人工智能在核酸药物研发中的应用和进展

胡子昂[1], 高利明[2], 余文颖[3*]

([1] 中国药科大学孟目的学院, 南京 211198; [2] 中国药科大学理学院, 南京 211198;
[3] 中国药科大学生命科学与技术学院, 南京 211198)

摘　要　近年来核酸药物领域蓬勃发展,正逐步成为小分子和抗体类药物后的第三代药物新模式。基于机器学习的人工智能技术进步迅速,可以极大推进核酸药物研发的进程。本文概述了核酸药物研发领域的人工智能算法、数据库、表征等基础,阐述了人工智能在核酸结构预测、小核酸药物设计等核酸药物研发环节中的应用和进展,旨在为人工智能和核酸药物交叉学科发展提供参考。

关键词　人工智能;核酸药物;核酸数据库;核酸表征;核酸结构预测;小核酸药物设计

中图分类号　R914;TP18　　**文献标志码**　A　　**文章编号**　1000−5048(2024)03−0335−12

doi: 10.11665/j.issn.1000−5048.2024033101

引用本文 胡子昂,高利明,余文颖. 人工智能在核酸药物研发中的应用和进展 [J]. 中国药科大学学报,2024,55(3):335 − 346.

Cite this article as: HU Zi'ang, GAO Liming, YU Wenying. Advances in the application of artificial intelligence in nucleic acid drug development[J]. *J China Pharm Univ*, 2024, 55(3): 335 − 346.

Advances in the application of artificial intelligence in nucleic acid drug development

HU Zi'ang[1], GAO Liming[2], YU Wenying[3*]

[1]Mudi Meng Honors College, China Pharmaceutical University, Nanjing 211198; [2]School of Science, China Pharmaceutical University, Nanjing 211198; [3]School of Life Science and Technology, China Pharmaceutical University, Nanjing 211198, China

Abstract　In recent years, the field of nucleic acid therapeutics has been flourishing, progressively establishing itself as the third generation of drug modalities following small molecules and antibody-based drugs. Artificial intelligence technology based on machine learning is advancing rapidly, which can significantly accelerate the development process of nucleic acid therapeutics. This review provides an overview of the foundational aspects of artificial intelligence algorithms, databases, and characterizations in the field of nucleic acid drug development. It elucidates the advances in the application of artificial intelligence in nucleic acid structural prediction, small nucleic acid drug design, and other research and development phases of nucleic acid therapeutics, aiming to offer some reference for the interdisciplinary development of artificial intelligence and nucleic acid drugs.

Key words　artificial intelligence; nucleic acid drugs; nucleic acid database; nucleic acid characterization; nucleic acid structure prediction; small nucleic acid drug design

癌症作为危及人类健康的重大疾病之一,至今仍未出现彻底攻克的有效手段。近年来兴起的基因治疗是一类有前景的、精确的治疗手段,以特定核酸序列的方式靶向致病基因。核酸成药正在开创新药研发的新领域,有望治疗癌症等各种基因特异性疾病[1]。核酸药物能够针对传统小分子或蛋白质/抗体药物无法作用的靶点,这一独特的特性使得核酸药物在近些年被广泛开发利用,且相较于小分子或蛋白质/抗体药物,核酸药物具有更短的研发周期和更广阔的治疗领域,在治疗人类疾病如癌症、病毒感染和遗传性疾病方面拥有巨大潜力。与此同时,以机器学习(machine learning, ML)、深度学习(deep learning, DL)为代表的人工智能在近几年有了极大的发展。本文首次以核酸药物研发领域

收稿日期　2024-03-31　*** 通信作者**　Tel: 18013840187　E-mail: ywy@cpu.edu.cn

的人工智能算法、数据库、表征等基础作为切入点，详细阐述了人工智能在核酸结构预测、小核酸药物设计等核酸药物研发环节中的应用和进展。

1 核酸药物研发中的人工智能基础

1.1 人工智能算法

在当前核酸药物研发领域，人工智能算法的应用呈现出多样化的特征，包括 ML、DL 等。ML 算法包括监督学习（supervised learning, SL）、无监督学习（unsupervised learning, UL）和强化学习（reinforcement learning, RL）。其中，SL 通过建立模型识别数据中的关联模式，以 k-近邻算法和决策树算法为代表算法，常用于解决分类问题，当用于解决回归问题时则多用线性回归和逻辑回归；UL 可以在数据未标记的情况下探索数据的内在结构和特征，多用于解决降维、聚类问题；RL 主要用于解决决策问题，通过智能体与环境的交互学习来最大化累积奖励。DL 算法的技术基础是深度神经网络（deep neural networks, DNN），包括卷积神经网络（convolutional neural networks, CNN），由卷积层、池化层、激活函数和全连接层组成，多用于二维结构图像的分割分类；循环神经网络（recurrent neural network, RNN），可以分为双向、深度循环神经网络和长短期记忆网络（long short-term memory, LSTM）；Transformer，序列建模任务中的主流算法，可用于自然语言处理，进而可以进行交互式核酸药物设计。深度强化学习（deep reinforcement learning, DRL）融合了 RL 和 DL，包括动态规划、蒙特卡洛、时间差分学习等基于值函数的算法，以及基于执行器评价器和深度确定性的策略梯度算法。此外，还有深度生成模型（deep generative model, DGM），包括变分自编码器、生成式对抗网络和流生成模型，能够学习数据分布并生成新数据样本，为核酸药物发现和设计提供思路方法。

1.2 核酸数据库

高质量的数据是人工智能的生命。随着新技术新方法在生物医药研究中的不断应用，生物大分子序列、结构等数据量高速增长，越来越多的数据库也应运而生。现有核酸数据库可以按核酸序列、碱基对相互作用、三维空间构象等核酸结构层级分为一级、二级、三级核酸数据库，表 1 对常用核酸数据库进行了总结。

表 1 常用核酸数据库

类别	名称	特点	地址
一级核酸数据库	GenBank	最高频核酸数据库，常用Entrez id检索访问，或用BLAST等工具对比序列[6]	ncbi.nlm.nih.gov/genbank
	ENA	提供全球核苷酸测序信息记录，涵盖原始测序数据、序列组装信息和功能注释[7]	ebi.ac.uk/ena/browser/home
	DDBJ	数据主要通过Sakura和MST工具完成[8]	ddbj.nig.ac.jp/index-e.html
二级核酸数据库	RefSeq	包含基因组、转录本和蛋白质的全面、非冗余、注释良好的参考序列和相关信息[9]	ncbi.nlm.nih.gov/refseq
	miRBase	microRNA数据库，可以分析microRNA基因组定位和挖掘microRNA序列间关系[10]	mirbase.org
	dbEST	GenBank分支之一，表达序列标签[6]	ncbi.nlm.nih.gov/genbank/dbest
	RNAcentral	非编码RNA序列数据库，整合了Ensembl、GENCODE、HGNC、lncRNAdb等51个数据库[11]	rnacentral.org
三级核酸数据库	NDB	专门收录核酸、核酸-蛋白质复合物等核酸相关结构数据和注释信息，提供特征分析工具和几何序列搜索、结构可视化等工具[12]	ndb-archive.rcsb.rutgers.edu
	NAKB	将NDB中的信息与附加序列、结构、功能和基于相互作用的注释集成到含核酸的三级结构，包括对所有核酸类型的等价计算，能够进行更精确的检索[13]	nakb.org
	wwPDB	管理归档了生物大分子结构数据，由RCSB PDB、PDBe、PDBj、BMRB等子数据库组成	wwpdb.org
	NPIDB	蛋白-核酸复合物结构数据库，侧重相互作用模式分类、相互作用界面水分子保守性信息[14]	ngdc.cncb.ac.cn/databasecommons/database/id/369
	PDIdb	蛋白-核酸复合物结构数据库，侧重高质量复合物结构、蛋白质-DNA界面间残基相互作用[15]	melolab.org/pdidb/web/content/home
	DNAproDB	蛋白-核酸复合物结构数据库，侧重生化特征[16]	dnaprodb.usc.edu
	PRIDB	蛋白-核酸复合物结构数据库，侧重非冗余蛋白-RNA相互作用界面综合性数据[17]	ngdc.cncb.ac.cn/databasecommons/database/id/614

