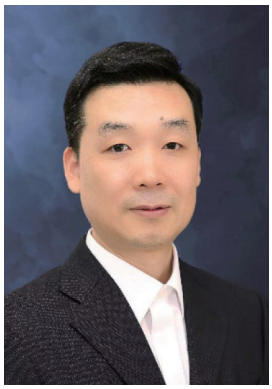
DOI: 10.13376/j.cbls/2021167
文章编号：1004-0374(2021)12-1483-10



乔宇，中科院深圳先进技术研究院研究员。从事计算机视觉、深度学习、机器人、生物信息等领域的研究开发。入选国家重点人才计划、科技部中青年科技创新领军人才等。发表学术论文200余篇，在JCR-Q1区期刊和CCF-A类会议发表论文90余篇；论文累计被引26 000余次，h-index为66。获授权发明专利40余项。以第一完成人获广东省技术发明一等奖、中科院卢嘉锡青年人才奖等，并获AAAI 2021杰出论文奖。

司同，中科院深圳先进技术研究院研究员，深圳合成生物研究重大科技基础设施(在建)总工艺师。研究方向为自动化合成生物技术。入选国家高层次人才(青年)、广东省“青年拔尖”等。课题组近年来主要研究工作包括：(1)全自动酵母基因组定向进化；(2)质谱成像用于细胞工厂高通量筛选。在*Nature Communications*、*JACS*等刊物发表论文40余篇，累计被引2 000余次，h-index为21。

合成生物数据库与大数据智能分析展望

胡如云#，陈永灿#，张建志，郭二鹏，付立豪，乔　宇*，司　同*
(中国科学院深圳先进技术研究院，深圳 518055)

摘　要：合成生物通过“设计 - 构建 - 测试 - 学习”闭环研究积累海量数据，推动合成生物数据储存、共享和分析等方面的发展。该文以合成生物数据库和数据智能分析为核心内容，描述了合成生物数据库建设的现状，讨论了合成生物数据质控和标准、存储和共享等方面的瓶颈问题和未来发展；另一方面，概述了人工智能技术在合成生物大数据智能分析方面的关键进展，讨论了系统建模、异构数据集成、智能设计与功能预测等方面的挑战与发展趋势。

关键词：合成生物学；数据库；大数据；机器学习；生物信息学
中图分类号：Q81；TP18；TP392　　**文献标志码**：A

收稿日期：2021-10-10
基金项目：国家自然科学基金面上项目(32071428)；中国科学院战略性先导B培育项目(XDPB18)
#共同第一作者
*通信作者：E-mail: yu.qiao@siat.ac.cn (乔宇)；tong.si@siat.ac.cn (司同)

Prospective of synthetic biology database and intelligent data analysis

HU Ru-Yun#, CHEN Yong-Can#, ZHANG Jian-Zhi, GUO Er-Peng, FU Li-Hao, QIAO Yu*, SI Tong*
(Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China)

Abstract: Synthetic biology accumulates massive data through the "design-build-test-learn" cycle, which has promoted the development of big data storage, sharing, and analysis. This prospective focuses on synthetic biology database and intelligent data analysis. For the former, we discuss the current status, bottleneck, and future development of data quality control and standards, storage and sharing in biology. For the latter, we highlight the use of artificial intelligence, particularly machine learning, in big data analysis in synthetic biology. We summarized the challenges and potentials of system modeling, heterogeneous data integration, intelligent design and function prediction.

Key words: synthetic biology; database; big data; machine learning; bioinformatics

人类基因组计划启动以来，以新一代测序技术和质谱技术为代表的各类组学技术飞速发展，推动基因组、转录组、表观遗传组、蛋白质组、代谢组等海量生命科学组学数据呈现指数级增长。与此同时，合成生物学基于工程学思想策略，在理性设计原则指导下，改造乃至从头合成具有特定功能的人造生命。合成生物研究提供了“建物致知”的崭新研究思想，开启了可定量、可计算、可预测及工程化的“会聚”研究新时代[1]。人们以标准化生物元器件和系统为对象，通过与高通量实验技术紧密结合，产生合成生物大数据。以细胞工厂为例，包括基于图论的分子结构转化特征数据、生物合成反应组合信息、催化元件性能数据、化学键能变化数据、基因线路调控数据、生物合成途径转化率数据、代谢与调控网络动态数据以及多组学表征数据等。

生物信息学是信息与系统科学和生命科学高度交叉的前沿学科，涉及多个学科领域，信息、控制与系统的理论、方法和技术在其中发挥着重要作用。在生物信息学诞生的阶段，其基础研究问题就是如何进行基因组数据的解读。在这之后，迎来了生物信息学发展的一个转型期 —— 后基因组时代。随着对生命系统的不断深入探究和各种其他高通量组学技术的产生和发展，生物信息学的研究范畴不断扩大，扩展到对各种组学数据 (转录组、蛋白质组、非编码 RNA 组、表观遗传组、代谢组、宏基因组等) 以及生物系统层面的解读，生命科学从定性科学转向定量科学[2]。

