 实验通常会检测超过一百万个 SNP，最终得到几百个疾病相关的 SNP。虽然 GWAS 已经极大的缩小了验证的范围，但进一步验证这些 SNP 所对应的基因仍需要耗费大量的时间和费用。同时，并不是所有的疾病相关的突变都可以通过 GWAS 找到，例如结构变异和上位性。因此，设计算法分析 GWAS 数据或是通过分析其他疾病相关的数据直接预测疾病基因，对识别疾病基因具有重大价值。

7.7.3 疾病基因预测的人工智能算法

目前，与人工智能相关的算法已经被广泛的应用于疾病基因的预测。这些算法将疾病基因预测的问题转化为一个二分类的问题，通过机器学习的方法，将疾病基因和非疾病基因进行分类。类似于其他分类问题，无监督学习、有监督学习和半监督学习都被用于疾病基因的预测。

无监督学习的方法主要通过聚类算法在生物分子网络中找出与

疾病密切相关的类（子网络），并将类中的基因预测为疾病基因。通常在聚类前，与疾病相关的数据例如 GWAS 数据和基因互表达数据会被用来对生物分子网络加权。聚类的目标则是在加权的网络中找出携带最多疾病相关信息的子网络。最常用的聚类算法是 Seed-growth 算法，该方法从单个节点出发，依次将蛋白质相互作用(protein-protein interaction, PPI)网络中的节点加入到类中，最终找到目标子网络。例如，EW_dmGWAS 首先使用 GWAS 数据和基因表达数据对 PPI 网络进行加权，之后通过 Seed-growth 算法，找出具有最小 p 值和最大差异互表达的子网络，并将其中的基因预测为疾病基因[443]。除了 Seed-growth 算法外，也可以通过其他聚类方法来得到疾病相关的子网络。例如，Wu 等人使用马尔科夫聚类在基因表达数据加权的 PPI 网络上进行聚类，找到和癌症相关的疾病基因[444]。

无监督学习的算法不需要已知疾病基因的信息，就可以对疾病基因进行预测，在早期疾病基因的预测以及 GWAS 数据的分析中起到了重要的作用。但由于未使用已知的疾病基因信息，无监督学习的准确率较低。同时，预测出的疾病基因彼此间没有优先级，增加了后续验证的成本。随着已被验证的疾病基因数量的上升，更多研究者开始使用半监督学习和有监督学习的方式对疾病基因进行预测。

考虑到疾病基因预测是一个 positive unlabeled learning 问题，在已知一部分疾病基因的前提下，我们既可以使用半监督学习的方法对疾病基因进行预测，例如使用贝叶斯分类器来判断未知基因是否为疾病基因[445]或是使用 Manifold learning 的方法预测疾病基因[446]，也可

以定义一部分非疾病基因（负样本），利用有监督学习的方式进行预测。定义非疾病基因的方法较多，常见方法是随机选择一部分未知的基因作为非疾病基因，并通过 bootstrapping 的方式选择多组负样本，根据多组数据中得到的平均结果对疾病基因进行预测[447]。虽然随机选择的负样本中有概率包含未知的疾病基因，但是考虑到疾病基因的个数远小于未知基因的个数，结合 bootstrapping 依旧可以获得相对准确的预测。除了随机选择非疾病基因外，一些研究也根据疾病之间的相似性或是特征向量之间的欧式距离来定义非疾病基因[448, 449]。

目前，很多分类器都被用于预测疾病基因，例如逻辑回归，支持向量机和随机树等经典算法。虽然在不同的样本容量和特征下，各个分类器的准确率有所不同，但提升预测准确率的关键还是提取出具有区分度的特征。现有的方法已经从不同的生物数据中为基因提取特征，包括 PPI 网络、基因表达数据、体细胞突变数据、文本数据、Gene Ontology 数据等。通常，为了提升预测准确率，我们希望将不同数据进行融合后提取特征。PPI 网络由于其自身的特性，常以加权的方式和各类数据融合。融合后的网络包含各种生物信息，直接使用中心性等网络结构指标就可提取出有价值的特征。另外，也可使用网络特征提取算法提取有用特征，例如利用 node2vec 从 PPI 网络中提取特征来预测疾病基因[450, 451]。此外，考虑到 PPI 网络的动态性以及现有 PPI 网络质量较低的特点，一些研究也使用基因表达数据来优化 PPI 网络，再从优化后的网络中提取特征，例如 Guan 等人通过构建样本专属的网路来为预测疾病基因[452]。除此之外，亦可根据 PPI 网络和其

他数据构造 metagraph，将多种数据的信息同时包含在一个网络中。网络的节点既可以是基因，也可以是基因相关的各种信息，适用于 metagraph 的特征提取方法可以被用来为基因提取特征，从而实现疾病基因的预测[453]。除了基于 PPI 网络的数据融合外，另一种常见的方法是将不同数据中提取的特征连接在一起，同时作为基因的特征来进行模型的训练。通常连接后的特征向量维度较高，研究者会使用特征提取的方式选择最具有区分度的特征，并将他们用在最终的预测中，例如 PUDI 根据 GO 和蛋白质域的数据提取出 47780 个特征后，使用特征提取选择出 4000 个特征并将这些特征用于支持向量机的训练[449]。

不论是基于 PPI 网络的加权还是特征连接的数据融合，都是对不同数据的线性融合，不同数据间的非线性关系无法使用这些方式学习出来。为了解决这个问题，一些研究通过深度学习来融合不同种类的数据，使用非线性的激活函数学习不同数据间的非线性关系，进而完成数据的融合，例如 Luo 等人使用深度置信网络(Deep Belief Net, DBN)融合从 PPI 网络和 GO 中提取的原始特征，并将融合后的特征用于疾病基因的预测。与 DBN 类似，深度自编码器(Deep Autoencode)也可以用来学习生物网络中的非线性特征[454]。

以上介绍了一些疾病基因预测的人工智能算法。随着人工智能领域的发展，新的算法可以帮助我们更好的融合不同种类的生物数据，提升疾病基因预测的准确率。相信在不久的将来，人工智能可以帮助我们更好的了解疾病和基因之间的关系，加速新药物的研发，促进医学发展。

7.8 人工智能在基因组分析中的发展前景

近年来，深度学习(Deep Learning, DL)开始逐渐应用到基因组学研究中，并在变异识别、转录调控、可变剪接等重要研究方向取得了成功的应用。端到端学习(end-to-end learning)由于能够将多个预处理步骤集成到模型中并能自动提取特征，成为了深度学习的一大优势。深度学习的另一优势在于能够有效处理多模态数据，非常适合对基因组学中的序列、基因表达、质谱、影像等多元异质数据进行整合分析，从而实现重要生物信息的提取。特别地，在影像数据分析方面，深度学习具有提取空间模式特征的能力，这也是其在图像分析领域取得了成功应用的重要因素之一。在模型开发方面，包括 TensorFlow、Keras、Pytorch 在内的框架已非常成熟且高度模块化，使得研究人员能够很容易开发出深度学习模型，大大的促进了深度学习的应用。

尽管深度学习成功应用于基因组分析，但其在基因组领域仍处于初步的阶段，仍然面临着诸多挑战。目前人类基因组学的一个特殊挑战是数据隐私，如何利用或构建新的人工智能方法，通过部署在不同的站点上的且共享公共参数站的模型同时训练本地数据，是保护遗传和医疗数据隐私的一种思路。另一个重要挑战是预测因果关联，相关的进展将极大地促进生物医学的发展。深度学习模型不易分析，难以理解的特性也让从模型中理解数据机制成为了挑战。

深度学习已成为基因组学数据分析的有力工具。我们挖掘深度学习工具理解基因组生物学的潜力，并期望深度学习领域会开发出越来越多的新技术，扩展到基因组学研究的各个方面，推动基因组分析的进步。

参考文献

[1] Kung JT, Colognori D, Lee JT. Long noncoding RNAs: past, present, and Future. Genetics, 2013, 193 (3): 651-669.

[2] Chen LL. Linking long noncoding RNA localization and function. Trends in Biochemical Sciences, 2016, 41(9): 761-772.

[3] Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. Science, 2005, 309(5740): 1559-63.

[4] Pasquinelli AE, Hunter S, Bracht J. Micrornas: a developing story. Current Opinion in Genetics and Development, 2005, 15(2): 200-5.

[5] Bentwich I. Prediction and validation of micrornas and their targets. Febs Letters 579, 2005 579(26): 5904-10.

[6] Hertel J, Stadler PF. Hairpins in a haystack: recognizing microrna precursors in comparative genomics data. Bioinformatics, 22(14): e197-202.

[7] Huang TH, Fan B, Rothschild MF, et al. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics, 2007, 8(1): 341.

[8] Nam JW, Shin KR, Han J, et al. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research, 2005, 33(11): 3570-3581.

[9] Sun L, Luo H, Bu D, Zhao G, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Research, 41(17): e166.

[10] Achawanantakun R, Chen J, Sun Y, et al. Lncrna-id: long non-coding RNA identification using balanced random forests. Bioinformatics, 2015, 31(24): 3897-905.

[11] Frith MC, Forrest AR, Nourbakhsh E, et al. The abundance of short proteins in the mammalian proteome. PLoS Genetics, 2006, 2(4): e52.

[12] Pian C, Zhang G, Chen Z, et al. LncRNApred: classification of long

non-coding RNAs and protein-coding transcripts by the ensemble algorithm with a new hybrid feature. PloS one, 2016, 11(5): e0154567.

[13] Yang C, Yang L, Zhou M, et al. Lncadeep: An ab initio lncrna identification and functional annotation tool based on deep learning. Bioinformatics, 2018, 34(22): 3825-3834.