其中, 世界三大一级核酸数据库 GenBank、ENA(European Nucleotide Archive)、DDBJ(DNA Data Bank of Japan), 每日交换、同步数据, 构成国际核酸序列数据库合作联盟(INSDC)[2]。三级核酸数据库包括专门核酸结构数据库、包含核酸结构的大分子数据库、蛋白-核酸复合物结构数据库。全球蛋白质数据库组织(wwPDB)管理归档了现有生物大分子结构数据, 主要由 RCSB PDB、PDBe、PDBj、BMRB 组成, RCSB PDB 中包含了通过 X 射线单晶衍射得到的 DNA 和 RNA 结构数据[3]、BMRB 中包含了核酸大分子的核磁共振实验数据[4] 等。

此外, 根据核酸研究(nucleic acids research, NAR)数据库统计报告[5], 2023 年有 90 个新数据库上线、82 个数据库更新。还有许多核酸数据库未详细介绍, 如 DNA 酶数据库 DNAmoreDB、mRNA 数据库 mirDIP 和 UTRdb、原核生物基因组数据库 ECDC 和 NRSub 等。

1.3　核酸表征方法

数据决定人工智能的上限, 而精准的数据表征可以最大程度发挥数据的价值。在当前人工智能算法大多是向量处理算法的背景下, 核酸表征都是从核酸的结构特征或性质特征出发, 将核酸数据转化为向量的同时尽量保留原始信息。目前, 核酸数据表征可以分为基于序列信息, 理化性质和二级、三级结构的特征表征, 下文对每类表征中最常用的 1~2 种表征进行重点介绍。

1.3.1　基于序列信息的表征　独热(one-hot)编码是一种常用的特征编码方式, 尤其适用于纯数据驱动的 DL 模型, 表征核酸序列时可以不依赖先验生物学知识。核酸序列原始数据中的“A”“C”“G”“T”可以转换为二进制向量, 它们可以分别被编码为(1,0,0,0):(0,1,0,0):(0,0,1,0):(0,0,0,1), 随后长度为 n 的核酸序列就可以被表征为 $4\times n$ 的矩阵, 如图 1 所示。经过转换, 核酸序列就可以作为 DL 模型的输入层。基于 CNN 预测蛋白质核酸相互作用偏好的 DeepBind[18] 是第一个使用 one-hot 编码表征核酸序列的 DL 模型。基于 CNN 和 RNN 预测蛋白质和 RNA 相互作用的 iDeepS[19] 更进一步, 使用 one-hot 编码表征核酸的序列和预测所得二级结构, 用于后续卷积操作。

	A	T	C	C	G	A	G	A	C	G	T	G	G	A	T	G
A	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0
C	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
G	0	0	0	0	1	0	1	0	0	1	0	1	1	0	0	1
T	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0

图 1　独热(one-hot)编码示意图

开放阅读框(open reading frame, ORF)也可以用作核酸特征表征。“阅读框”指双链基因序列翻译氨基酸的不同种可能性; “开放”指完整基因序列中用于翻译氨基酸的区域。ORF 有 3 种不同定义, Sieber 等[20] 将不同定义研究比较, 认为“长度能被 3 整除、以终止密码子为界”的定义更适用于真核和原核生物, 如图 2 所示, ORF 的这种定义对剪切位点没有影响。ORF 表征核酸的特征包括最长开放阅读框的长度, 最长开放阅读框在整个序列中的占比, 序列是否完整包含起始和终止密码子等。

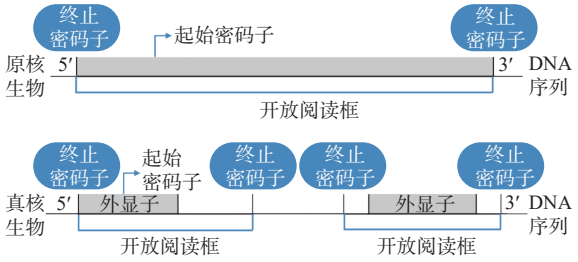

图 2　开放阅读框(ORF)定义示意图

此外, 还有多种基于序列信息的核酸表征, 例如基于核苷酸出现频率的 k-mer 特征表征[21], 基于 RNA 序列短基序频率和自然语言信息熵理论描述核酸序列的熵密度分布特征表征[22], 终止密码子特征表征和 GC 含量特征表征等。

1.3.2　基于理化性质的表征　稀疏编码通过与核酸理化性质联系, 实现核酸特征表征。Meher 等[23] 把核苷酸编码为三维向量(x, y, z), 3 个坐标分别根据核苷酸是嘌呤或嘧啶、是胺衍生物或酮衍生物、是强氢键或弱氢键相互作用, 用 1 或 0 表示, 即编码 A、T、G、C 的三维向量分别是(1,1,1):(0,0,1):(1,0,0):(0,1,0)。

伪蛋白是指, 针对非编码 RNA, 根据氨基酸翻译规则人为翻译的非自然肽链。Yang 等[24] 研究表明, 伪蛋白序列和理化性质与构成真实蛋白的肽链有一定差异, 据此伪蛋白特征可以对核酸进行表征。Cock 等[25] 研究揭示, 伪蛋白表征核酸的特征包括: 相对分子质量、等电点、等电点与相对分子质量比值的对数、氨基酸亲水平均值、不稳定指数等。

另外，常见的基于理化性质的核酸表征还有双核苷酸自相关性特征表征、伪二核苷酸组成特征表征、电子-离子相互作用赝势特征表征等。

1.3.3 基于二级、三级结构的表征 One-hot 编码策略也可以用于编码 RNA 二级结构特征实现核酸表征[26]。Danaee 等[27] 开发的工具 bpRNA 就可以通过解析 RNA 二级结构特征信息得到通用注释，其解析的二级结构包括：茎(S)、段(X)、凸起(B)、发夹环(H)、内部环(I)、外部环(E)、多环(M)等。基于此，one-hot 编码策略就可以编码二级结构[28]，例如茎(S)编码为$(1,0,0,0,0,0,0)^{T}$，而后就可以将 RNA 编码为 7 行多列的矩阵。此外，还有多种基于二级结构特征的表征方法。例如，基于二级结构的保守性，对 RNA 与其同家族 RNA 存在同源结构的情况进行打分和表征[24]；根据转录本的最小自由能、或根据配对碱基数量和未配对碱基数量等二级结构描述符进行特征表征[28]；Han 等[29] 还综合上述多种特征，提出了多尺度的二级结构信息表征核酸，包括基于稳定性的低层次特征、基于碱基配对情况的中层次特征和基于结构-核苷酸序列的高层次特征；另外，还有基于 DL 方法的 RNA 结构预测工具 UFold[30] 也使用二级结构进行核酸表征[31]。