在当今大数据时代，生命科学领域的数据产出能力在各学科中处于领先位置，以基因组学和蛋白质组学数据为核心的组学大数据增长速度远超很多其他领域。一方面，各种新的组学技术不断发展，产生新的数据类型。每一种新的生物实验技术的产生，都需要设计新的生物信息分析方法，开发面向特定实验技术 / 高通量组学技术以及解读特定数据类型的计算机软件或者处理流程。另一方面，生物数据具有小样本高维度的特点。对某一个样本可进行各种高通量的组学测序，获得各个组学层面的特征表示，借助计算机科学和系统科学的方法手段，进行多组学的数据整合分析[3]。在新阶段的生物信息学研究中，提供数据的整合建模是一个很重要的能力：从数据中获得新的知识，发现新的生物学结论[4]。作为生物信息学发展的重要趋势，数据量迅速增大，数据类型不断增加，对生物信息学方法提出了大量新挑战；组学技术使越来越多层面的生物机理被揭示出来，系统生物学研究越来越走向对生物调控机理的定量认识和建模；同时，对生物系统认识的深化和合成生物学、基因编辑技术的不断突破，使得合成基因线路与系统的理论和技术有很大发展。

1 合成生物数据库和数据智能分析现有状况和水平

1.1 合成生物数据库现状

发达国家政府较早重视生命与健康大数据的收集、分析和应用。美国国家生物技术信息中心 (NCBI) 经过 30 多年对全世界生物技术数据的收集，积累了全世界最大的生命与健康数据库 (如 GenBank、PubMed、SRA、dbGap 等) 和软件资源 (如 BLAST、e-Utilities 等)[5]。目前，GenBank 数据库中存储的数据达 25 亿条，总数据量已达 15 TB 以上；PubMed 数据库包含超过 3 300 万生物医学文献。欧洲生物信息学研究所 (EBI) 目前已建成世界上最全面的分

子生物学数据库集合，2019 年原始数据已超过 300 PB。日本 DNA 数据库 (DDBJ) 目前自有数据量约为 15 PB。NCBI、EBI 和 DDBJ 共同成立了国际核酸序列数据库联盟 (INSDC)，是国际公共领域数据共享方面最著名的组织之一 [6]。此外，瑞士生物信息学研究所 (SIB) 数据库涵盖了生命科学的不同领域，包括基因组、蛋白质组、医药健康、进化、结构生物学和系统生物学等。

目前，我国各种类型的生命大数据中心也相继建成。具有代表性的包括：(1) 深圳国家基因库生命大数据平台 (China National GeneBank DataBase, CNGBdb)，整合了来源于国家基因库、NCBI、EBI、DDBJ 等平台的数据，元信息达 10 TB 以上；(2) 上海生物医学大数据中心，以中国科学院上海生命科学研究院自产数据为主；(3) 北京基因组研究所 (国家生物信息中心)，有近 25 PB 的存储资源，2018 年被期刊 *Nucleic Acids Research* 列为与美国 NCBI、欧洲 EBI 齐名的全球核心数据中心 [7-8]。

目前生物数据共享存在不同模式。NCBI 和 EBI 等机构通过数据递交服务汇聚了大量的数据资源，并通过网络提供数据共享；英国国家队列 UK Biobank (UKB) 等依托大型科研项目产出的数据，提供分级共享，满足不同类型的科研需求；介于这两者之间，中小型研究团队利用自身的数据采集能力和整合能力，建立了大量的种类繁多、规模悬殊、质量参差不齐的数据库和知识库，提供数据查询、浏览、下载服务，部分数据库还提供在线分析服务。同时，还有按照数据类型 (如基因组、转录组、蛋白质组等)、底盘 (如细菌、真菌、植物)、研究目的 (如基因线路、细胞工厂) 等方式建设的数据库 (表 1)，在推进数据共享方面发挥了巨大的作用 [9]。

1.2　合成生物数据智能分析发展概述

人工智能方法正在成为合成生物研究的强大工具。2006 年，Hinton 和 Salakhutdinov[38] 在 *Science* 发文提出深度学习概念。得益于多层神经网络从大规模训练数据中学习的能力，深度学习在计算机视觉 [39]、自然语言处理 [40]、游戏 [41] 等领域都取得了巨大成功，在一些特定任务上的表现甚至超过了人类。人工智能方法能从生物实验产生的海量数据中挖掘人脑不易发现的重要特征，正成为生命科学研究的强大工具。在基因组学领域，DeepBind 模型利用卷积神经网络来预测 DNA 序列上特定蛋白质的结合位点 [42]。DeepSEA 模型使用深度神经网络来解释非编码区变异的调控与功能 [43]。DeepTACT 模型借鉴集成学习思想预测启动子与其他基因调控元件的相互作用 [44]。在蛋白质折叠领域，DeepMind 公司利用深度学习方法开发的蛋白质结构预测软件 AlphaFold2 将蛋白质骨架的预测精度提高到了冷冻电镜的精度水平 [45]。在酶功能分析领域，卷积神经网络、递归神经网络等深度学习模型表现出极大的潜力 [46-47]。在免疫学领域，深度学习方法在 MHC 蛋白 - 抗原多肽相互作用预测、多抗原表位预测等方面取得重要进展 [48-49]。