[14] Kanehisa M, Goto S, Furumichi M, et al. Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research, 2010, 38(Database issue): D355-60.

[15] Croft D, Mundo AF, Haw R, et al. The reactome pathway knowledgebase. Nucleic Acids Research, 2014, 42: D472-D477.

[16] Deshpande S, Shuttleworth J, Yang J, et al. Plit: an alignment-free computational tool for identification of long non-coding rnas in plant transcriptomic datasets. Computers in Biology and Medicine, 2019, 105: 169-181.

[17] Agarwal V, Bell GW, Nam JW, et al. Predicting effective microRNA target sites in mammalian mRNAs. Elife, 2015, doi: 10.7554/eLife.05005.

[18] Betel D, Wilson M, Gabow A, et al. The microRNA.org resource: targets and expression. Nucleic Acids Research, 2008, 36(Database issue): D149-53.

[19] Kertesz M, Iovino N, Unnerstall U, et al., The role of site accessibility in microRNA target recognition. Nature Genetics. 2007, 39(10): 1278-84.

[20] Majoros WH, Lekprasert P, Mukherjee N, et al. MicroRNA target site identification by integrating sequence and binding information. Nature Methods, 2013, 10(7): 630-3.

[21] Erhard F, Dölken L, Jaskiewicz L, et al. PARma: identification of microRNA target sites in AGO-PAR-CLIP data. Genome Biology, 2013, 14(7): R79.

[22] Khorshid M, Hausser J, Zavolan M, et al. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. Nature Methods, 2013, 10(3): 253-5.

[23] Wang KC. Molecular mechanisms of long noncoding RNAs. Molecular cell. 2011, 43(6): 904-14.

[24] Li JH, Liu S, Zhou H, et al. StarBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Research. 2014, 42(Database issue): D92-7.

[25] Zhou Z, Shen Y, Khan MR, et al. LncReg: a reference resource for lncRNA-associated regulatory networks. Database (Oxford), 2015, 2015, pii: bav083.

[26] Jiang Q. LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. Nucleic Acids Research, 2015, 43(Database issue): D193-6.

[27] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, et al. ViennaRNA package 2.0. algorithms for molecular biology. Algorithms for Molecular Biology, 2011, 6:26.

[28] Chan C Y, Lawrence C E, Ding Y. Structure clustering features on the Sfold Web server. Bioinformatics, 2005, 21(20): 3926-3928.

[29] Wang L, Liu Y, Zhong X, et al. DMfold: a novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. Frontiers in Genetics, 2019, 10:143.

[30] Ren J, Rastegari B, Hoos HH. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. Rna-a Publication of the Rna Society, 2005, 11(10): 1494-504.

[31] Lathauwer LD, Moor BD, Vandewalle J. A multilinear singular value decomposition. Society for Industrial and Applied Mathematics, 2000.

[32] Zhan ZH, Jia LN, Zhou Y, et al. BGFE: a deep learning model for ncRNA-protein interaction predictions based on improved sequence information. International Journal of Molecular Sciences, 2019, 20(4): 978.

[33] Pan X, Fan YX, Yan J, et al. IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC genomics, 2016, 17(1): 582.

[34] Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature

biotechnology, 2015, 33(8): 831.

[35] Wang Y, Chen X, Liu ZP, et al. De novo prediction of RNA-protein interactions from sequence information. Molecular BioSystems, 2013, 9(1): 133-142.

[36] Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of molecular biology, 1981. 147(1): 195-197.

[37] Altschul S. Basic local alignment search tool. Journal of Molecular Biology, 1990, 215(3): 403-410.

[38] Liao W, Ren J, Wang K, et al. Alignment-free transcriptomic and metatranscriptomic comparison using sequencing signatures with variable length markov chains. Scientific Reports, 2016, 6: 37243.

[39] Zielezinski A, Vinga S, Almeida J, et al. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biologyogy, 2017. 18(1): 186.

[40] Segata N1, Waldron L, Ballarini A, et al., Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods, 2012. 9(8): 811-814.

[41] Leimena MM, Ramiro-Garcia J, Davids M, et al. A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics, 2013, 14: 530.

[42] Marchetti A, Schruth DM, Durkin CA, et al. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. Proceedings of the National Academy of Sciences, 2012, 109(6): E317-25.

[43] Karlin S, Mrázek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. Journal of Bacteriology, 1997, 179(12): 3899-913.

[44] Ren J. Alignment-free sequence analysis and applications. Annual Review of Biomedical Data Science, 2018, 1: 93-114.

[45] Song K, Ren J, Reinert G, et al. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings in

Bioinformatics, 2013, 15(3): 343-353.

[46] Qi J, Luo H, Hao B. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Researchearch, 2004, 32(suppl_2): W45-7.

[47] Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Researchearch, 2009, 37(suppl_2): W174-8.

[48] Zuo G1, Hao B. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. Genomics, Proteomics & Bioinformatics, 2015, 13(5):321-31.

[49] Reinert G, Chew D, Sun F, et al. Alignment-free sequence comparison (I): statistics and power. Journal of Computational Biology, 2009, 16(12): 1615-34.

[50] Wan L, Reinert G, Sun F, et al. Alignment-free sequence comparison (II): theoretical power of comparison statistics. Journal of Computational Biology, 2010, 17(11): 1467-90.

[51] Zuo G1, Xu Z, Hao B. Shigella strains are not clones of escherichia coli but sister species in the genus Escherichia. Genomics, Proteomics & Bioinformatics, 2013. 11(1): 61-5.

[52] Li Q, Xu Z, Hao B. Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations. Journal of Biotechnology, 2010, 149(3): 115-9.

[53] Zuo G, Qi J, Hao B. Polyphyly in 16S rRNA-based LVTree versus monophyly in whole-genome-based CVTree. Genomics, Proteomics & Bioinformatics, 2018, 16(5): 310-319.

[54] Ren J, Song K, Deng M, et al., Inference of Markovian properties of molecular sequences from NGS data and applications to comparative genomics. Bioinformatics, 2016. 32(7): 993-1000.

[55] Jiang B, Song K, Ren J, et al. Comparison of metagenomic samples using sequence signatures. BMC Genomics, 2012, 13(1): 730.

[56] Wang Y, Liu L, Chen L, et al. Comparison of metatranscriptomic samples based on k-tuple frequencies. PloS One, 2014, 9: e84348.

[57] Ahlgren NA, Ren J, Lu YY, et al. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Researchearch, 2016, 45(1): 39-53.

[58] Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biologyogy, 2016, 17(1): 132.

[59] Indyk P, Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, 1998, ACM.

[60] Murray KD, Webers C, Ong CS, et al. KWIP: the k-mer weighted inner product, a de novo estimator of genetic similarity. PLoS Computational Biology, 2017, 13(9): e1005727.

[61] Sarmashghi S, Bohmann K, P Gilbert MT, et al. Skmer: assembly-free and alignment-free sample identification using genome skims. Genome Biologyogy, 2019, 20(1): 34.

[62] Zielezinski A, Girgis HZ, Bernard G, et al. Benchmarking of alignment-free sequence comparison methods. BioRxiv, 2019, 20(1):144.

[63] Zheng W, Yang L, Genco RJ, et al. SENSE: Siamese neural network for sequence embedding and alignment-free comparison. Bioinformatics, 2018, 35(11): 1820-1828.

[64] Giancarlo R, Rombo SE, Utro F. Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies. Briefings in Bioinformatics, 2013, 15(3): 390-406.

[65] Wang Z, Wang Y, Fuhrman JA, et al. Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences. Briefings in Bioinformatics, 2019, pii: bbz025.

[66] Pop M. Genome assembly reborn: recent computational challenges. Briefings in Bioinformatics, 2009, 10(4): 354-366.

[67] Myers EW. A whole-genome assembly of Drosophila. Science, 2000, 287(5461): 2196-2204.

[68] Batzoglou S, Jaffe DB, Stanley K, et al. ARACHNE: a whole-genome shotgun assembler. Genome Research, 2002. 12(1): 177-189.

[69] Nurk S, Meleshko D, Korobeynikov A, et al. metaSPAdes: a new versatile metagenomic assembler. Genome Research, 2017, 27(5): 824-834.

[70] Peng Y, Leung HC, Yiu SM, et al. Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics, 2011, 27(13): i94-101.

[71] Namiki T, Hachiya T, Tanaka H, et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Researchearch, 2012, 40(20): e155-e155.

[72] Kislyuk A, Bhatnagar S, Dushoff J, et al. Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics, 2009. 10(1): 316.

[73] Kelley DR, Salzberg SL. Clustering metagenomic sequences with interpolated Markov models. BMC Bioinformatics, 2010. 11(1): 544.

[74] Wang Y, Wang K, Lu YY, et al. Improving contig binning of metagenomic data using oligonucleotide frequency dissimilarity. Bmc Bioinformatics, 2017, 18(1): 425.

[75] Brown CT, Sharon I, Thomas BC, et al., Genome resolved analysis of a premature infant gut microbial community reveals a Varibaculum cambriense genome and a shift towards fermentation-based metabolism during the third week of life. Microbiome, 2013, 1(1): 30.

[76] Wrighton KC, Thomas BC, Sharon I, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science, 2012, 337(6102): 1661-1665.

[77] Abe T, Hamano Y, Ikemura T. Visualization of genome signatures of eukaryote genomes by batch-learning self-organizing map with a special emphasis on drosophila genomes. BioMed Research International, 2014, 2014: 985706.

[78] Laczny CC, Sternal T, Plugaru V, et al. VizBin-an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome, 2015, 3(1): 1.