当前基于核酸三级结构的表征方法较少，大多从蛋白质结构表征方法迁移而来，它们在核酸三级结构表征中同样适用，主要有以下几种：(1)距离矩阵，通过三级结构中各个核苷酸之间的物理距离进行特征表征；(2)拓扑协同特征，三级结构具有环状结构、折叠等复杂的空间关系，可以通过计算环系统的数量和类型等不同级别的拓扑特征来进行核酸表征；(3)嵌入技术，利用自编码器等 DL 模型，可以直接从三级结构数据中学习得到低维表示，这种表征方法可以自动捕捉数据中的复杂模式，不需要手动设计特征。随着人工智能技术的全面发展以及对蛋白结构领域相对成熟的表征方法的经验借鉴，核酸三级结构表征方法必将有所突破。

2 人工智能在核酸药物研发中的应用与进展

目前，人工智能在核酸药物研发领域的应用重点集中在核酸结构预测和小核酸药物设计方面。人工智能先进的算法和模型推进了核酸基础研究进程，加速了小核酸药物的发现，极大地提高了研发效率。

2.1 人工智能在核酸结构预测中的应用

2.1.1 核酸结构预测概述 核酸结构预测是指，通过已知核酸序列，预测其二级、三级，甚至四级结构。自 1978 年第 1 个核酸分子晶体结构被解出到现在，由于核酸骨架动态变化的灵活性和带负电磷酸基团在结晶中的互斥作用[32]，解析核酸晶体结构仍比解析蛋白质困难得多，因此，推进核酸结构预测研究具有极高必要性。当前，核酸结构预测方法可以分为经典计算方法和基于 ML 的人工智能方法。

2.1.2 基于经典计算的核酸结构预测方法 基于经典计算的方法可以分为核酸二级结构预测方法和三级结构预测方法。

(1) 二级结构预测

核酸二级结构预测的最底层原理是碱基互补配对和碱基堆积力。由于生物体内核酸的碱基互补配对原则，DNA 普遍形成双螺旋结构，RNA 则会形成茎环、假结、三叶草结构等更复杂多样的二级结构。因此，现有研究侧重于 RNA 二级结构的精准预测，主要分为单序列预测和多序列对比预测。

当前单序列预测方法主要有最小自由能方法、动态规划算法和基于抽样统计的方法。最小自由能方法[33] 认为真实世界的核酸二级结构必定拥有最小的或偶尔次最小的吉布斯自由能，例如基于 Zuker-Stiegle 算法的 RNAfold 和 Mfold，其中 Mfold 还加入了酶切位点、化学反应性等数据来提高预测准确性，这一类算法的缺点是在处理长序列核酸时耗费较大。动态规划算法[34] 利用递归和记忆化技术来预测核酸序列碱基配对最大化的结构，例如 pknotsRG，这类算法不用穷举序列总长度的所有可能结构，只需从最短片段出发，确定所有可能结构中自由能最低的构象然后保存，继而逐步延伸到更长片段，但这类算法在预测假结结构时会受限。抽样统计方法[35] 将输入核酸序列的所有二级结构经过 Bolzmann 分布处理后，使用分配函数条件概率随机抽样并进行聚类，最后得到代表性结构，例如 Sfold，这类方法也可以用于小核酸药物的理性设计。

多序列对比法的开发是为了解决单序列方法中的热力学参数不完全符合真实体内环境、忽略核酸碱基的化学修饰、未包含核酸共进化信息等[36] 局限性问题。多序列对比法基于客观事实：生物体进化过程中，核酸的二级、三级结构保守程度大于一

级序列[37]。当一个碱基突变时, 这个区域势必进一步再次突变补偿互作体系, 维持结构稳定性。因此多序列对比法需要多条同源核酸序列的数据集。这类方法可以根据对比和预测的顺序进行分类。其中, 先对比后预测的方法先通过多序列对比产生多条核酸对比结果, 之后寻找保守序列, 再使用单序列方法将其折叠成共有结构, 这类方法包括结合最小自由能和共进化信息的 RNAalifold、基于随机上下文无关文法的 Pfold、基于随机上下文无关文法和热力学的 PETfold、结合热力学和交互信息内容打分的 ILM; 同时对比和预测的方法大多基于限制性 Sankoff 算法, 由序列对比和动态规划算法结合而成, 这类方法包括 Foldalign、Dynalign、PMcomp; 先预测后对比的方法使用相对并不广泛, 使用场景大多是在序列保守性信息完全缺失时, 先使用单序列方法预测核酸结构, 再运用简单树匹配算法对齐结构, 这类方法首先需要确保单序列预测的结构足够准确, 才能再进行后续分析, 包括 RNA forester、MARNA 等[37]。

（2）三级结构预测

三级结构预测方法主要在二级结构的基础上预测双螺旋以外区域的结构, 目前可以分为使用同源模板的方法、仅基于量子力学的方法和基于片段拼合的方法。同源建模方法[38]认为同源核酸分子虽然序列有差异, 但都生成相似的三级结构, 因此找到同源模板即可进行建模预测, 这种方法包括 ModeRNA、SWISS-MODEL 等, 它们的优点是方便引入用户自定义的各种限制。从头建模方法[39]则不需要模板, 这种方法仅依靠最基础的量子力学等物理学定律来模拟核酸构象变化, 也因此在构象采样和计算能量等步骤中具有耗费大量计算资源的缺点, 这种方法包括 NAST、Vfold、DMD、SimRNA 等。片段拼合方法[40]从已知的三级结构中切割片段, 在二级结构等信息的指导下对片段进行组合, 最后还可以使用基于物理或统计的打分函数评价组合的结构, 这种方法包括 RNA2D3D、RNA Composer 等。

2.1.3　基于 ML 的核酸结构预测方法　在过去的二十年中, 预测计算精度和速度并没有显著提高。直到近年来 RNA 序列数据的爆炸式增长以及 ML 技术的进步, 最新的基于 ML 的方法在准确性和适用性方面超越了经典计算方法[41]。根据参与核酸结构预测环节的不同, 基于 ML 的核酸结构预测方法可以分为基于 ML 的打分方案, 基于 ML 的预处理和后处理, 以及基于 ML 的全过程预测。其中所有基于 ML 的方法都以 SL 的方式训练模型[42], 根据已知的成对的输入输出, 通过调整模型参数来学习, 将输入的特征映射到输出的函数。其中许多算法使用自由能参数、编码的 RNA 序列、序列模式或进化信息作为关键特征, 其结果可以是输出碱基是否成对的分类标签或自由能的连续值[42]。

基于 ML 的方法通常训练一个 ML 模型, 生成新的评分方案取代传统方法中的评分方案, 方法框架如图 3 所示。根据得分含义, 基于 ML 的评分方案可以分为自由能参数优化方法、加权方法、概率方法。其中自由能参数优化方法是当前最流行的方法[41]。近几年一些 ML 技术已经对能量模型中的参数进行了细化, 利用已知的热力学数据或 RNA 二级结构数据, 得到更丰富准确的特征表示。Xia 等[43]首先使用已知的热力学数据训练了一个线性回归模型 INN-HB, 来推断一些热力学参数。然而这种方法会导致在计算其他参数之前, 一些结构元素的范围已经被固定, 限制了参数集整体考虑的可能性范围。为了克服这个问题, Andronescu 等[44]提出了约束生成方法来估计自由能参数, 使用不同类型约束确保参考结构比同一序列的备选方案能量更低。该方法获得的 F-score 比标准 Turner 参数高 7%。随后他们进一步改进方法, 使用更大的数据集[45]提出了一个损失增强的 LAM-CG 模型和 Boltzmann-likelihood 模型, 能够做到对参数施加约束, 当结构越不准确, 其自由能与训练集中参考结构自由能之