深度学习在合成生物研究中已有概念性验证报道。2019 年，哈佛大学 Church 团队利用无监督深度学习方法开发氨基酸序列统一表示方法 UniRep，提高了蛋白质功能预测的准确度，并可用于蛋白质设计 [50]。利用深度生成式模型，研究者初步实现了多肽序列 [51]、酶序列 [52] 和 TCR[53] 序列的生成。清华大学汪小我团队利用深度对抗网络成功设计了全新的高表达大肠杆菌启动子元件，为生物调控元件的设计和优化提供了新的手段 [54]。斯坦福大学 Smolke 团队运用卷积神经网络，成功设计了具有高活性、高多样性的酿酒酵母启动子 [55]。相比于经验式设计方法，深度学习在合成元件设计的多样性、成功率方面已经显现出独特优势。

合成生物数据智能分析 [50, 56-58] 与生物设计 [59-62] 发展历程总结见图 1。

2　现有应用的瓶颈问题及未来应用中的关键问题

现有合成生物数据库分散、内容完整度差、缺乏统一标准，如何构建标准化合成生物数据库 [32-33]，进而构建全面、准确的合成生物知识库 [63-64]，是亟需解决的关键技术问题 (图 2)。例如，建立标准化命名访问服务 [65-66]，面临着数据来源分散、语法表达不一致等数据质量问题；多源异构合成生物数据库整合过程 [32]，面临着数据抽取、冗余清除、非完整或非一致合成生物数据清洗与转换等问题所带来的挑战；实体链接与知识补全 [64]，面临着合成生物领域缺乏足够的链接预测标签数据等问题，需要利用少量的标记数据有效地训练高质量链接预测模型。针对文献报道、公共数据库、实验结果等多源异构数据的整合和集成，需要解决人工标注训练数据稀少、自动标注易产生噪声关系标签、流水线式训练易导致错误传播等问题所带来的挑战。

当前，以深度学习为代表的人工智能方法，其分布外泛化 (推广) 能力差，从而要求数据具有独

表1 部分代表性合成生物学数据库

类别	数据库名称(网站链接)	简介
数据类型		
基因组	NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/)	综合性核苷酸序列数据库，由美国国家生物技术信息中心 (NCBI)维护，与日本DNA数据库(DDBJ)以及欧洲生物信息研究所(EBI)共同构成了国际核酸序列数据库，3个数据库之间以日为单位进行数据交换[10]
	NCBI Genome (https://www.ncbi.nlm.nih.gov/genome/)	综合性基因组数据库，包括序列、图谱、染色体、组装和注释等信息[10]
	FungiDB (http://fungidb.org/fungidb/)	真核病原体基因组学数据库(EuPathDB)的一部分，汇集了>100种真菌/卵菌的基因组、转录组、蛋白质组和表型等相关数据[11]
转录组	GEO (https://www.ncbi.nlm.nih.gov/geo/)	由NCBI维护的基因表达数据库，收录了全球研究机构提交的高通量基因表达数据[12]
	ArrayExpress (https://www.ebi.ac.uk/arrayexpress/)	收录了芯片和高通量测序的相关数据[13]
	M3D (http://m3d.mssm.edu/)	微生物基因表达数据库[14]
蛋白质组	Uniprot (https://www.uniprot.org/)	信息最丰富、资源最广的蛋白质数据库。由Swiss-Prot、TrEMBL和PIR-PSD三大数据库的数据组合而成[15]
	PRIDE (https://www.ebi.ac.uk/pride/)	欧洲生物信息研究所建立的主要基于质谱鉴定数据的蛋白质组学数据库[16]
	PROSITE (http://www.expasy.ch/prosite/)	集合了生物学具有显著意义的蛋白位点和序列模式，并可根据这些位点和模式快速分析未知蛋白的家族归属[17]
	MitoMiner (http://mitominer.mrc-mbu.cam.ac.uk)	集合了哺乳动物、斑马鱼及酵母线粒体蛋白质组数据[18]
代谢组	METLIN (http://metlin.scripps.edu)	在三种不同碰撞能量下(10、20和40 V)系统地汇集了超过13 000种化学标准品的高分辨串联质谱信息[19]
	MassBank (http://massbank.jp/)	基于代谢物化学标准品得到的质谱图，包含质谱仪设置、采集情况等[20]
	HMDB (http://www.hmdb.ca/)	包含人体代谢产物的详细信息[21]
底盘		
细菌	EcoCyc (https://www.ecocyc.org/)	大肠杆菌K12菌株基因组数据库，包括基因、蛋白质、基因间蛋白质组信息[22]
	CyanoBase (http://genome.microbedb.jp/cyanobase)	集胞蓝细菌的基因组数据库[23]
	SubtiList (http://genolist.pasteur.fr/SubtiList/)	枯草芽孢杆菌基因组数据库[24]
真菌	FGSC (http://fgsc.net/)	真菌遗传学信息中心[25]
	SGD (https://www.yeastgenome.org/)	酿酒酵母基因组数据库，汇集基因功能注释、突变体表型及关联文章等[26]
植物	MOsDB (http://mips.gsf.de/proj/rice)	水稻基因组数据库[27]
	SoyBase (https://www.soybase.org/)	大豆基因组学和分子生物学数据库[28]
	TAIR (https://www.arabidopsis.org/)	拟南芥基因组和注释数据库[29]
研究目的		
基因线路	FBDB (https://biosensordb.ucsd.edu/)	生物传感器数据库[30]
	The iGEM Parts Registry (http://parts.igem.org)	标准化的生物元件库[31]
	BioMaster (http://www.biomaster-uestc.cn)	基于BioBrick元件库整合包括UniProt、KEGG、BioGRID、BRENDA等10个数据库，注释元件功能、互作、文献等[32]
	SynBioHub (https://synbiohub.org/)	基于Web的合成生物学存储库，兼容SBOL标准，用户能够浏览、上传和共享合成生物学设计[33]
细胞工厂	KEGG (https://www.kegg.jp/)	整合基因组、化学和系统功能信息的数据库，可把完整测序基因组中得到的基因目录与更高级别的细胞、物种和生态系统水平的系统功能关联起来[34]
	BRENDA (https://brenda-enzymes.org/)	生化反应数据库[35]
	Metacyc (https://metacyc.org/)	非冗余、经实验验证的代谢途径和酶元件数据库[36]