[79] Wang Y, Hu H, Li X. MBBC: an efficient approach for metagenomic binning

based on clustering. BMC Bioinformatics, 2015, 16(1): 36.

[80] Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. Journal of Computational Biology, 2011, 18(3): 523-534.

[81] Nielsen HB, Almeida M, Juncker AS, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature biotechnology, 2014, 32(8): 822.

[82] Alneberg J, Bjarnason BS, de Bruijn I, et al. Binning metagenomic contigs by coverage and composition. Nature Methods, 2014, 11: 1144-1146.

[83] Lu YY, Chen T, Fuhrman JA, et al., COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment, and paired-end read LinkAge. Bioinformatics, 2017, 33(6): 791-798.

[84] Kang DD, Froula J, Egan R, et al. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ, 2015, 3: e1165.

[85] Nissen JN. Binning microbial genomes using deep learning. BioRxiv, 2018: 490078.

[86] Ren, J., et al., Identifying viruses from metagenomic data by deep learning. arXiv preprint arXiv:1806.07810, 2018.

[87] Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Computational Biology, 2012. 8(9): e1002687.

[88] Deng Y, Jiang YH, Yang Y, et al. Molecular ecological network analyses. BMC Bioinformatics, 2012, 13(1): 113.

[89] Faust K, Sathirapongsasuti JF, Izard J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Computational Biology, 2012. 8(7): e1002606.`

[90] Kelley DR, Liu B, Delcher AL, et al. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Researchearch, 2011. 40(1): e9.

[91] Keegan KP, Glass EM, Meyer F. MG-RAST, a metagenomics service for

analysis of microbial community structure and function. Methods in Molecular Biology, 2016, 1399: 207-33.

[92] Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Researchearch, 2010. 38(20): e191-e191.

[93] Al-Ajlan A1, El Allali A. CNN-MGP: convolutional neural networks for metagenomics gene prediction. Interdisciplinary Sciences: Computational Life Sciences, 2018: 1-8.

[94] Arango-Argoty G, Garner E, Pruden A, et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome, 2018, 6(1): 23.

[95] Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 2012, 490(7418): 55.

[96] Wang Y, Fu L, Ren J, et al. Identifying Group-Specific Sequences for Microbial Communities Using Long k-mer Sequence Signatures. Frontiers in Microbiology, 2018, 9:872.

[97] Fioravanti D, Giarratano Y, Maggio V, et al. Phylogenetic convolutional neural networks in metagenomics. BMC Bioinformatics, 2018, 19(2): 49.

[98] Galkin F. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. BioRxiv, 2018: 507780.

[99] Proctor, LM, et al. The Integrative Human Microbiome Project. Nature, 2019, 569(7758): 641-648.

[100] Serrano MG, Parikh HI, Brooks JP, et al. Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. Nature Medicine, 2019, 25(6): 1001-1011.

[101] Fettweis, J.M. The vaginal microbiome and preterm birth. Nature Medicine, 2019.

[102] Lloyd-Price J, Arze C, Ananthakrishnan AN, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature, 2019, 569(7758): 655-662.

[103] Franzosa EA, Sirota-Madi A, Avila-Pacheco J, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature Microbiology, 2019, 4(2): 293-305.

[104] Zhou W, Sailani MR, Contrepois K, et al. Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature, 2019, 569(7758): 663-671.

[105] Schüssler-Fiorenza Rose SM, Contrepois K, Moneghetti K, et al. A longitudinal big data approach for precision health. Nature Medicine, 2019, 25(5): 792-804.

[106] Qin J, Li R, Raes J, Arumugam M, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2009, 464: 59-65.

[107] Qin N, Yang F, Li A, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature, 2014. 513(7516): 59.

[108] Strandwitz P, Kim KH, Terekhova D, et al. GABA-modulating bacteria of the human gut microbiota. Nature Microbiology, 2019, 4(3): 396.

[109] Strati F, et al. New evidences on the altered gut microbiota in autism spectrum disorders. Microbiome, 2017, 5(1): 24.

[110] Hill-Burns EM, et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. Movement Disorders, 2017, 32(5): 739-749.

[111] Vogt NM, et al. Gut microbiome alterations in Alzheimer’s disease. Scientific Reports, 2017, 7(1): 13537.

[112] Zhang X, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nature Medicine, 2015, 21(8): 895.

[113] Zackular JP, et al. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prevention Research, 2014. 7(11): 1112-1121.

[114] Arthur JC, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science, 2012, 338(6103): 120-123.

[115] Mao Q, et al. Interplay between the lung microbiome and lung cancer. Cancer letters, 2018, 415: 40-48.

[116] Alexander JL, et al. Gut microbiota modulation of chemotherapy efficacy and toxicity. Nature Reviews Gastroenterology & Hepatology, 2017, 14(6): 356.

[117] Guglielmi G. How gut microbes are joining the fight against cancer. Nature, 2018, 557(7706): 482-484.

[118] Routy B, et al. Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. Science, 2018, 359(6371): 91-97.

[119] Surawicz CM, et al. Guidelines for diagnosis, treatment, and prevention of Clostridium difficile infections. The American Journal of Gastroenterology, 2013, 108(4): 478.

[120] Cui B, et al. Step-up fecal microbiota transplantation strategy: a pilot study for steroid-dependent ulcerative colitis. Journal of translational medicine, 2015, 13(1): 298.

[121] Rossen NG, et al. Findings from a randomized controlled trial of fecal transplantation for patients with ulcerative colitis. Gastroenterology, 2015, 149(1): 110-118.

[122] Bak SH, et al. Fecal microbiota transplantation for refractory Crohn's disease. Intestinal Research, 2017, 15(2): 244.

[123] Tian H, et al. Treatment of slow transit constipation with fecal microbiota transplantation. Journal of Clinical Gastroenterology, 2016, 50(10): 865-870.

[124] Johnsen PH, et al. Faecal microbiota transplantation versus placebo for moderate-to-severe irritable bowel syndrome: a double-blind, randomised, placebo-controlled, parallel-group, single-centre trial. The Lancet Gastroenterology & Hepatology, 2018, 3(1): 17-24.

[125] Ren YD, et al. Fecal microbiota transplantation induces hepatitis B virus e-antigen (HBeAg) clearance in patients with positive HBeAg after long-term antiviral therapy. Hepatology, 2017, 65(5): 1765-1768.

[126] Kang DW, et al. Microbiota Transfer Therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. Microbiome, 2017, 5(1): 10.

[127] Nealson KH, Venter JC. Metagenomics and the global ocean survey: what's in

it for us, and why should we care? The ISME Journal, 2007, 1(3): 185.

[128] Han M, et al. Agricultural risk factors influence microbial ecology in Honghu Lake. Genomics, Proteomics & Bioinformatics, 2019, 17(1): 76-90.

[129] Bahram M, et al. Structure and function of the global topsoil microbiome. Nature, 2018, 560(7717): 233.

[130] Crits-Christoph A, et al. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature, 2018, 558(7710): 440.

[131] Glassman SI, et al. Decomposition responses to climate depend on microbial community composition. Proceedings of the National Academy of Sciences, 2018, 115(47): 11994-11999.

[132] Suzuki S, et al. Unusual metabolic diversity of hyperalkaliphilic microbial communities associated with subterranean serpentinization at The Cedars. The ISME Journal, 2017, 11(11): 2584.

[133] 白洋, et al. 农作物微生物组: 跨越转化临界点的现代生物技术. 中国科学院院刊, 2017, 32(3): 260-265.

[134] Zhang J, et al. Root microbiota shift in rice correlates with resident time in the field and developmental stage. Science China Life Sciences, 2018: 1-9.

[135] Xu J, et al. The structure and function of the global citrus rhizosphere microbiome. Nature Communications, 2018, 9(1): 4894.

[136] Garrido-Oter R, et al. Modular traits of the rhizobiales root microbiota and their evolutionary relationship with symbiotic rhizobia. Cell Host & Microbe, 2018, 24(1): 155-167.

[137] Müller DB, et al. The plant microbiota: systems-level insights and perspectives. Annual Review of Genetics, 2016, 50: 211-234.

[138] Sobel J, et al. BeerDeCoded: the open beer metagenome project. F1000Research, 2017 6:1676.

[139] Jünemann S, et al. Bioinformatics for NGS-based metagenomics and the application to biogas research. Journal of Biotechnology, 2017, 261: 10-23.

[140] Bashir Y, Pradeep Singh S, Kumar Konwar B. Metagenomics: an application based perspective. Chinese Journal of Biology, 2014, 2014.

[141] Ramírez C, Romero J. The microbiome of Seriola lalandi of wild and aquaculture origin reveals differences in composition and potential function. Frontiers in Microbiology, 2017, 8: 1844.

[142] Zeng S, et al. Composition, diversity and function of intestinal microbiota in pacific white shrimp (Litopenaeus vannamei) at different culture stages. PeerJ, 2017, 5: e3986.

[143] Gobet A, et al. Seasonal and algal diet-driven patterns of the digestive microbiota of the European abalone Haliotis tuberculata, a generalist marine herbivore. Microbiome, 2018, 6(1): 60.

[144] Ahmed, et al. “Distributed Natural Large Scale Graph Factorization,” In Proceedings of the 22nd International Conference on World Wide Web, 2013, 37-48.

[145] Grover, Leskovec J. “node2vec: Scalable feature learning for networks,” In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 855-864.

[146] Micheli. “Neural network for graphs: A contextual constructive approach.” IEEE Transactions on Neural Networks, 2009,20(3): 498–511.

[147] Perozzi R, et al. “Deepwalk: Online learning of social representations,” In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, 701-710.