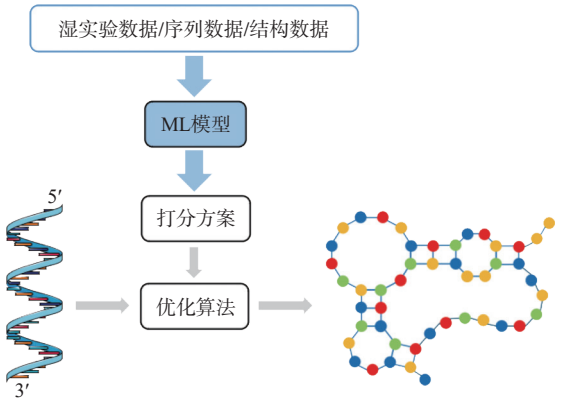

图 3　基于 ML 打分方案的核酸二级结构预测方法框架。可以使用湿实验室数据、核酸序列数据或核酸结构数据来训练 ML 模型以获得打分方案

间的差值越大。此外，自由能参数优化方法所确定的参数是热力学性质的，可以直接嵌入其他能量模型的算法中，如 miRNA 靶标预测算法[46] 和 RNA 折叠动力学模拟算法[47]。

ML 也可用于预处理或后处理，方法框架如图 4 所示，预处理时可以用于选择合适预测方法或参数。Hor 等[48] 提出了一种基于支持向量机的工具，选择预测方法的依据是：不同 RNA 序列具有不同特征，每种预测方法在特定不同的 RNA 物种中效果最好。Zhu 等[49] 假设不同 RNA 序列遵循不同折叠规则，提出一种基于随机上下文无关文法的模型，用于在 RNA 二级结构预测之前识别最可能的折叠规则。由于不同预测方法会返回不同结构，ML 模型用于后处理可以确定预测结果中最可能的结构。Andrews 等[50] 结合图论使用决策树方法表示 RNA 图形结构，而后使用图形不变量作为输入特征，训练了一个多层感知器判断其是否为 RNA-like 结构。

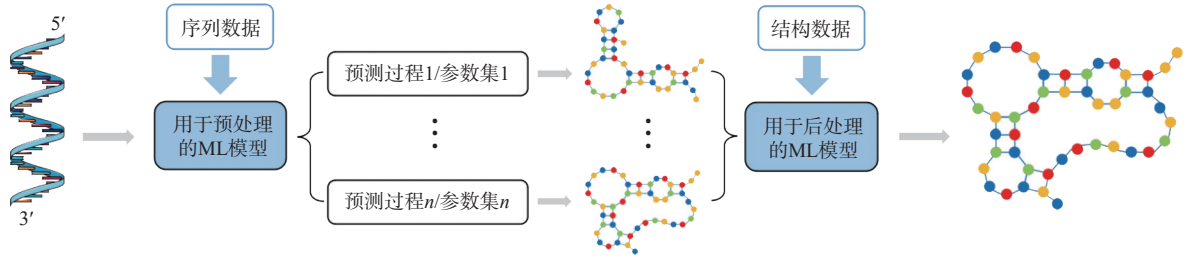

图 4 基于 ML 预处理或后处理的核酸二级结构预测方法框架。在核酸二级结构预测中，由序列数据训练得到的 ML 模型可用于预处理，用来选择合适的预测方法或一组合适的参数；由结构数据训练得到的 ML 模型也可以提供一种方法来确定预测结果中最可能的结构

ML 还可以直接参与核酸结构的预测过程，实现从序列到结构的端到端（end-to-end）预测，预测方法框架如图 5 所示。Singh 等[51] 开发的 SPOT-RNA 是第一个预测 RNA 二级结构的基于 DL 的 end-to-end 模型，将 RNA 序列接触矩阵作为输入，采用 CNN、二维双向 LSTM、全连接层模块组成的混合深度网络，先经过 14565 个非冗余 RNA 数据集训练，而后在 226 个高精度 RNA 结构上进行迁移学习。多个外部测试集的测试表明，SPOT-RNA 的 RNA 结构预测性能显著优于基于评分的方法，且能用于预测非经典、非嵌套碱基配对。SPOT-RNA2[52] 使用进化驱动的序列数据和突变耦合作为网络输入，同样使用迁移学习策略，得到比第一代更好的性能。序列信息之外，核酸的形状数据、共进化数据等也可以融入 DL 模型[53−54]。

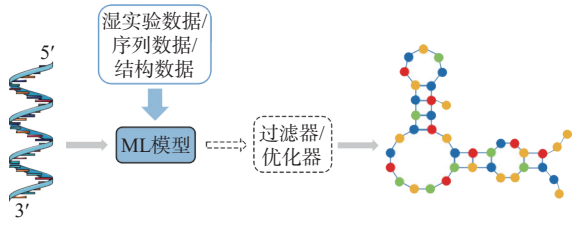

图 5 基于 ML 全过程预测的核酸二级结构预测方法框架。通过湿实验数据、核酸序列数据或核酸结构数据进行训练 ML 模型以端到端的方式直接用于预测核酸二级结构，还可以加入过滤器或优化器获得最优的核酸二级结构。

在基于 ML 的核酸三级结构预测中，Townshend 等[55] 仅从 18 个 RNA 结构出发，构建旋转和平移等变性 DNN 模型，训练得到的原子旋转等变评分器大幅提升了 RNA 结构全盲预测的准确性。由于输入参数仅仅是原子坐标和原子类型，不包含 RNA 结构其他信息，此方法可以推广到结构生物学、化学和材料学等领域。Wang 等[56] 开发的 3dRNA 使用最小二级元素，通过模板方法从 RNA 序列和二级结构中建模 RNA 三级结构。Wang 等[57] 开发的 trRosettaRNA 包括两个主要步骤，通过 Transformer 网络进行一维二维几何形状预测，以及通过能量最小化进行的三维结构折叠。还有其他模型如 MELD-DNA[58] 和 RoseTTAFoldNA[59] 实现了核酸和蛋白质复合物结构的准确预测。

2.2 人工智能在小核酸药物设计中的应用

2.2.1 小核酸药物设计概述 小核酸药物由治疗特定疾病、诱导特定功能的核苷酸序列构成，通常是核苷酸数小于 30 的短链 RNA，通过与靶标 mRNA 形成相互作用来调节基因或抑制沉默基因。根据小核酸药物设计模式，可以分为靶向核酸设计、靶向蛋白质设计和编码蛋白质设计[1]。

针对靶向核酸进行设计的小核酸药物主要有

单链反义寡核苷酸(antisense oligonucleotides, ASOs)和小干扰 RNA(small interfering RNA, siRNA)。ASOs 是人工合成的寡核苷酸, 通常长度为 12~30 个核苷酸。ASOs 作用机制主要有两种[1]: 与靶 mRNA 结合, 使靶 mRNA 更容易被核酸水解酶识别和降解; 调节 RNA 剪接, 与靶 mRNA 高特异性结合, 通过空间位阻效应调控基因转录过程, 影响靶 mRNA 正常剪接过程。siRNA 则是双链 RNA, 比 ASOs 更难进入细胞, 通常有 20~27 个碱基对, 有抑制基因表达的作用[1]。RNA 干扰分为 3 个阶段: 起始阶段 dsRNA 被 RNA 酶 Dicer 等切割成 siRNA; 效应阶段 siRNA 和内切核酸酶一起形成 RNA 诱导沉默复合物(RNA-induced silencing complex, RISC), siRNA 解链, 正义链降解, 反义链与靶 mRNA 结合, 而后 RISC 的切割蛋白 Ago-2 降解靶 mRNA; 扩增阶段 siRNA 作为 RNA 引物在 RNA 聚合酶作用下再次形成 dsRNA, 循环往复。