表1 部分代表性合成生物学数据库(续表)

类别	数据库名称(网站链接)	简介
	Laser (https://bitbucket.org/jdwinkler/laser_release/)	整合了菌株代谢工程信息，包括培养条件、基因型及其他相关信息[37]

图1 合成生物数据智能分析[50, 56-58]与生物设计[59-62]发展历程

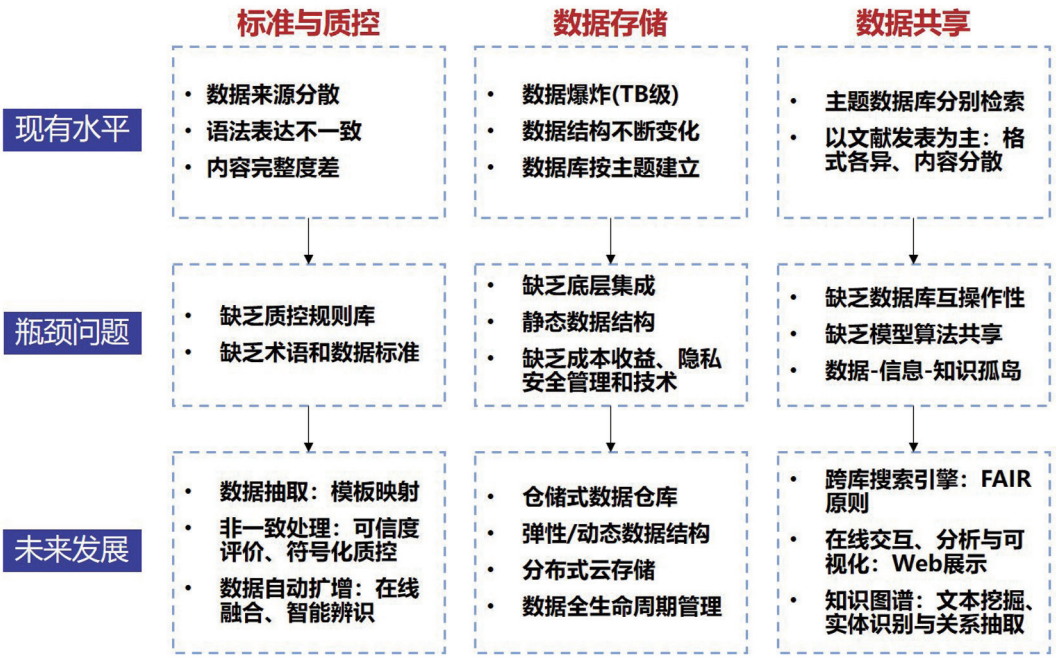


图2 合成生物数据库现有水平、瓶颈问题及未来发展方向

立同分布的性质且能较好地覆盖问题相关的全空间。这与合成生物问题的巨大空间和实验测试能力不足形成主要矛盾，成为当前人工智能方法在合成生物领域应用的瓶颈问题 (图 3)。因此，如何提高深度学习模型的泛化能力，是亟需解决的关键技术问题。例如，自监督学习可以充分利用无标签数据学习生物对象的有效表示，弥补实验测试能力不足 [50]；主动学习通过多轮次主动采样与实验形成闭环迭代，可以提高学习的样本效率，降低对实验测试能力的要求 [60]；强化学习通过与环境 (实验或者适应度地形模型) 交互，可以实现生物对象的设计优化或实验条件的优化 (实验设计)，以提高合成生物设计能力或实验测试能力 [67]；知识驱动的学习，通过利用问题相关知识，设计更适合问题特点的数学表示，补充模型的输入特征，搭建更加合理和可解释的模型架构，构造具有生物学意义的损失函数 [68]。