[148] Scarselli F, et al. “The graph neural network model.” IEEE Transactions on Neural Networks, 2009,20(1): 60-80.

[149] Dai H, et al. “Learning steady-states of iterative algorithms over graphs.” In Proceedings of the International Conference on Machine Learning, 2018, 1114-1122.

[150] Ribeiro LFR, et al. “struc2vec: Learning node representations from structural identity.” In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, 385-394.

[151] Bruna J, et al. “Spectral networks and locally connected networks on graphs.” In Proceedings of International Conference on Learning Representations, 2014.

[152] Lee JB, et al. “Graph classification using structural attention.” In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018, 1666-1674.

[153] Agrawal M, et al. “Large-scale analysis of disease pathways in the human interactome,” In Pacific Symposium on Biocomputing, 2018, 23: 111.

[154] Defferrard M, et al. “Convolutional neural networks on graphs with fast localized spectral filtering.” In Advances in Neural Information Processing Systems, 2016, 3844-3852.

[155] Henaff M, et al. “Deep convolutional networks on graph-structured data.” arXiv preprint arXiv:1506.05163, 2015.

[156] Simonovsky M, Komodakis N. “Dynamic edgeconditioned filters in convolutional neural networks on graphs.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[157] Zhang M, Cui Z, Neumann M, et al. “An end-to-end deep learning architecture for graph classification.” In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[158] Zitnik M, Leskovec J. “Predicting multicellular function through multi-layer tissue networks.” Bioinformatics, 2017, 33(14), i190-i198.

[159] Zitnik M, Agrawal M, Leskovec J. “Modeling polypharmacy side effects with graph convolutional networks.” Bioinformatics, 2018, 34(13), i457-i466.

[160] Langfelder P, Horvath S. “WGCNA: an R package for weighted correlation network analysis.” BMC Bioinformatics, 2008, 9(559).

[161] Levie R, Monti F, Bresson X, et al. “Cayleynets: Graph convolutional neural networks with complex rational spectral filters,” IEEE Transactions on Signal Processing, 2017, 67(1): 97-109.

[162] Kearnes S, McCloskey K, Berndl M, et al. “Molecular graph convolutions: moving beyond fingerprints,” Journal of Computer-aided Molecular Design, 2016, 30(8): 595-608.

[163] S. Rhee, S. Seo, and S. Kim, “Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype

classification," 2017, arXiv preprint arXiv:1711.05859.

[164] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proceedings of the International Conference on Learning Representations, 2017.

[165] W. Huang, T. Zhang, Y. Rong, and J. Huang, "Adaptive sampling towards fast graph representation learning," in Advances in Neural Information Processing Systems, 2018, pp. 4563–4572.

[166] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 135-144.

[167] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in Proceedings of the International Conference on Learning Representations, 2015.

[168] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in Advances in Neural Information Processing Systems, 2018, pp. 4801–4811.

[169] Venter J C, Adams M D, Myers E W, et al. The sequence of the human genome. science, 2001, 291 (5507): 1304-1351.

[170] Consortium I H G S. Initial sequencing and analysis of the human genome. Nature, 2001, 409 (6822): 860.

[171] Collins F S, Patrinos A., Jordan E., et al. New goals for the U.S. Human Genome Project: 1998-2003. Science, 1998, 282 (5389): 682-689.

[172] Butcher S P. Target discovery and validation in the post-genomic era. Neurochemical Research, 2003, 28 (2): 367-371.

[173] Dawid I B, Wahli W. Application of recombinant DNA technology to questions of developmental biology: A review. Developmental Biology, 1979, 69 (1): 305-328.

[174] Hsu P D, Lander E S, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell, 2014, 157 (6): 1262-1278.

[175] Beerli R R, Schopfer U., Dreier B., et al. Chemically regulated zinc finger transcription factors. Journal of Biological Chemistry, 2000, 275 (42): 32617-32627.

[176] Marina B, Mary G, Golic K G, et al. Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. Genetics, 2002, 161 (3): 1169-1175.

[177] Marina B, Kelly B, Trautman J K, et al. Enhancing gene targeting with designed zinc finger nucleases. Science, 2003, 300 (5620): 764-764.

[178] Urnov F D, Miller J C, Ya-Li L, et al. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature, 2005, 435 (7042): 646-651.

[179] Beerli R R, Dreier B., Barbas C F. Positive and negative regulation of endogenous genes by designed transcription factors. Proceedings of the National Academy of Sciences of the United States of America, 2000, 97 (4): 1495-1500.

[180] Moscou M J, Bogdanove A J. A simple cipher governs DNA recognition by TAL effectors. Science, 2009, 326 (5959): 1501-1501.

[181] Jens B, Heidi S, Sebastian S, et al. Breaking the code of DNA binding specificity of TAL-type III effectors. Science, 2010, 326 (5): 1509-1432.

[182] Miller J C, Tan S, Qiao G, et al. A TALE nuclease architecture for efficient genome editing. Nature biotechnology, 2010, 29 (2): 143.

[183] Sander J D, Dahlborg E J, Goodwin M J, et al. Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). Nature methods, 2010, 8 (1):67.

[184] Alexandre J, Gwendoline D, Julien V, et al. Comprehensive analysis of the specificity of transcription activator-like effector nucleases. Nucleic Acids Researchearch, 2014, 42 (8): 5390.

[185] Doudna J A, Emmanuelle C. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science, 2014, 346 (6213): 1258096.

[186] Hyongbum K, Jin-Soo K. A guide to genome engineering with programmable

nucleases. Nature Reviews Genetics, 2014, 15 (5): 321-334.

[187] 杨发誉, 葛香连, 谷峰. 新型靶向基因组编辑技术研究进展. 中国生物工程杂志, 2014, 34 (02): 98-103.

[188] 陈永昌, 牛昱宇, 季维智. 通过 CRISPR/Cas9 和 TALENs 介导的基因打靶技术获得基因修饰的猴模型. 中国细胞生物学学报, 2014, 36 (05): 557-560.

[189] Sander J D, J Keith J. CRISPR-Cas systems for editing, regulating and targeting genomes. Nature Biotechnology, 2014, 32 (4): 347-355.

[190] Graham D B, Root D E. Resources for the design of CRISPR gene editing experiments. Genome Biologyogy, 2015, 16 (1): 260.

[191] Hartenian E, Doench J G. Genetic screens and functional genomics using CRISPR/Cas9 technology. Febs Journal, 2015, 282 (8): 1383-1393.

[192] Ishino Y, Shinagawa H, Makino K, et al. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. Journal of bacteriology, 1987, 169 (12): 5429-5433.

[193] Rodolphe B, Christophe F, Hélène D, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science, 2007, 315 (5819): 1709-1712.

[194] Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science, 2012, 337 (6096): 816-821.

[195] Mojica F, Diez-Villasenor C, Garcia-Martinez J, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of Molecular Evolution, 2005, 60 (2): 174-182.

[196] Pourcel C., Salvignol G., Vergnaud G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology, 2005, 151 (3): 653-663.

[197] Mojica F J, Díez-Villaseñor C, Soria E, et al. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and

mitochondria. Molecular microbiology, 2000, 36 (1): 244-246.

[198] Jansen R, Embden J D V, Gaastra W, et al. Identification of genes that are associated with DNA repeats in prokaryotes. Molecular microbiology, 2002, 43 (6): 1565-1575.

[199] Haft D H, Selengut J, Mongodin E F, et al. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS computational biology, 2005, 1 (6): e60.

[200] Bolotin A, Quinquis B, Sorokin A, et al. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology, 2005, 151 (8): 2551-2561.

[201] Brouns S J J, Jore M M, Magnus L, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science, 2008, 321 (5891): 960-964.

[202] Marraffini L A, Sontheimer E J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science, 2008, 322 (5909): 1843-1845.

[203] Hale C R, Zhao P, Olson S, et al. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell, 2009, 139 (5): 945-956.

[204] Deveau H, Barrangou R, Garneau J E, et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of bacteriology, 2008, 190 (4): 1390-1400.

[205] Garneau J E, Marie-Ève D, Manuela V, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature, 2010, 468 (7320): 67.

[206] Deltcheva E, Chylinski K, Sharma C M, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature, 2011, 471 (7340): 602.

[207] Rimantas S, Giedrius G, Christophe F, et al. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Researchearch, 2011, 39 (21): 9275-9282.

[208] Giedrius G, Rodolphe B, Philippe H, et al. Cas9-crRNA ribonucleoprotein

complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109 (39): 15539-15540.

[209] Cong L, Ran F A, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. Science, 2013, 339 (6121): 819-823.

[210] Mali P, Yang L, Esvelt K M, et al. RNA-guided human genome engineering via Cas9. Science, 2013, 339 (6121): 823-826.

[211] Sander J D, Joung J K. CRISPR-Cas systems for editing, regulating and targeting genomes. Nature biotechnology, 2014, 32 (4): 347.

[212] Makarova K S, Wolf Y I, Alkhnbashi O S, et al. An updated evolutionary classification of CRISPR–Cas systems. Nature Reviews Microbiology, 2015, 13 (11): 722.

[213] Heler R, Samai P, Modell J W, et al. Cas9 specifies functional viral targets during CRISPR–Cas adaptation. Nature, 2015, 519 (7542): 199.

[214] Wei Y, Terns R M, Terns M P. Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. Genes & development, 2015, 29 (4): 356-361.

[215] Nuñez J K, Kranzusch P J, Noeske J, et al. Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nature structural & molecular biology, 2014, 21 (6): 528.