靶向蛋白质的小核酸药物设计使用核酸适配体。核酸适配体由 20~50 个核苷酸组成, 可与蛋白质特定位点结合, 调节其功能。核酸适配体起效快速、作用可逆, 能够辅助麻醉调节凝血, 有望在外科手术和急诊科室中发挥作用。

mRNA 作为小核酸药物可以生成需要的蛋白质。mRNA 注射进体内被细胞吸收识别后, 会启动蛋白质合成程序, 生成所需蛋白质[60]。目前这类小核酸药物主要应用在疫苗领域, 例如预防传染病的 mRNA 疫苗表达感染性病原体抗原, 诱导强效细胞体液免疫应答; 癌症 mRNA 疫苗用来表达肿瘤相关抗原, 刺激细胞免疫清除癌细胞。

目前共有 15 款小核酸药物上市, 但由于小核酸药物主要作用于细胞内靶点, 其设计具有许多挑战, 包括: 易被非特异性 RNA 酶水解; 常带负电荷, 很难穿过细胞膜; 易被生物体防御系统识别, 导致急性免疫反应, 甚至炎症因子风暴等。因此, 开发安全高效的小核酸药物设计方法需要人工智能的助力。基于人工智能的小核酸药物设计方法主要分为: 基于 ML 的设计方法和基于 DL 的设计方法, 后者又可以分为基于 DNN 的方法、基于 DRL 的方法和基于 DGM 的方法。

2.2.2　*基于 ML 的小核酸药物设计方法*　基于 ML 的设计方法主要包括动态探索、模拟退火、约束规划、多目标元启发式算法、支持向量机等方法。

动态探索属于初始算法之一。Churkin 等[61] 开发的 RNAInverse 模型使用简单的自适应游走, 通过比较变异 RNA 序列的最小能量折叠与目标结构来计算碱基对之间的距离, 实现碱基对距离的最小化。Andronescu 等[62] 开发的 RNA-SSD 算法首先将结构分为子结构再进行自适应游走, 以便减小搜索空间的大小。Hampson 等[63] 提出新的动态探索策略, 在算法早期减少折叠次数, 尝试不增加运行时间, 探索更多样的设计空间, 这种方法的缺点是没有特定的最佳参数集。Busch 等[64] 研发的 INFO-RNA 首先使用动态规划生成序列, 估计目标结构的最小能量序列, 然后使用模拟退火执行随机搜索。Taneda 等[65] 的 MODENA 使用遗传算法生成初始集合, 然后使用交叉移动, 同一位置的两个候选解相互交换, 或单点突变执行随机搜索, 最后使用对所得序列的结构稳定性和相似性打分的目标函数判断集合。百度公司研发的 LinearDesign[66] 是 mRNA 序列优化模型, 通过动态规划算法, 联合优化序列稳定性和密码子翻译效率指标, 设计具有最佳折叠稳定性和密码子的 mRNA。LinearDesign 突破了高稳定性设计瓶颈, 可以优化编码单克隆抗体等所有治疗性蛋白的 mRNA。Zhao 等[67] 使用 LinearDesign 设计具有最佳折叠稳定性和密码子使用的 mRNA, 开发了一种用于 CHB 治疗的编码乙型肝炎表面抗原的 LNP-mRNA 疫苗, 兼具强免疫原性和持续抗病毒作用。

模拟退火是一种概率优化算法, 适用于具有大量局部最优解的复杂优化问题。设计核酸序列的模拟退火方法包括 SIMARD[68]、ERD[69] 和 RNAPredict[70], 这些方法都旨在返回折叠后接近目标结构的 RNA 序列。这类方法有以下特点[71]: 缺点包括只使用单一模拟退火冷却, 最多只涵盖两个模拟退火变体的四个 RNA 设计问题, 几何调度参数的不敏感; 优点包括对数冷却调度可以解决其他调度无法解决的 RNA 设计问题, 可以识别 RNA 设计自适应和非自适应调度中的常见问题等。

Minuesa 等[72] 开发了基于约束编程的核酸设计算法 MoiRNAiFold, 包括新变量类型、启发式和大邻域搜索的重启策略, 可以处理数十种设计约束和质量措施, 并改进了如翻译效率计算等基因表达的核糖核酸调节控制功能, 但不能预测趾扣开关结构的功效。MoiRNAiFold 专注于 RNA 核糖调节

体，所设计的 RNA 序列在体外和体内都具有功能，为从头生成复杂 RNA 设计提供了一个强大工具。

Rubio-Largo 等[73]设计了多目标元启发式算法 m2dRNAs，考虑了目标和预测结构之间的相似性作为约束，以及 3 个目标函数：系综自由能配分函数；整体多样性函数；核苷酸组成函数。因此 m2dRNAs 可以提供稳定的 RNA 序列，并确保结构预测的可靠性，避免过度偏差。m2dRNAs 与 RNAinverse、RNA-SSD、INFO-RNA、MODENA、NUPACK、fRNAkenstein、RNAiFOLD 等其他已发表的 RNA 逆折叠方法进行了比较，发现其性能优于当前其他方法。

Chiba 等[74]开发了目前最大的反义核酸药物数据库 eSkip-Finder，首个基于 ML 方法预测外显子跳跃效率的工具，使用了支持向量回归器，为抗肌萎缩蛋白 mRNA 靶外显子的相对跳跃功效构建了一个预测模型。eSkip-Finder 收集了外显子跳跃药物的序列、活性等信息数据，目的是研发外显子跳跃治疗遗传神经和肌肉疾病的新药。

2.2.3 基于 DNN 的小核酸药物设计方法 基于 DNN 的设计方法包括 CNN、深度卷积去噪神经网络（deep convolutionary denosing neural network, DCDNN）、CNN 与 LSTM 结合、LSTM 等方法。

Han 等[75]利用 DL 算法 CNN 开发了一种预测器用于 siRNA 设计，并探索不同基序对基因沉默的影响，在 CNN 模型的卷积层中，将卷积核设计为基序检测器，自动学习 siRNA 多模基序的潜在特征模式。这一类特征更抽象，对分类更有利。测试结果表明该模型的 Pearson 相关系数为 0.717，比 Biopredsi、DSIR 和 siRNApred 算法分别高出 13.81%、16.78% 和 5.91%。因此，模型可以探索 siRNA 多模基序对疗效的贡献，并获取序列局部特征中有价值信息的特征模式。

Chuai 等[76]提出了一个计算平台 DeepCRISPR，基于 DCDNN 的自动编码器设计向导 RNA（small guide RNA, sgRNA）。DeepCRISPR 将 sgRNA 靶上预测和靶外预测进行统一，自动化识别可能影响 sgRNA 敲除效果的序列和表观遗传特征。CRISPR 可用于同时预测 sgRNA 靶向敲除效果和全基因组脱靶谱；DeepCRISPR 则可以解释和优化 CRISPR 的目标内、外设计。

Tasdelen 等[77]提出了一种基于 pre-miRNA 顺序结构和空间结构的混合 DL 方法，通过集成 CNN 和 LSTM 这两种不同的神经网络，来实现 pre-miRNA 的分类任务。CNN 自动从输入数据中提取特征，从而解决了手动特征提取的问题。在卷积输入数据的 CNN 层之后，LSTM 层则用于执行时间建模。该方法使用 Keras 库在 Python 中实现，模型后端为 TensorFlow。

Im 等[78]开发了一个生成模型，利用 LSTM 构建与目标蛋白结合的单链核酸，其生成的多个目标蛋白的 DNA 和 RNA 序列具有很高的特异性，生成序列中的基序与已知的蛋白质结合基序相似。此方法可用于生成与靶蛋白结合的核酸序列，特别是还可以用于构建具有高亲和力和特异性的与目标蛋白结合的潜在适配体初始池，有助于设计高效体外实验。