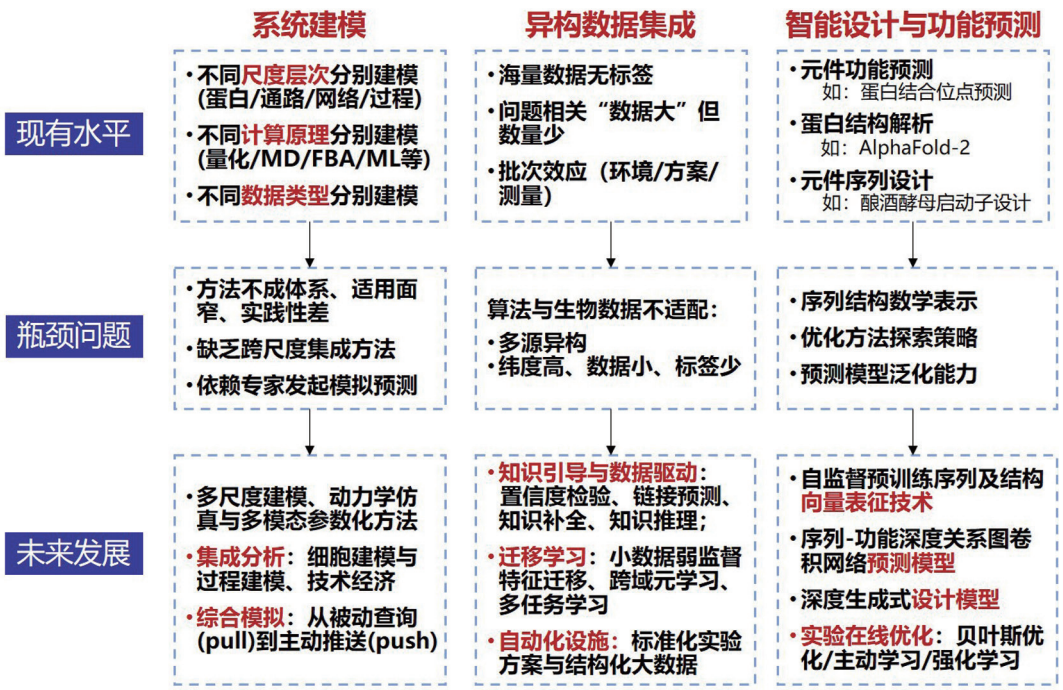


图3 合成生物数据智能分析现有水平、瓶颈问题及未来发展方向

3 未来5~15年需要优先发展的方向及领域

3.1 合成生物数据库和知识图谱

合成生物数据库和知识图谱方面，需要建立适应大数据时代的新的技术和资源体系，建设面向合成生物研究的数据仓库、数据库和知识图谱等，用于合成生物大数据的标准化存储、共享和挖掘分析等 (图 2)。未来重点的研究方向和需要突破的关键技术如下。

3.1.1 数据标准、质量控制

合成生物学的标准主要集中在术语集和数据标准，不同的标准之间相对独立，对数据产出过程、分析过程的规范性表述较少。在建设合成生物大数据平台时，应制定定量描述合成生物元件及其互作网络的标准，建立数据质控规则库和标准化命名访问体系，解决数据来源分散、语法表达不一致等引起的数据质量问题。针对现有合成生物数据库内容完整度差、存在噪声等问题，开发基于模板映射的数据准确抽取方法和基于可信度的非一致数据检测与处理方法，实现标准化命名服务。以此为基础，开发合成生物结构化数据库建库技术，实现对已有多源异构数据库的整合。

针对高通量实验技术，建立公开和在线实验数据动态更新的合成生物数据库，以及数据库内容快速访问。面向公开数据库和合成生物平台产生的测序、质谱等高通量实验数据，开发在线融合、智能辨识等方法，能够针对特定合成生物目标进行动态搜索和数据集构建，实现数据的自动扩增。针对合成生物元件来源分散、语法表达不一致带来的数据访问可控性差、响应不及时等问题，构造符号化的数据质控规则库，建立合成生物元件及属性的标准化命名访问服务体系，构建基于 Spark 的分布式云存储定量合成生物信息数据库。与此同时，系统性探索过程参数 (实验条件和测试手段等) 如何引入实验数据噪声进而影响数据处理和挖掘，制定标准操作流程、报告规范和质控方法，建立不同实验产生、共享、集成合成生物大数据的方法。

3.1.2 整合式数据仓库

传统的数据模型和数据组织方式无法满足合成生物海量数据的结构、数量快速增长以及数据结构不断变化的管理需求；有必要突破传统的严格按照一类数据建设一个数据库的模式，采用新的仓储式的数据仓库模式 [9]。在底层数据结构上以整合为导向，按照合成生物设计、研究对象、底盘细胞类型、环境等信息，以及时间、空间信息，预留不同类型的数据之间的联系，形成弹性的数据结构，支持数据结构动态调整，为数据库集成、整合和挖掘奠定基础。同时，在建设合成生物大数据平台时，TB