[216] Staals R H, Agari Y, Maki-Yonekura S, et al. Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. Molecular cell, 2013, 52 (1): 135-145.

[217] Staals R H, Zhu Y, Taylor D W, et al. RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Molecular cell, 2014, 56 (4): 518-530.

[218] Huo Y, Nam K H, Ding F, et al. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nature structural & molecular biology, 2014, 21 (9): 771.

[219] Samai P, Pyenson N, Jiang W, et al. Co-transcriptional DNA and RNA

cleavage during type III CRISPR-Cas immunity. Cell, 2015, 161 (5): 1164-1174.

[220] Makarova K S, Anantharaman V, Aravind L, et al. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. Biology direct, 2012, 7 (1): 40.

[221] Makarova K S, Anantharaman V, Grishin N V, et al. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. Frontiers in genetics, 2014, 5 102.

[222] Nam K H, Kurinov I, Ke A. Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca2+-dependent double-stranded DNA binding activity. Journal of Biological Chemistry, 2011, 286 (35): 30759-30768.

[223] Arslan Z, Wurm R, Brener O, et al. Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. Nucleic Acids Researchearch, 2013, 41 (12): 6347-6359.

[224] Sinkunas T, Gasiunas G, Fremaux C, et al. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. The EMBO journal, 2011, 30 (7): 1335-1342.

[225] Makarova K S, Grishin N V, Shabalina S A, et al. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biology direct, 2006, 1 (1): 7.

[226] Zetsche B, Gootenberg J, Abudayyeh O, et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell, 2015, 163 (3): 759-771.

[227] Zetsche B, Heidenreich M, Mohanraju P, et al. Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. Nature biotechnology, 2017, 35 (1): 31.

[228] Sternberg S H, Redding S, Jinek M, et al. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature, 2014, 507 (7490): 62.

[229] Martin J, Fuguo J, Taylor D W, et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science, 2014, 343 (6176): 1247997.

[230] Xuebing W, Scott D A, Kriz A J, et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nature Biotechnology, 2014, 32 (7): 670-676.

[231] Schwank G, Koo B K, Sasselli V, et al. Functional Repair of CFTR by CRISPR/Cas9 in Intestinal Stem Cell Organoids of Cystic Fibrosis Patients. Cell Stem Cell, 2013, 13 (6): 653-658.

[232] Gratz S J, Ukken F P, C Dustin R, et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. Genetics, 2014, 196 (4): 961-971.

[233] Slaymaker I M, Gao L, Zetsche B, et al. Rationally engineered Cas9 nucleases with improved specificity. Science, 2016, 351 (6268): 84-88.

[234] Chuai G-H, Wang Q-L, Liu Q. In silico meets in vivo: towards computational CRISPR-based sgRNA design. Trends in biotechnology, 2017, 35 (1): 12-21.

[235] Doench J G, Fusi N, Sullender M, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature biotechnology, 2016, 34 (2): 184.

[236] Xu H, Xiao T, Chen C-H, et al. Sequence determinants of improved CRISPR sgRNA design. Genome research, 2015, 25 (8): 1147-1157.

[237] Tim W, Wei J J, Sabatini D M, et al. Genetic screens in human cells using the CRISPR-Cas9 system. Science, 2013, 343 (6166): 80-84.

[238] Doench J G, Hartenian E, Graham D B, et al. Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nature biotechnology, 2014, 32 (12): 1262.

[239] Chari R, Mali P, Moosburner M, et al. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nature Methods, 2015, 12 (9): 823-826.

[240] Lei S, Qi, Larson M H, Gilbert L A, et al. Repurposing CRISPR as an

RNA-guided platform for sequence-specific control of gene expression. Cell, 2013, 152 (5): 1173-1183.

[241] Chen B, Gilbert L A, Cimini B A, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell, 2013, 155 (7): 1479-1491.

[242] Honglei L, Zheng W, Antonia D, et al. CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. Bioinformatics, 2015, 31 (22): 3676-3678.

[243] Zhu L J, Holmes B R, Aronin N, et al. CRISPRseek: a bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. PloS one, 2014, 9 (9): e108424.

[244] Sangsu B, Jeongbin P, Jin-Soo K. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics, 2014, 30 (10): 1473-1475.

[245] Park J, Bae S, Kim J S. Cas-Designer: a web-based tool for choice of CRISPR-Cas9 target sites. Bioinformatics, 2015, 31 (24): btv537.

[246] Prykhozhij S V, Vinothkumar R, Daniel G, et al. CRISPR multitargeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. Plos One, 2015, 10 (3): e0119372.

[247] Wong N, Liu W, Wang X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biologyogy, 2015, 16 (1): 218.

[248] Chuai G, Yang F, Yan J, et al. Deciphering relationship between microhomology and in-frame mutation occurrence in human CRISPR-based gene knockout. Molecular Therapy Nucleic Acids, 2016, 5 (6): e323.

[249] Chuai G, Ma H, Yan J, et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biologyogy, 2018, 19 (1): 80.

[250] Florian H, Grainne K, Michael B. E-CRISP: fast CRISPR target site identification. Nature Methods, 2014, 11 (2): 122-123.

[251] Macpherson C R, Scherf A. Flexible guide-RNA design for CRISPR applications using Protospacer Workbench. Nature Biotechnology, 2015, 33 (8):

805.

[252] Lei Y, Lu L, Liu H Y, et al. CRISPR-P: A Web Tool for Synthetic Single-Guide RNA Design of CRISPR-System in Plants. Molecular Plant, 2014, 7 (9): 1494-1496.

[253] Peng D, Tarleton R L. EuPaGDT: a web tool tailored to design CRISPR guide RNAs for eukaryotic pathogens. Microbial Genomics, 2015, 1 (4):

[254] Lee C M, Cradick T J, Fine E J, et al. Nuclease target site selection for maximizing on-target activity and minimizing off-target effects in genome editing. Molecular Therapy, 2016, 24 (3): 475-487.

[255] Duan J, Lu G, Xie Z, et al. Genome-wide identification of CRISPR/Cas9 off-targets in human genome. Cell research, 2014, 24 (8): 1009.

[256] O'geen H, Henry I M, Bhakta M S, et al. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. Nucleic Acids Researchearch, 2015, 43 (6): 3389-3404.

[257] Kuscu C, Arslan S, Singh R, et al. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nature biotechnology, 2014, 32 (7): 677.

[258] Kim D, Bae S, Park J, et al. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nature methods, 2015, 12 (3): 237.

[259] Tsai S Q, Zheng Z, Nguyen N T, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nature biotechnology, 2015, 33 (2): 187.

[260] Wang X, Wang Y, Wu X, et al. Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. Nature biotechnology, 2015, 33 (2): 175.

[261] Frock R L, Hu J, Meyers R M, et al. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nature biotechnology, 2015, 33 (2): 179.

[262] Crosetto N, Mitra A, Silva M J, et al. Nucleotide-resolution DNA

double-strand break mapping by next-generation sequencing. Nature Methods, 2013, 10 (4): 361-365.

[263] Koike-Yusa H, Li Y, Tan E P, et al. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nature Biotechnology, 2014, 32 (3): 267.

[264] 蔡昌祖, 周悦欣, 朱诗优, et al. 通过哺乳细胞 CRISPR/Cas9 敲除文库实现高通量功能性基因筛选. 中国细胞生物学学报, 2014, 36 (07): 853-856.

[265] Bae S, Kweon J, Kim H S, et al. Microhomology-based choice of Cas9 nuclease target sites. Nature methods, 2014, 11 (7): 705.

[266] Fusi N, Smith I, Doench J, et al. In silico predictive modeling of CRISPR/Cas9 guide efficiency. BioRxiv, 2015, 021568.

[267] Güell M, Yang L, Church G M. Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA). Bioinformatics, 2014, 30 (20): 2968-2970.

[268] Shi J, Wang E, Milazzo J P, et al. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nature biotechnology, 2015, 33 (6): 661.

[269] Shah A N, Davey C F, Whitebirch A C, et al. Rapid reverse genetic screening using CRISPR in zebrafish. Nature methods, 2015, 12 (6): 535.

[270] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics, 2009, 25 (14): 1754-1760.

[271] Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2. Nature methods, 2012, 9 (4): 357.

[272] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:13126034, 2013,

[273] Abadi S, Yan W X, Amar D, et al. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS computational biology, 2017, 13 (10): e1005807.

[274] Hsu P D, Scott D A, Weinstein J A, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nature Biotechnology, 2013, 31 (9): 827.

[275] Yanfang F, Foden J A, Cyd K, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nature Biotechnology, 2013, 31 (9): 822.

[276] Haeussler M, Schönig K, Eckert H, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biologyogy, 2016, 17 (1): 148.

[277] Kim D, Kim S, Kim S, et al. Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. Genome research, 2016, 26 (3): 406-415.

[278] Ran F A, Cong L, Yan W X, et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature, 2015, 520 (7546): 186.

[279] Ma H, Tu L-C, Naseri A, et al. CRISPR-Cas9 nuclear dynamics and target recognition in living cells. J Cell Biol, 2016, 214 (5): 529-537.

[280] 张远望, 人工智能与应用. 中国科技纵横 2015, (20), 22-22.

[281] Yerushalmy, J., Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Reports (1896-1970) 1947, 1432-1449.

[282] Kulikowski, C. A., Artificial Intelligence Methods and Systems for Medical Consultation. IEEE Transactions on Pattern Analysis & Machine Intelligence 2013, PAMI-2 (5), 464-476.

[283] Szolovits, P.; Patil, R. S.; Schwartz, W. B., Artificial intelligence in medical diagnosis. Annals of internal medicine 1988, 108 (1), 80-87.