2.2.4 基于 DRL 的小核酸药物设计方法 基于 DRL 的设计方法当前也有所进展。RNA 结构特性决定功能，因此小核酸药物设计挑战之一是识别 RNA 中导致其折叠成特定结构的模式和序列，此过程称为 RNA 反向折叠。Runge 等[79]采用 DRL 算法 LEARNA，使用奖励机制驱动算法。在给定目标结构的情况下，算法按顺序设计整个 RNA 序列，在 20 个 CPU 核上对 65 000 个不同 RNA 设计任务进行 1 h 的 Meta 学习后，可以得到扩展程序 Meta-LEARNA。Meta-LEARNA 学习了许多 RNA 设计问题的单一策略，通过使用贝叶斯优化方法解决架构搜索和超参数优化问题，在神经结构空间中对策略网络、训练过程中的超参数和决策过程中的制定来进行联合优化，可以适用于新的 RNA 设计。Eastman 等[80]设计了一种 RL 算法，给定目标的二级结构，在算法内部设计一个可以折叠到该结构的新序列。它采用了一种高级图卷积架构，允许将单个模型应用于任意长度的任意目标结构，在对随机生成的目标进行训练后，在 Eterna100 基准上进行了测试，发现它的性能优于目前所有其他算法。

2.2.5 基于 DGM 的小核酸药物设计方法 目前，相较于基于 DGM 的小分子药物设计方法，基于 DGM 的小核酸药物设计方法还相对较少。Iwano 等[81]开发了 RaptGen，一种用于生成核酸适配体的变分自编码器模型。RaptGen 利用一个轮廓隐藏的马尔可夫模型解码器来有效地表示 motif 序列，在 motif 信息的基础上将模拟序列数据嵌入到

低维潜在空间中, 并使用两个独立的核酸适配体数据集进行了序列嵌入, 成功地从潜在空间生成了适配体, RaptGen 还可以生成一个截断的适配体, 并且可以根据贝叶斯优化应用于活性引导的适配体生成。此外, 在更广泛的核酸研究领域, DGM 的应用也在逐渐增多。Sumi 等[82] 开发了 RfamGen, 一种利用变分自编码器框架结合协变模型架构的 DGM 模型, 可以高效地设计 RNA 家族序列。Gupta 等[83] 提出了一个基于生成式对抗网络的产生 DNA 序列的反馈-循环机制, 可以生成编码抗菌肽的合成基因, 以及优化合成基因的二级结构。Linder 等[84] 开发了 Fast SeqProp, 它通过可微适应度预测器进行高效通用序列优化, 可以结合变异自动编码器等多种正则化技术, 保持序列设计的置信度。这些更广泛的核酸研究可以为开发新的小核酸药物设计方法提供启发。

3　总结与展望

人工智能正在赋能核酸药物研发越来越多的环节, 但总体仍处于起步阶段, 还面临着许多挑战。例如对于核酸二级结构预测模型, 过拟合是一个非常重要的问题。过拟合模型在与训练数据相似的测试 RNA 上表现良好, 但泛化能力较差。模型只记住了训练集核酸的二级结构, 并没有学习折叠机制。Sato 等[85] 的研究表明, E2Efold 在新发现的 RNA 家族上表现不佳, 可能存在严重过拟合现象。同样, Rivas 等[86] 报道, 在与训练集结构不相似的一组 RNA 上测试时, ContextFold 的 F-score 也降低了 24%。又如 siRNA 药物设计有一个环节是 siRNA 脱靶效应的预测, 即规避 siRNA 对非靶基因的影响。但当前 siRNA 设计模型中进行脱靶预测工作的最常用的基于序列对比原理的工具 BLAST 会不可避免地忽略部分潜在脱靶情况。尽管一直以来有许多基于热力学等其他原理的 siRNA 脱靶效应预测模型被开发, 如 Picky[87]、RIsearch2[88]、Batch RNAi selector[89] 等, 但它们的预测性能并没有显著超越 BLAST, 目前仍有待进一步开发性能更好、预测全面的集成模型。

挑战与机遇并存。例如在核酸药物递送领域, 人工智能的应用就相对较少。可电离脂质纳米颗粒(lipid nanoparticles, LNP)是核酸递送的主流技术, 已经在 mRNA 新冠疫苗研发中得到产业化验证。Xu 等[90] 开发了 AI-Guided Ionizable Lipid Engineering(AGILE)平台, 首次利用 DL 和组合化学的协同来探索可离子化脂质分子搜索空间, 通过预训练的 DNN 方法学习大量小分子化合物的结构知识, 利用自我监督的方法学习辨别区分脂质结构。经过微调和高通量筛选, AGILE 能够准确识别具有较高 mRNA 转染效力的新型脂质结构。这项研究首次证明了 DL 在加速和定制化 LNP 开发方面的潜力。同时, 核酸药物治疗领域的蛋白替代疗法、再生疗法等新疗法正在带来丰富的研究需求, 有待应用人工智能技术实现突破性进展。

随着人工智能新技术、生物实验新数据、核酸表征新方法的发展, 人工智能与核酸药物研发这个新兴交叉领域一定能释放巨大潜能, 成为新药研发的重要支柱和核心技术。

References

[1] DeWeerdt S. RNA therapies explained[J]. *Nature*, 2019, **574**(7778): S2-S3.

[2] Cochrane G, Karsch-Mizrachi I, Takagi T, *et al*. The international nucleotide sequence database collaboration[J]. *Nucleic Acids Res*, 2016, **44**(D1): D48-D50.

[3] Zardecki C, Duarte JM, Bi C, *et al*. RCSB PDB next-generation data delivery and search services[J]. *Acta Crystallogr A*, 2020, A**76**: a70.

[4] Romero PR, Kobayashi N, Wedell JR, *et al*. BioMagResBank (BMRB) as a resource for structural biology[J]. *Methods Mol Biol*, 2020, **2112**: 187-218.

[5] Rigden DJ, Fernández XM. The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection[J]. *Nucleic Acids Res*, 2023, **51**(D1): D1-D8.

[6] Benson DA, Cavanaugh M, Clark K, *et al*. GenBank[J]. *Nucleic Acids Res*, 2018, **46**(D1): D41-D47.

[7] Yuan D, Ahamed A, Burgin J, *et al*. The European nucleotide archive in 2023[J]. *Nucleic Acids Res*, 2024, **52**(D1): D92-D97.

[8] Ara T, Kodama Y, Tokimatsu T, *et al*. DDBJ update in 2023: the MetaboBank for metabolomics data and associated metadata[J]. *Nucleic Acids Res*, 2024, **52**(D1): D67-D71.

[9] Haft DH, Badretdin A, Coulouris G, *et al*. RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes[J]. *Nucleic Acids Res*, 2024, **52**(D1): D762-D769.

[10] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function[J]. *Nucleic Acids Res*, 2019, **47**(D1): D155-D162.

[11] The RNAcentral Consortium. RNAcentral: a hub of information for non-coding RNA sequences[J]. *Nucleic Acids Res*, 2019, **47**(D1): D221-D229.

[12] Coimbatore Narayanan B, Westbrook J, Ghosh S, *et al*. The Nucleic Acid Database: new features and capabilities[J]. *Nucleic Acids Res*, 2014, **42**(D1): D114-D122.

[13] Berman HM, Lawson CL, Schneider B. Developing community resources for nucleic acid structures[J]. *Life*, 2022, **12**(4): 540.