量级的数据下载需求对数据下载、单库检索等数据共享手段提出了严峻的挑战。因此，在延续按照主题 (数据类型、物种、研究领域) 组织数据的基础上，引入跨库搜索引擎、可视化、在线分析等在线交互技术，通过更加准确地返回用户数据访问结果的方式，提高数据共享效率。

3.1.3 数据共享

随着数据类型和规模的日益扩大，如何存储、组织、访问存放在不同平台上的不同类型的生物医学数据成为新的挑战。为此，研究者提出 FAIR 原则，即可发现 (findable)、可访问 (accessible)、互操作 (interoperable) 和重用 (reusable)[69]。基于 FAIR 原则，采用搜索引擎等技术，可以突破传统的以主题为基础建设的数据库的局限性，对不同数据中心的数据资源提供统一检索服务，实现以搜索引擎为核心的数据跨库整合，更好地满足用户一站式的数据共享需求。

除了搜索技术外，数据可视化、在线分析也是用户利用数据的重要手段。新的可视化技术，包括 HTML5、JavaScript 等 Web 展示技术在数据平台中的应用越来越广泛，可用于大分子展示、基因组浏览器、代谢网络解析等 [70]。此外，依托数据库的分子序列、分子结构、调控及相互作用网络等数据，数据库根据自身特点，集成了序列比对、多序列比对、结构相似性比较、网络结构分析等在线分析的工具，也极大地加强了数据的可交互性。

3.1.4 知识图谱

构建合成生物知识库与知识图谱，为知识引导和数据驱动紧密结合的合成生物研究奠定基础 [71]。利用来自文献报道、公共数据库、实验结果等的多源异构数据，研究基于深度学习的语义关系匹配、置信度检验、实体链接等方法，开发合成生物知识图谱，构建更全面、准确的合成生物知识库。在合成生物元件命名、实体识别和关系抽取方面，利用远程监督学习来缓解人工标注训练数据稀少问题；开发多任务学习框架，联合训练实体识别和关系抽取模型，来减轻流水线式训练导致的错误传播。在实体链接与知识补全方面，采用基于分布式表示空间的实体链接方案，运用联合学习、正则化技术和矩阵协同分解等技术。

3.2 数据智能分析

数据智能分析方面，为了提升理性设计人工生命的能力，需要深度集成传统生物信息技术与新型人工智能方法，实现数据驱动的“设计 - 构建 - 测试 - 学习”智能闭环，在系统建模、异构数据集成、智能设计与功能预测等方面进行关键技术突破 (图 3)。

3.2.1 系统建模

现有合成生物研究主要针对单一组分、单一基因线路以及单一细胞器等开展计算模拟。随着组学大规模数据的积累、信息理论的应用，以及化学和工程科学等多学科的交叉和融合，系统、整合、跨尺度研究细胞内不同组分和结构的功能与互作机制成为可能。细胞功能的系统整合研究是在对细胞内所有组分进行鉴定和认识的基础上，描绘出细胞的系统结构，包括生物大分子相互作用网络和细胞内亚结构间的互作系统，构造出初步的细胞系统模型，通过不断地设定和实施新干预实验，对模型进行修订和精练，最终获得一个理想的模型，使其理论预测能够反映出细胞的系统功能和真实性。

例如，针对具有目标代谢功能的细胞工厂，需要开发一系列生信与智能设计方法。元件层面，需要确定一组可以将容易获得的分子转化为高价值产品的酶蛋白，找出与底盘细胞高度适配的酶蛋白组合，结合调控元件优化其化学计量学；途径层面，基因组规模的代谢模型通过重建生物体的完全代谢反应网络，将基因型与表型联系起来，用于定义理论生产限度以及设计和测试计算机中的新微生物菌株，预测细胞生长、通量分布、产物合成，并指导宿主设计。细胞功能实现的系统整合研究可以推动对生命基本单元 —— 细胞的功能机制的深入认识，对于未来的人造细胞、合成生命以及新型生物产业发展如细胞工厂、细胞治疗等均具有重要的意义。

3.2.2 异构数据集成

合成生物数据具有小样本、高维度、多源异构等特点，需要针对性开发新的生物信息和机器学习理论和方法 [4]。例如，不同层次合成生物对象数据的集成需要推断多个实体之间的关系，这类关系可能存在于相同实体 (如蛋白质 - 蛋白质互作、基因 - 基因共表达等) 或不同实体 (如序列 - 结构 - 功能之间的关系、元件 - 模块 - 底盘适配等) 之间。实体关系推断可以利用多矩阵下的协同分解、二部图的随机游走等机器学习算法进行关联性预测。面对数据源分布不一致、直接相关数据缺乏的问题，一方面可以借鉴共训练、多层面等学习模式，消除不同实验技术、实验批次之间的异质性；另一方面可以基于多任务学习、迁移学习等模型，集成利用不同物种、不同细胞系或者不同实验条件下同类生物学问题的相关数据。

3.2.3 智能设计与功能预测

现有方法多面向基因、蛋白质等研究中的某个特定任务或环节，不成体系且适用面窄，与合成生物实践适配性差。针对这些困难，需要研发与合成生物研究高度适配的人工智能方法体系。具体如下：