[284] Süt, N.; Çelik, Y., Prediction of mortality in stroke patients using multilayer perceptron neural networks. Turkish Journal of Medical Sciences 2012, 42 (5), 886-893.

[285] Bentley, P.; Ganesalingam, J.; Jones, A. L. C.; Mahady, K.; Epton, S.; Rinne, P.; Sharma, P.; Halse, O.; Mehta, A.; Rueckert, D., Prediction of stroke thrombolysis outcome using CT brain machine learning. NeuroImage: Clinical 2014, 4, 635-640.

[286] Yan, H.; Jiang, Y.; Zheng, J.; Peng, C.; Li, Q., A multilayer perceptron-based

medical decision support system for heart disease diagnosis. Expert Systems with Applications 2006, 30 (2), 272-281.

[287] Panday, P.; Godara, N., Decision support system for cardiovascular heart disease diagnosis using improved multilayer perceptron. International Journal of Computer Applications 2012, 45 (8).

[288] Mojarad, S.; Dlay, S. S.; Woo, W. L.; Sherbet, G. V., Cross validation evaluation for breast cancer prediction using multilayer perceptron neural networks. American Journal of Engineering and Applied Sciences 2011.

[289] Streba, C. T.; Vere, C. C.; Sandulescu, L. D.; Saftoiu, A.; Gheonea, D. I.; Streba, L.; Rogoveanu, I., Sa1000 Focal Liver Lesions Classification by Artificial Neural Networks and Support Vector Machines Employing Dynamic Imaging Data. Gastroenterology 2014, 146 (5), S-933.

[290] Vorontsov, E.; Tang, A.; Roy, D.; Pal, C. J.; Kadoury, S., Metastatic liver tumour segmentation with a neural network-guided 3D deformable model. Medical & biological engineering & computing 2017, 55 (1), 127-139.

[291] Li, S.; Jiang, H.; Pang, W., Joint multiple fully connected convolutional neural network with extreme learning machine for hepatocellular carcinoma nuclei grading. Comput. Biol. Med. 2017, 84, 156-167.

[292] James, S. L.; Henderson, E. E.; Shatzel, J. J.; Dickson, R., Mo1903 machine learning classifiers: A novel approach to predicting bleeding risk in hospitalized cirrhotic patients. Gastroenterology 2015, 148 (4), S-1079.

[293] Reddy, R.; Imler, T. D., Artificial Neural Networks are Highly Predictive for Hepatocellular Carcinoma in Patients with Cirrhosis. Gastroenterology 2017, 152 (5), S1193.

[294] Gao, S.; Peng, Y.; Guo, H.; Liu, W.; Gao, T.; Xu, Y.; Tang, X., Texture analysis and classification of ultrasound liver images. Bio-Med. Mater. Eng. 2014, 24 (1), 1209-1216.

[295] Chen, Y.; Luo, Y.; Huang, W.; Hu, D.; Zheng, R.-q.; Cong, S.-z.; Meng, F.-k.; Yang, H.; Lin, H.-j.; Sun, Y., Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic

hepatitis B. Comput. Biol. Med. 2017, 89, 18-23.

[296] Lemoine, M.; Thursz, M.; Mallet, V.; Shimakawa, Y., Diagnostic accuracy of the gamma-glutamyl transpeptidase to platelet ratio (GPR) using transient elastography as a reference. Gut 2017, 66 (1), 195-196.

[297] Lu, X.-J.; Li, X.-H.; Yuan, Z.-X.; Sun, H.-Y.; Wang, X.-C.; Qi, X.; Zhang, X.; Sun, B., Assessment of liver fibrosis with the gamma-glutamyl transpeptidase to platelet ratio: a multicentre validation in patients with HBV infection. Gut 2018, 67 (10), 1903-1904.

[298] Johanson, J.; Frakes, J.; Eisen, D., Computer-assisted analysis of abrasive transepithelial brush biopsies increases the effectiveness of esophageal screening: a multicenter prospective clinical trial by the EndoCDx Collaborative Group. Digestive diseases and sciences 2011, 56 (3), 767-772.

[299] Chan, D. K.; Zakko, L.; Visrodia, K. H.; Leggett, C. L.; Lutzke, L. S.; Clemens, M. A.; Allen, J. D.; Anderson, M. A.; Wang, K. K., Breath testing for Barrett's esophagus using exhaled volatile organic compound profiling with an electronic nose device. Gastroenterology 2017, 152 (1), 24-26.

[300] Chan, D. K.; Lutzke, L. S.; Clemens, M. A.; Leggett, C. L.; Wang, K. K., 299 Detection of Barrett's Esophagus by Non-invasive Breath Screening of Exhaled Volatile Organic Compounds Using an Electronic-Nose Device. Gastroenterology 2016, 150 (4), S67.

[301] Săftoiu, A.; Vilmann, P.; Gorunescu, F.; Janssen, J.; Hocke, M.; Larsen, M.; Iglesias–Garcia, J.; Arcidiacono, P.; Will, U.; Giovannini, M., Efficacy of an artificial neural network–based approach to endoscopic ultrasound elastography in diagnosis of focal pancreatic masses. Clinical Gastroenterology and Hepatology 2012, 10 (1), 84-90. e1.

[302] Burge, P. S.; Pantin, C. F.; Newton, D. T.; Gannon, P. F.; Bright, P.; Belcher, J.; Mccoach, J.; Baldwin, D. R.; Burge, C. B., Development of an expert system for the interpretation of serial peak expiratory flow measurements in the diagnosis of occupational asthma. Midlands Thoracic Society Research Group. Occupational & Environmental Medicine 1999, 56 (11), 758-764.

[303] Gautier, V.; Redier, H.; Pujol, J. L.; Bousquet, J.; Proudhon, H.; Michel, C.; Daures, J. P.; Michel, F. B.; Godard, P., Comparison of an expert system with other clinical scores for the evaluation of severity of asthma. European Respiratory Journal 1996, 9 (1), 58.

[304] Matsuki, Y.; Nakamura, K.; Watanabe, H.; Aoki, T.; Nakata, H.; Katsuragawa, S.; Doi, K., Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: evaluation with receiver operating characteristic analysis. Ajr American Journal of Roentgenology 2002, 178 (3), 657.

[305] Mcculloch, C. C.; Kaucic, R. A.; Mendonça, P. R.; Walter, D. J.; Avila, R. S., Model-based detection of lung nodules in computed tomography exams. Thoracic computer-aided diagnosis. Academic Radiology 2004, 11 (3), 258-266.

[306] Pavlou, A. K.; Turner, A. P., Sniffing out the truth: clinical diagnosis using the electronic nose. Clinical Chemistry & Laboratory Medicine 2000, 38 (2), 99-112.

[307] Kanis, J. A., Diagnosis of osteoporosis and assessment of fracture risk. Lancet 2002, 359 (9321), 1929-1936.

[308] Henderson, J. E.; Goltzman, D., The Osteoporosis Primer: The Osteoporosis Primer. Clinical Endocrinology 2010, 54 (1), 133-134.

[309] Ongphiphadhanakul, B.; Rajatanavin, R.; Chailurkit, L.; Piaseu, N.; Teerarungsikul, K.; Sirisriro, R.; Komindr, S.; Pauvilai, G., Prediction of low bone mineral density in postmenopausal women by artificial neural network model compared to logistic regression model. Journal of the Medical Association of Thailand= Chotmaihet thangphaet 1997, 80 (8), 508-515.

[310] Tafraouti, A.; El Hassouni, M.; Toumi, H.; Lespessailles, E.; Jennane, R. In Osteoporosis diagnosis using fractal analysis and support vector machine, 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems, IEEE: 2014; pp 73-77.

[311] Iliou, T.; Anagnostopoulos, C.-N.; Stephanakis, I. M.; Anastassopoulos, G., A

novel data preprocessing method for boosting neural network performance: a case study in osteoporosis prediction. Information Sciences 2017, 380, 92-100.

[312] Liu, Q.; Cui, X.; Chou, Y.-C.; Abbod, M. F.; Lin, J.; Shieh, J.-S., Ensemble artificial neural networks applied to predict the key risk factors of hip bone fracture for elders. Biomed. Signal Process. Control 2015, 21, 146-156.

[313] Yu, X.; Ye, C.; Xiang, L., Application of artificial neural network in the diagnostic system of osteoporosis. Neurocomputing 2016, 214, 376-381.

[314] Vohora DGS. Pharmaceutical Medicine and Translational Clinical Research. Academic Press, 2018.

[315] 刘琦. 人工智能与药物研发. 第二军医大学学报 2018; 39: 869-72.

[316] Smietana K, Siatkowski M, Moller M. Trends in clinical success rates. Nat Rev Drug Discov 2016; 15: 379-80.

[317] Feisheng Zhong, Jing Xing, Xutong Li, et al. Artificial intelligence in drug design. Science China (Life Sciences), 2018, 61 (10):59-72.

[318] Keller T H , Pichota A , Yin Z. A practical view of ‘druggability’. Current Opinion in Chemical Biology, 2006, 10(4):357-361.

[319] Bakheet T M , Doig A J . Properties and identification of human protein drug targets. Bioinformatics, 2009, 25(4):451-457.

[320] Huang C , Zhang R , Chen Z , et al. Predict potential drug targets from the ion channel proteins based on SVM. Journal of Theoretical Biology, 2010, 262(4):750-756.

[321] Ya-Wei Z , Zhen-Dong S , Wuritu Y , et al. IonchanPred 2.0: A Tool to Predict Ion Channels and Their Types. International Journal of Molecular Sciences, 2017, 18(9):1838.