[14] Zanegina O, Kirsanov D, Baulin E, *et al*. An updated version of NPIDB includes new classifications of DNA-protein complexes and their families[J]. *Nucleic Acids Res*, 2016, **44**(D1): D144-D153.

[15] Norambuena T, Melo F. The Protein-DNA interface database[J]. *BMC Bioinformatics*, 2010, **11**: 262.

[16] Sagendorf JM, Markarian N, Berman HM, *et al*. DNAproDB: an expanded database and web-based tool for structural analysis of DNA-protein complexes[J]. *Nucleic Acids Res*, 2020, **48**(D1): D277-D287.

[17] Lewis BA, Walia RR, Terribilini M, *et al*. PRIDB: a Protein-RNA interface database[J]. *Nucleic Acids Res*, 2011, **39**(D1): D277-D282.

[18] Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning[J]. *Nat Biotechnol*, 2015, **33**(8): 831-838.

[19] Pan XY, Rijnbeek P, Yan JC, *et al*. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks[J]. *BMC Genomics*, 2018, **19**(1): 511.

[20] Sieber P, Platzer M, Schuster S. The definition of open reading frame revisited[J]. *Trends Genet*, 2018, **34**(3): 167-170.

[21] Kirk JM, Kim SO, Inoue K, *et al*. Functional classification of long non-coding RNAs by k-mer content[J]. *Nat Genet*, 2018, **50**(10): 1474-1482.

[22] Liu YC, Guo JT, Hu GQ, *et al*. Gene prediction in metagenomic fragments based on the SVM algorithm[J]. *BMC Bioinformatics*, 2013, **14**(Suppl 5): S12.

[23] Meher PK, Sahu TK, Gahoi S, *et al*. Evaluating the performance of sequence encoding schemes and machine learning methods for splice sites recognition[J]. *Gene*, 2019, **705**: 113-126.

[24] Yang S, Wang Y, Zhang SQ, *et al*. NCResNet: noncoding ribonucleic acid prediction based on a deep resident network of ribonucleic acid sequences[J]. *Front Genet*, 2020, **11**: 90.

[25] Cock PJ, Antao T, Chang JT, *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics[J]. *Bioinformatics*, 2009, **25**(11): 1422-1423.

[26] Song JM, Tian SW, Yu L, *et al*. MD-MLI: prediction of miRNA-lncRNA interaction by using multiple features and hierarchical deep learning[J]. *IEEE/ACM Trans Comput Biol Bioinform*, 2022, **19**(3): 1724-1733.

[27] Danaee P, Rouches M, Wiley M, *et al*. bpRNA: large-scale automated annotation and analysis of RNA secondary structure[J]. *Nucleic Acids Res*, 2018, **46**(11): 5381-5394.

[28] Blumenthal DM, Singal G, Mangla SS, *et al*. Predicting non-adherence with outpatient colonoscopy using a novel electronic tool that measures prior non-adherence[J]. *J Gen Intern Med*, 2015, **30**(6): 724-731.

[29] Han SY, Liang YC, Ma Q, *et al*. LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property[J]. *Brief Bioinform*, 2019, **20**(6): 2009-2027.

[30] Fu LY, Cao YX, Wu J, *et al*. UFold: fast and accurate RNA secondary structure prediction with deep learning[J]. *Nucleic Acids Res*, 2022, **50**(3): e14.

[31] Vicens Q, Kieft JS. Thoughts on how to think (and talk) about RNA structure[J]. *Proc Natl Acad Sci U S A*, 2022, **119**(17): e2112677119.

[32] Lin LN, Sheng J, Huang Z. Nucleic acid X-ray crystallography via direct selenium derivatization[J]. *Chem Soc Rev*, 2011, **40**(9): 4591-4602.

[33] Zuker M. On finding all suboptimal foldings of an RNA molecule[J]. *Science*, 1989, **244**(4900): 48-52.

[34] Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA[J]. *Proc Natl Acad Sci U S A*, 1980, **77**(11): 6309-6313.

[35] McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure[J]. *Biopolymers*, 1990, **29**(6/7): 1105-1119.

[36] Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches[J]. *BMC Bioinformatics*, 2004, **5**: 140.

[37] Aigner K, Dreßen F, Steger G. Methods for predicting RNA secondary structure[M]//Leontis N, Westhof E. *RNA 3D Structure Analysis and Prediction*. Berlin, Heidelberg: Springer, 2012: 19-41.

[38] Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems[J]. *SIAM J Appl Math*, 1985, **45**(5): 810-825.

[39] Šponer J, Bussi G, Krepl M, *et al*. RNA structural dynamics As captured by molecular simulations: a comprehensive overview[J]. *Chem Rev*, 2018, **118**(8): 4177-4338.

[40] Martinez HM, Maizel JV Jr, Shapiro BA. RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA[J]. *J Biomol Struct Dyn*, 2008, **25**(6): 669-683.

[41] Zhao Q, Zhao Z, Fan XY, *et al*. Review of machine learning methods for RNA secondary structure prediction[J]. *PLoS Comput Biol*, 2021, **17**(8): e1009291.

[42] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects[J]. *Science*, 2015, **349**(6245): 255-260.

[43] Xia T, SantaLucia J Jr, Burkard ME, *et al*. Thermodynamic pa-

rameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs[J]. *Biochemistry*, 1998, **37**(42): 14719-14735.

[44] Andronescu M, Condon A, Hoos HH, *et al*. Efficient parameter estimation for RNA secondary structure prediction[J]. *Bioinformatics*, 2007, **23**(13): i19-i28.

[45] Andronescu M, Condon A, Hoos HH, *et al*. Computational approaches for RNA energy parameter estimation[J]. *RNA*, 2010, **16**(12): 2304-2318.

[46] Rehmsmeier M, Steffen P, Hochsmann M, *et al*. Fast and effective prediction of microRNA/target duplexes[J]. *RNA*, 2004, **10**(10): 1507-1517.

[47] Tang XY, Thomas S, Tapia L, *et al*. Simulating RNA folding kinetics on approximated energy landscapes[J]. *J Mol Biol*, 2008, **381**(4): 1055-1067.

[48] Hor CY, Yang CB, Chang CH, *et al*. A tool preference choice method for RNA secondary structure prediction by SVM with statistical tests[J]. *Evol Bioinform Online*, 2013, **9**: 163-184.

[49] Zhu Y, Xie ZY, Li YZ, *et al*. Research on folding diversity in statistical learning methods for RNA secondary structure prediction[J]. *Int J Biol Sci*, 2018, **14**(8): 872-882.

[50] Andrews D, Guggenberger P. Asymptotics for stationary very nearly unit root processes[J]. *J Time Ser Anal*, 2008, **29**(1): 203-212.

[51] Singh J, Hanson J, Paliwal K, *et al*. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning[J]. *Nat Commun*, 2019, **10**(1): 5407.

[52] Singh J, Paliwal K, Zhang TC, *et al*. Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning[J]. *Bioinformatics*, 2021, **37**(17): 2589-2600.

[53] Calonaci N, Jones A, Cuturello F, *et al*. Machine learning a model for RNA structure prediction[J]. *NAR Genom Bioinform*, 2020, **2**(4): lqaa090.

[54] Willmott D, Murrugarra D, Ye Q. Improving RNA secondary structure prediction via state inference with deep recurrent neural networks[EB/OL]. *arXiv*, 2019: 1906.10819. http://arxiv.org/abs/1906.10819.

[55] Townshend RJL, Eismann S, Watkins AM, *et al*. Geometric deep learning of RNA structure[J]. *Science*, 2021, **373**(6558): 1047-1051.