(1) 开发基因型到表现型的关系预测模型。针对合成生物对象序列维度高、样本数据量小、合成任务多样等挑战，以合成生物大数据库和知识库为基础，研究小数据和弱监督下的特征迁移、多任务学习和跨域元学习等算法，开发基于自监督预训练的基因 / 蛋白质序列及结构的向量表征方法，开发基因型 - 表现型间的深度关系图卷积网络模型，实现特定生物功能的精准预测。

(2) 建立合成生物元件系统的生成式设计算法。针对合成生物设计空间巨大、海量实验试错成本高、效率低等挑战，采用数据驱动和知识引导相结合的方式，开发特定合成生物对象的生成式设计和优化模型，如变分自编码器、生成对抗网络等深度模型，生成在高维空间中具有相似特征的人工序列，并利用功能预测算法对其进行筛选。

(3) 开发实验方案在线智能优化方法。针对合成生物实验中辅助元件和底盘细胞选择等难题，利用贝叶斯优化等方法，实现实验方案的动态优化配置，采用主动学习和强化学习思想，结合自动化合成生物研究平台的在线数据生成能力，实现人工智能模型与实验的迭代优化，达到目标功能所需试错规模比传统方法大幅减少。

[参 考 文 献]

[1] 赵国屏. 合成生物学: 开启生命科学“会聚”研究新时代. 中国科学院院刊, 2018, 33: 1135-49

[2] Gauthier J, Vincent AT, Charette SJ, et al. A brief history of bioinformatics. Brief Bioinform, 2019, 20: 1981-96

[3] Li Y, Chen L. Big biological data: challenges and opportunities. Genomics Proteomics Bioinformatics, 2014, 12: 187-9

[4] 刘琦. 生物信息学研究的思考. 中国计算机学会通讯, 2016, 12: 49-53

[5] Sayers EW, Beck J, Brister JR, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 2020, 48: D9-16

[6] 鲍一明, 薛勇彪. 生命与健康大数据现状和展望. 中国科学院院刊, 2018, 33: 861-5

[7] BIG Data Center Members. Database resources of the BIG Data Center in 2018. Nucleic Acids Res, 2018, 46: D14-20

[8] Rigden DJ, Fernandez XM. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Res, 2018, 46: D1-7

[9] 张国庆, 李亦学, 王泽峰, 等. 生物医学大数据发展的新挑战与趋势. 中国科学院院刊, 2018, 33: 853-60

[10] Sayers EW, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Res, 2021, 49: D92-6

[11] Basenko EY, Pulman JA, Shanmugasundram A, et al. FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J Fungi, 2018, 4: 39

[12] Clough E, Barrett T. The Gene Expression Omnibus database. Methods Mol Biol, 2016, 1418: 93-110

[13] Sarkans U, Fullgrabe A, Ali A, et al. From ArrayExpress to BioStudies. Nucleic Acids Res, 2021, 49: D1502-6

[14] Faith JJ, Driscoll ME, Fusaro VA, et al. Many Microbe Microarrays Database: uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic Acids Res, 2008, 36: D866-70

[15] Bateman A, Martin MJ, Orchard S, et al. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res, 2019, 47: D506-15

[16] Perez-Riverol Y, Csordas A, Bai JW, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res, 2019, 47: D442-50

[17] Hulo N, Bairoch A, Bulliard V, et al. The PROSITE database. Nucleic Acids Res, 2006, 34: D227-30

[18] Smith AC, Robinson AJ. MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. Nucleic Acids Res, 2019, 47: D1225-8

[19] Xue JC, Guijas C, Benton HP, et al. METLIN MS2 molecular standards database: a broad chemical and biological resource. Nat Methods, 2020, 17: 953-4

[20] Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom, 2010, 45: 703-14

[21] Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res, 2018, 46 : D608-17

[22] Keseler IM, Collado-Vides J, Santos-Zavaleta A, et al. EcoCyc: a comprehensive database of *Escherichia coli* biology. Nucleic Acids Res, 2011, 39: D583-90

[23] Nakamura Y, Kaneko T, Tabata S. CyanoBase, the genome database for *Synechocystis* sp. strain PCC6803: status for the year 2000. Nucleic Acids Res, 2000, 28: 72

[24] Moszer I, Jones LM, Moreira S, et al. SubtiList: the reference database for the *Bacillus subtilis* genome. Nucleic Acids Res, 2002, 30: 62-5

[25] McCluskey K, Wiest A, Plamann M. The Fungal Genetics Stock Center: a repository for 50 years of fungal genetics research. J Biosci, 2010, 35: 119-26

[26] Cherry JM, Hong EL, Amundsen C, et al. *Saccharomyces* Genome Database: the genomics resource of budding yeast. Nucleic Acids Res, 2012, 40: D700-5

[27] Karlowski WM, Schoof H, Janakiraman V, et al. MOsDB: an integrated information resource for rice genomics. Nucleic Acids Res, 2003, 31: 190-2