[322] Lin H , Ding H . Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. Journal of Theoretical Biology, 2011, 269(1):64-69.

[323] Chen W , Lin H . Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. Computers in Biology and Medicine, 2012, 42(4):504-507.

[324] Xin-Xin C , Hua T , Wen-Chao L , et al. Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. BioMed Research International, 2016, 2016:1-8.

[325] Huan Y , Hua T , Xin-Xin C , et al. Identification of Secretory Proteins in\r, Mycobacterium tuberculosis\r, Using Pseudo Amino Acid Composition. BioMed Research International, 2016, 2016:1-7.

[326] Lai H Y , Chen X X , Chen W , et al. Sequence-based predictive modeling to identify cancerlectins. Oncotarget, 2017, 8(17).

[327] Bakhtiarizadeh M R , Moradi-Shahrbabak M , Ebrahimi M , et al. Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. Journal of Theoretical Biology, 2014, 356:213-222.

[328] Kayvanjoo A , Ebrahimi M , Haqshenas G . Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC Research Notes, 2014, 7(1):565.

[329] Zhao Y W , Lai H Y , Tang H , et al. Prediction of phosphothreonine sites in human proteins by fusing different features. Scientific Reports, 2016, 6:34817.

[330] Hardy L W , Peet N P . The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. Drug Discovery Today, 2004, 9(3):117-126.

[331] Li Z C , Zhong W Q , Liu Z Q , et al. Large-scale identification of potential drug targets based on the topological features of human protein–protein interaction network. Analytica Chimica Acta, 2015, 871:18-27.

[332] Tang H , Su Z D , Wei H H , et al. Prediction of cell-penetrating peptides with feature selection techniques. Biochemical and Biophysical Research Communications, 2016:S0006291X16309536.

[333] Feng P M , Ding H , Chen W , et al. Naïve Bayes Classifier with Feature Selection to Identify Phage Virion Proteins. Computational and Mathematical Methods in Medicine, 2013, 2013(2):530696.

[334] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al.

Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 2019.

[335] 龚家瑜 (2013), 基于数据挖掘的药物靶标发现方法研究, 华东理工大学.

[336] Schneider M. A rational approach to maximize success rate in target discovery. Arch Pharm (Weinheim)2004; 337: 625-33.

[337] Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov 2002; 1: 727-30.

[338] Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature 2009; 462: 175-81.

[339] Garcia-Serna R, Ursu O, Oprea TI, Mestres J. iPHACE: integrative navigation in pharmacological space. Bioinformatics 2010; 26: 985-6.

[340] 吴曾睿 (2018), 基于网络的药物设计方法发展及其在抗癌药物研究中的应用, 华东理工大学.

[341] Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol 2012; 8: e1002503.

[342] Costa PR, Acencio ML, Lemke N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. BMC Genomics 2010; 11 Suppl 5: S9.

[343] Bravo A, Pinero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics 2015; 16: 55.

[344] 王磊 (2018), 基于机器学习的药物—靶标相互作用预测研究, 中国矿业大学.

[345] Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today 2017; 22: 1680-5.

[346] Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med 2012; 4: 125ra31.

[347] Dakshanamurthy S, Issa NT, Assefnia S, Seshasayee A, Peters OJ, Madhavan S,

et al. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem 2012; 55: 6832-48.

[348] Searls DB. Data integration: challenges for drug discovery. Nat Rev Drug Discov 2005; 4: 45-58.

[349] Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol 2009; 5: e1000450.

[350] Iwata H, Sawada R, Mizutani S, Yamanishi Y. Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. J Chem Inf Model 2015; 55: 446-59.

[351] Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 2018.

[352] 张永祥, 程肖蕊, 周文霞. 药物重定位——网络药理学的重要应用领域. 中国药理学与毒理学杂志 2012; 26: 779-86.

[353] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004; 3: 673-83.

[354] Guney E , Menche J , Vidal M , et al. Network-based in silico drug efficacy screening. Nature Communications, 2016, 7:10331.

[355] Menche J, Sharma A, Kitsak M, et al. Uncovering disease-disease relationships through the incomplete interactome. Science, 2015, 347(6224):1257601.

[356] Zitnik M, Nguyen F, Wang B, et al. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. Information Fusion, 2018: 71-91.

[357] Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nature Communications, 2017, 8(1):573.

[358] Wu Z, Wang Y, Chen L. Network-based drug repositioning. Molecular Biosystems, 2013, 9(6):1268-1281.

[359] Zhao S , Li S . A co-module approach for elucidating drug-disease associations

and revealing their molecular basis. Bioinformatics, 2012, 28(7):955-961.

[360] Fung K W , Jao C S , Demner-Fushman D . Extracting drug indication information from structured product labels using natural language processing. Journal of the American Medical Informatics Association, 2013, 20(3):482-488.

[361] Wang F , Zhang P , Cao N , et al. Exploring the associations between drug side-effects and therapeutic indications. Journal of Biomedical Informatics, 2014, 51:15-23.

[362] Wang S , Peng J . Network-assisted target identification for haploinsufficiency and homozygous profiling screens. PLoS Computational Biology, 2017, 13(6):e1005553.

[363] Mizutani S, Pauwels E, Stoven V, et al. Relating drug-protein interaction network with drug side effects. Bioinformatics, 2012, 28(18): i522-i528.

[364] Iorio F, Bosotti R, Scacheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci USA, 2010, 107(8): 14621-6.

[365] Wang W, Yang S, Zhang X, et al. Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics, 2014, 30(20): 2923-2930.

[366] Yang F, Xu J, Zeng J. Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data. Pac Symp Biocomput, 2014, 19(2): 148-159.

[367] Zhang X, Li L, Ng M K, et al. Drug-target interaction prediction by integrating multiview network data. Computational Biology & Chemistry, 2017, 69: S1476927117301950.

[368] Gönen M, Kaski S. Kernelized Bayesian Matrix Factorization. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36(10): 2047-2060.

[369] Lee S, Zhang C, Liu Z, et al. Network analyses identify liver-specific targets for treating liver diseases. Molecular Systems Biology, 2017, 13(8): 938.

[370] Fu G, Ding Y, Seal A, et al. Predicting drug target interactions using meta-path-based semantic network analysis. BMC Bioinformatics, 2016, 17(1): 160.

[371] Zitnik M, Zupan B. [WORLD SCIENTIFIC Proceedings of the Pacific Symposium-Kohala Coast, Hawaii, USA (4-8 January 2016)] Biocomputing 2016-COLLECTIVE PAIRWISE CLASSIFICATION FOR MULTI-WAY ANALYSIS OF DISEASE AND DRUG DATA. Pac Symp Biocomput: 81-92.

[372] Wu C C, Asgharzadeh S, Triche T J, et al. Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning. Bioinformatics, 2010, 26(6): 807-813.

[373] Han K, Jeng E E, Hess G T, et al. Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. Nature Biotechnology, 2017.

[374] Sun Y, Sheng Z , Ma C , et al. Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. Nature Communications, 2015, 6: 8481.

[375] Woo H G, Choi J H, Yoon S, et al. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. Nature Communications, 2017, 8(1): 839.

[376] Chen X, Ren B, Chen M, et al. NLLSS: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning. Plos Computational Biology, 2016, 12(7): e1004975.

[377] Lewis R, Guha R, Korcsmaros T, et al. Synergy Maps: exploring compound combinations using network-based visualization. Journal of Cheminformatics, 7, 1(2015-08-01), 2015, 7(1): 36.

[378] A community computational challenge to predict the activity of pairs of compounds. Nature Biotechnology, 2014, 32(12): 1213-1222.

[379] Takeda T, Hao M, Cheng T, et al. Predicting drug-drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. Journal of

Cheminformatics, 2017, 9(1): 16.

[380] Shi J, Li J, Gao K, et al. Predicting combinative drug pairs towards realistic screening via integrating heterogeneous features. BMC bioinformatics, 2017, 18(12): 409.

[381] Zitnik M, Zupan B. Data Fusion by Matrix Factorization. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2015, 37(1): 41-53.

[382] Li X, Xu Y, Cui H, et al. Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles. Artificial intelligence in medicine, 2017, 83.

[383] Ferdousi R, Safdari R, Omidi Y. Computational prediction of drug-drug interactions based on drugs functional similarities. Journal of Biomedical Informatics, 2017, 70: 54-64.

[384] Zhang W, Chen Y, Liu F, et al. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC Bioinformatics, 2017, 18(1): 18.

[385] Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 2018, 34(13): i457-i466.

[386] Ryu J Y, Kim H U, Lee S Y. Deep learning improves prediction of drug-drug and drug-food interactions. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(18): E4304-E4311.

[387] 贾志龙 (2016), 基于转录组数据的药物重定位, 国防科学技术大学.

[388] 智能医疗未来发展的三种积极趋势——《人工智能：驯服赛维坦》.

[389] 中国人工智能医疗白皮书. 2019.

[390] Pharmaceutical Giants Are Chasing R&D Deals With Smaller AI Startups. https://www.biopharmatrend.com/post/79-pharmaceutical-giants-are-chasing-rd-deals-with-smaller-ai-startups/ last accessed).

[391] Mao W, Diggavi S, Kannan S. Models and information-theoretic bounds for nanopore sequencing. IEEE International Symposium on Information Theory, 2017: 2458-2462.

[392] Rhoads A, Au K F. PacBio Sequencing and Its Applications. Genomics

Proteomics & Bioinformatics, 2015, 13(5): 278-289.

[393] Jain M, Olsen H E, Paten B, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biologyogy, 2016, 17(239): 1-11.

[394] Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics & Bioinformatics, 2016, 14(5): 265-279.