[56] Wang J, Wang J, Huang YZ, *et al*. 3dRNA v2.0: an updated web server for RNA 3D structure prediction[J]. *Int J Mol Sci*, 2019, **20**(17): 4116.

[57] Wang WK, Feng CJ, Han RM, *et al*. trRosettaRNA: automated prediction of RNA 3D structure with transformer network[J]. *Nat Commun*, 2023, **14**(1): 7266.

[58] Esmaeeli R, Bauzá A, Perez A. Structural predictions of protein-DNA binding: MELD-DNA[J]. *Nucleic Acids Res*, 2023, **51**(4): 1625-1636.

[59] Baek M, McHugh R, Anishchenko I, *et al*. Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA[J]. *Nat Methods*, 2024, **21**(1): 117-121.

[60] Chaudhary N, Weissman D, Whitehead KA. mRNA vaccines for infectious diseases: principles, delivery and clinical translation[J]. *Nat Rev Drug Discov*, 2021, **20**(11): 817-838.

[61] Churkin A, Retwitzer MD, Reinharz V, *et al*. Design of RNAs: comparing programs for inverse RNA folding[J]. *Brief Bioinform*, 2018, **19**(2): 350-358.

[62] Andronescu M, Fejes AP, Hutter F, *et al*. A new algorithm for RNA secondary structure design[J]. *J Mol Biol*, 2004, **336**(3): 607-624.

[63] Hampson DJD, Tsang HH. Incorporating dynamic exploration strategy for RNA design[C]//2018 IEEE Symposium Series on Computational Intelligence (SSCI). Bangalore, India. IEEE, 2018: 1041-1048.

[64] Busch A, Backofen R. INFO-RNA: a fast approach to inverse RNA folding[J]. *Bioinformatics*, 2006, **22**(15): 1823-1831.

[65] Taneda A. MODENA: a multi-objective RNA inverse folding[J]. *Adv Appl Bioinform Chem*, 2011, **4**: 1-12.

[66] Zhang H, Zhang L, Lin A, *et al*. Algorithm for optimized mRNA design improves stability and immunogenicity[J]. *Nature*, 2023, **621**(7978): 396-403.

[67] Zhao HJ, Shao XY, Yu YT, *et al*. A therapeutic hepatitis B mRNA vaccine with strong immunogenicity and persistent virological suppression[J]. *NPJ Vaccines*, 2024, **9**(1): 22.

[68] Sav S, Hampson DJD, Tsang HH. SIMARD: a simulated annealing based RNA design algorithm with quality pre-selection strategies[C]//2016 IEEE Symposium Series on Computational Intelligence (SSCI). Athens, Greece. IEEE, 2016: 1-8.

[69] Esmaili-Taheri A, Ganjtabesh M. ERD: a fast and reliable tool for RNA design including constraints[J]. *BMC Bioinformatics*, 2015, **16**: 20.

[70] Wiese K, Deschenes A, Hendriks A. RnaPredict: an evolutionary algorithm for RNA secondary structure prediction[J]. *IEEE/ACM Trans Comput Biol Bioinform*, 2008, **5**(1): 25-41.

[71] McBride R, Tsang HH. Examination of annealing schedules for RNA design[C]//2020 IEEE Congress on Evolutionary Computation (CEC). Glasgow, UK. IEEE, 2020: 1-8.

[72] Minuesa G, Alsina C, Garcia-Martin JA, *et al*. MoiRNAiFold: a novel tool for complex in silico RNA design[J]. *Nucleic Acids Res*, 2021, **49**(9): 4934-4943.

[73] Rubio-Largo Á, Vanneschi L, Castelli M, *et al*. Multiobjective metaheuristic to design RNA sequences[J]. *IEEE Trans Evol Comput*, 2019, **23**(1): 156-169.

[74] Chiba S, Lim KRQ, Sheri N, *et al*. eSkip-Finder: a machine learning-based web application and database to identify the op-

timal sequences of antisense oligonucleotides for exon skipping[J]. *Nucleic Acids Res*, 2021, **49**(W1): W193-W198.

[75] Han Y, He F, Tan X, *et al*. Effective small interfering RNA design based on convolutional neural network[C]//2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Kansas City, MO. IEEE, 2017: 16-21.

[76] Chuai GH, Ma HH, Yan JF, *et al*. DeepCRISPR: optimized CRISPR guide RNA design by deep learning[J]. *Genome Biol*, 2018, **19**(1): 80.

[77] Tasdelen A, Sen BH. A hybrid CNN-LSTM model for pre-miRNA classification[J]. *Sci Rep*, 2021, **11**(1): 14125.

[78] Im J, Park B, Han K. A generative model for constructing nucleic acid sequences binding to a protein[J]. *BMC Genomics*, 2019, **20**(Suppl 13): 967.

[79] Runge F, Stoll D, Falkner S, *et al*. Learning to design RNA[EB/OL]. *arXiv*, 2018: 1812.11951. http://arxiv.org/abs/1812.11951

[80] Eastman P, Shi J, Ramsundar B, *et al*. Solving the RNA design problem with reinforcement learning[J]. *PLoS Comput Biol*, 2018, **14**(6): e1006176.

[81] Iwano N, Adachi T, Aoki K, *et al*. Generative aptamer discovery using RaptGen[J]. *Nat Comput Sci*, 2022, **2**(6): 378-386.

[82] Sumi S, Hamada M, Saito H. Deep generative design of RNA family sequences[J]. *Nat Methods*, 2024, **21**(3): 435-443.

[83] Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions[J]. *Nat Mach Intell*, 2019, **1**: 105-111.

[84] Linder J, Seelig G. Fast activation maximization for molecular sequence design[J]. *BMC Bioinformatics*, 2021, **22**(1): 510.

[85] Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration[J]. *Nat Commun*, 2021, **12**(1): 941.

[86] Rivas E. The four ingredients of single-sequence RNA secondary structure prediction. A unifying perspective[J]. *RNA Biol*, 2013, **10**(7): 1185-1196.

[87] Chen X, Liu P, Chou HH. Whole-genome thermodynamic analysis reduces siRNA off-target effects[J]. *PLoS One*, 2013, **8**(3): e58326.

[88] Alkan F, Wenzel A, Palasca O, *et al*. RIsearch2: suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets[J]. *Nucleic Acids Res*, 2017, **45**(8): e60.

[89] Iyer S, Deutsch K, Yan XW, *et al*. Batch RNAi selector: a standalone program to predict specific siRNA candidates in batches with enhanced sensitivity[J]. *Comput Methods Programs Biomed*, 2007, **85**(3): 203-209.

[90] Xu Y, Ma SH, Cui HT, *et al*. AGILE platform: a deep learning-powered approach to accelerate LNP development for mRNA delivery[EB/OL]. *bioRxiv*, 2023. doi: 10.1101/2023.06.01.543345.

［专家介绍］余文颖，博士、研究员、博士生导师、国家高层次青年人才。江苏省抗衰老学会青年分会副主任委员、中国抗癌协会肿瘤重症医学专业委员会青年委员、美国化学会会员、美国药学会会员。2002–2006 年就读于中国药科大学国家生命科学与技术人才培养基地班，“4+2”本硕连读。2006 年免试推荐为中国药科大学药物化学硕士研究生，2008 年获理学硕士学位。同年获美国俄亥俄州立大学（The Ohio State University）药学院全额奖学金资助留学，2013 年获俄亥俄州立大学药物化学博士学位和统计学二学位，并获得 JACK. L. BEAL 优秀博士生奖学金。2013 年底作为海外人才引进于中国药科大学国家重点实验室。在药物化学领域一区杂志 *Journal of Medicinal Chemistry* 等发表多篇文章。授权世界、美国和中国专利十余项。