[395] Cali D S, Kim J S, Ghose S, et al. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. Briefings in Bioinformatics, 2017, 1-18.

[396] Ewing B, Hillier L, Wendl M C, et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research, 1998, 8(3): 175-185.

[397] Travers K, Chin C, Rank D, et al. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Researchearch, 2010, 38(15): 159-168.

[398] Jang-il Sohn, Jin-Wu Nam, The present and future of de novo whole-genome assembly, Briefings in Bioinformatics, 2018,19, 23-40.

[399] Angeleri E, Apolloni B, Falco D, et al. DNA fragment assembly using neural prediction techniques. International journal of neural systems, 1999, 9(06): 523-544.

[400] Krachunov M, Nisheva M, Vassilev D. Machine learning models in error and variant detection in high-variation high-throughput sequencing datasets. Procedia Computer Science, 2017, 108: 1145-1154.

[401] Palmer L E, Dejori M, Bolanos R, et al. Improving de novo sequence assembly using machine learning and comparative genomics for overlap correction. BMC bioinformatics, 2010, 11(1): 33.

[402] Zhu X, Leung H C M, Chin F Y L, et al. PERGA: a paired-end read guided de novo assembler for extending contigs using SVM and look ahead approach. PloS one, 2014, 9(12): e114253.

[403] Wang, Q., Yu, H., Zhao, Z., & Jia, P. EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. Bioinformatics, 2015, 31(15), 2591-2594.

[404] Lanc, Irena, and Scott Emrich. “An unsupervised learning approach to assembly validation”. 2013 IEEE 3rd International Conference on Computational Advances in Bio and medical Sciences (ICCABS). IEEE, 2013.

[405] Kuhring M, Dabrowski P W, Piro V C, et al. SuRankCo: supervised ranking of contigs in de novo assemblies. BMC bioinformatics, 2015, 16(1): 240.

[406] Cacho A, Smirnova E, Huzurbazar S, et al. A comparison of base-calling algorithms for Illumina sequencing technology. Briefings in bioinformatics, 2015, 17(5): 786-795.

[407] Haotian Teng, Minh Duc Cao, Michael B Hall, Tania Duarte, Sheng Wang, Lachlan J M Coin. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning, GigaScience, 2018, 7.

[408] Tamas S, Golovchenko J A. De novo sequencing and variant calling with nanopores using PoreSeq Nature Biotechnology, 2015, 33(10): 1087-1091.

[409] Torracinta R, Mesnard L, Levine S, et al. Adaptive Somatic Mutations Calls with Deep Learning and Semi-Simulated Data, 2016.

[410] Krueger, Felix, and Simon R. Andrews. “Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications”. bioinformatics 27.11, 2011: 1571-1572.

[411] Flusberg, Benjamin A., et al. “Direct detection of DNA methylation during single-molecule, real-time sequencing”. Nature methods 7.6, 2010: 461.

[412] Stoiber, Marcus H., et al. “De novo identification of DNA modifications enabled by genome-guided nanopore signal processing”. bioRxiv, 2016: 094672.

[413] Liu, Qian, et al. “NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data”. bioRxiv, 2018: 277178.

[414] Simpson, Jared T., et al. "Detecting DNA cytosine methylation using nanopore sequencing." nature methods 14.4, 2017: 407.

[415] Rand, Arthur C., et al. “Mapping DNA methylation with high-throughput nanopore sequencing”. Nature methods 14.4, 2017: 411.

[416] McIntyre, Alexa BR, et al. “Nanopore detection of bacterial DNA base modifications”. bioRxiv, 2017: 127100.

[417] Angermueller, Christof, et al. “DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning”. Genome Biologyogy 18.1, 2017: 67.

[418] Fu, Laiyi, Qinke Peng, and Ling Chai. “Predicting DNA methylation states with hybrid information based deep-learning model”. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019.

[419] Wang, Yiheng, et al. “Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks”. Scientific reports 6, 2016: 19598.

[420] Ni, Peng, et al. “DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning”. Bioinformatics, 2019.

[421] Liu, Qian, et al. “Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data”. Nature communications 10.1, 2019: 2449.

[422] Peng, Jianhao, Idoia Ochoa, and Olgica Milenkovic. “E2M: A Deep Learning Framework for Associating Combinatorial Methylation Patterns with Gene Expression”. bioRxiv, 2019: 527044.

[423] Li HD, Menon R, Omenn GS, and Guan Y. The emerging era of genomic data integration for analyzing splice isoform function,Trends in Genetics, 2014, 340-347.

[424] Barutcuoglu, Z., Schapire, R. E., & Troyanskaya, O. G. Hierarchical multi-label prediction of gene function. Bioinformatics, 2006, 22(7), 830-836.

[425] Vinayagam, A., Kfnig, R., Moormann, J., Schubert, F., Eils, R., Glatting, K. H., & Suhai, S. Applying support vector machines for gene ontology based gene function prediction. BMC bioinformatics, 2004, 5(1), 116.

[426] Li, Z., Liao, B., Li, Y., Liu, W., Chen, M., & Cai, L. Gene function prediction

based on combining gene ontology hierarchy with multi-instance multi-label learning. RSC advances, 2018, 8(50), 28503-28509.

[427] Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proceedings of the National Academy of Sciences, 2003, 100(14), 8348-8353.

[428] Guan, Y., Myers, C. L., Hess, D. C., Barutcuoglu, Z., Caudy, A. A., & Troyanskaya, O. G. Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biologyogy, 2008, 9(1), S3.

[429] Kulmanov M, Khan M A, Hoehndorf R. DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics, 2017, 34(4): 660-668.

[430] Gligorijevic V, Barot M, Bonneau R. deepNF: Deep network fusion for protein function prediction. Bioinformatics, 2017.

[431] Cao R, Freitas C, Chan L, et al. ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. Molecules, 2017, 22(10): 1732.

[432] Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet. 2006; 7: 29-59.

[433] Alipanahi, B., Delong, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learningBiotechnol. 2015, 33, 831-838.

[434] Lee N K, Azizan F L, Wong Y S, et al. DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery. Biotechnology & Biotechnological Equipment, 2018:1-10.

[435] Hache H. et al. Reverse engineering of gene regulatory networks: a comparative study. EURASIP J Bioinform. Syst. Biol., 2009, 617281.

[436] Jiguo, C., et al. Modeling gene regulation network using ordinary differential equation. In: Next Generation Microarray. Bioinformatics, 2012: 185-197.

[437] Patel, N., Wang, J.T.L.: Semi-supervised prediction of gene regulatory networks using machine learning algorithms. J. Biosci. 2015, 40(4), 731-740.

[438] Jisha, A., Jereech, A.S.: Gene regulatory network: a semi supervised approach. In: International Conference on Electronics Communication and Aerospace Technology ICECA, 2017.

[439] Mordelet, F., Vert, J.P.: SIRENE: supervised inference of regulatory network. Bioinformatics 24, 2018, i76-i82.

[440] Gillani, Z., et al. Compare SVM: supervised support vector machine (SVM) inference of gene regularity network. BMC Bioinform, 2014.

[441] Mandal S. et al. Large-scale recurrent neural network based modelling of gene regulatory network using cuckoo search-flower pollination algorithm. Adv. Bioinformatics, 2016: 5283937.

[442] Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G., & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. Proceedings of the National Academy of Sciences, 2015, 112(1), 118-123.

[443] Wang Q, Fish J A, Gilman M, et al. Xander: employing a novel method for efficient gene-targeted metagenomic assembly. Microbiome, 2015, 3(1): 32.

[444] Wu, G., & Stein, L. A network module-based method for identifying cancer prognostic signatures. Genome Biologyogy, 2012, 13(12), R112.

[445] Kim, J., Kim, J. J., & Lee, H. An analysis of disease-gene relationship from Medline abstracts by DigSee. Scientific reports, 2017, 7, 40154.

[446] Luo, P., Li, Y., Tian, L. P., & Wu, F. X. Enhancing the prediction of disease-gene associations with multimodal deep learning. Bioinformatics, 2019.

[447] Mordelet, F., & Vert, J. P. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. BMC bioinformatics, 2011, 12(1), 389.

[448] Luo, P., Tian, L. P., Ruan, J., & Wu, F. X. Disease gene prediction by integrating PPI networks, clinical RNA-Seq data and OMIM data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2019, 16(1), 222-232.

[449] Yang, P., Li, X. L., Mei, J. P., Kwoh, C. K., & Ng, S. K. Positive-unlabeled learning for disease gene identification. Bioinformatics, 2012, 28(20), 2640-2647.

[450] Ata, S. K., Ou-Yang, L., Fang, Y., Kwoh, C. K., Wu, M., & Li, X. L. Integrating node embeddings and biological annotations for genes to predict disease-gene associations. BMC systems biology, 2018, 12(9), 138.

[451] Luo, P., Xiao, Q., Wei, P. J., Liao, B., & Wu, F. Identifying disease-gene associations with graph-regularized manifold learning. Frontiers in genetics, 2019, 10, 270.

[452] Guan, Y., Gorenshteyn, D., Burmeister, M., Wong, A. K., Schimenti, J. C., Handel, M. A., Troyanskaya, O. G. Tissue-specific functional networks for prioritizing phenotype and disease genes. PLoS computational biology, 2012, 8(9), e1002694.

[453] Ata, S. K., Fang, Y., Wu, M., Li, X. L., & Xiao, X. Disease gene classification with metagraph representations. Methods, 2017, 131, 83-92.

[454] Cao, S., Lu, W., & Xu, Q. Deep neural networks for learning graph representations. In Thirtieth AAAI Conference on Artificial Intelligence, 2016.