[28] Grant D, Nelson RT. SoyBase: a comprehensive database

for soybean genetic and genomic data[M]//Nguyen H, Bhattacharyya M. The soybean genome. Compendium of plant genomes. Cham: Springer, 2017: 193-211
[29] Rhee SY, Beavis W, Berardini TZ, et al. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Res, 2003, 31: 224-8
[30] Greenwald EC, Mehta S, Zhang J. Genetically encoded fluorescent biosensors illuminate the spatiotemporal regulation of signaling networks. Chem Rev, 2018, 118: 11707-94
[31] Galdzicki M, Rodriguez C, Chandran D, et al. Standard biological parts knowledgebase. PLoS One, 2011, 6: e17005
[32] Wang BB, Yang HY, Sun JN, et al. BioMaster: an integrated database and analytic platform to provide comprehensive information about BioBrick parts. Front Microbiol, 2021, 12: 593979
[33] McLaughlin JA, Myers CJ, Zundel Z, et al. SynBioHub: a standards-enabled design repository for synthetic biology. ACS Synth Biol, 2018, 7: 682-8
[34] Kanehisa M. The KEGG database. Novartis Found Symp, 2002, 247: 91-101
[35] Schomburg I, Chang AJ, Hofmann O, et al. BRENDA: a resource for enzyme data and metabolic information. Trends Biochem Sci, 2002, 27: 54-6
[36] Caspi R, Foerster H, Fulcher CA, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res, 2006, 34: D511-6
[37] Winkler JD, Halweg-Edwards AL, Gill RT. Quantifying complexity in metabolic engineering using the LASER database. Metab Eng Commun, 2016, 3: 227-33
[38] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504-7
[39] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM, 2017, 60: 84-90
[40] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv, 2020: 2005.14165
[41] Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484-9
[42] Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 2015, 33: 831-8
[43] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods, 2015, 12: 931-4
[44] Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res, 2019, 47: e60
[45] Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. Nature, 2021, 596: 590-6
[46] Kim HK, Min S, Song M, et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nat Biotechnol, 2018, 36: 239-41
[47] Sequeira AM, Rocha M. Recurrent deep neural networks for enzyme functional annotation[M]//Rocha M, Fdez-Riverola F, Mohamad MS, et al. Practical applications of computational biology & bioinformatics, 15th International Conference (PACBB 2021). Cham: Springer, 2022, 325: 62-73
[48] Chen B, Khodadoust MS, Olsson N, et al. Predicting HLA class II antigen presentation through integrated deep learning. Nat Biotechnol, 2019, 37: 1332-43
[49] Yang Z, Bogdan P, Nazarian S. An *in silico* deep learning approach to multi-epitope vaccine design: a SARS-CoV-2 case study. Sci Rep, 2021, 11: 3238
[50] Alley EC, Khimulya G, Biswas S, et al. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods, 2019, 16: 1315-22
[51] Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions. Nat Mach Intell, 2019, 1: 105-11
[52] Repecka D, Jauniskis V, Karpus L, et al. Expanding functional protein sequence spaces using generative adversarial networks. Nat Mach Intell, 2021, 3: 324-33
[53] Davidsen K, Olson BJ, DeWitt WS, et al. Deep generative models for T cell receptor protein sequences. Elife, 2019, 8: e46935
[54] Wang Y, Wang H, Wei L, et al. Synthetic promoter design in *Escherichia coli* based on a deep generative network. Nucleic Acids Res, 2020, 48: 6403-12
[55] Kotopka BJ, Smolke CD. Model-driven generation of artificial yeast promoters. Nat Commun, 2020, 11: 2113
[56] Hopf TA, Ingraham JB, Poelwijk FJ, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol, 2017, 35: 128-35
[57] Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods, 2018, 15: 816-22
[58] Hsu C, Nisonoff H, Fannjiang C, et al. Combining evolutionary and assay-labelled data for protein fitness prediction. bioRxiv, 2021, doi: 10.1101/2021.03.28.437402
[59] Romero PA, Krause A, Arnold FH. Navigating the protein fitness landscape with Gaussian processes. Proc Natl Acad Sci U S A, 2013, 110: E193-201
[60] HamediRad M, Chao R, Weisberg S, et al. Towards a fully automated algorithm driven platform for biosystems design. Nat Commun, 2019, 10: 5150
[61] Wittmann BJ, Yue Y, Arnold FH. Informed training set design enables efficient machine learning-assisted directed protein evolution. Cell Syst, 2021, 12: 1026-45.e7
[62] Biswas S, Khimulya G, Alley EC, et al. Low-N protein engineering with data-efficient deep learning. Nat Methods, 2021, 18: 389-96
[63] Callahan TJ, Tripodi IJ, Pielke-Lombardo H, et al. Knowledge-based biomedical data science. Annu Rev Biomed Data Sci, 2020, 3: 23-41
[64] Mante J, Hao YK, Jett J, et al. Synthetic biology knowledge system. ACS Synth Biol, 2021, 10: 2276-85
[65] Galdzicki M, Clancy KP, Oberortner E, et al. The Synthetic Biology Open Language (SBOL) provides a

community standard for communicating designs in synthetic biology. Nat Biotechnol, 2014, 32: 545-50
[66] McLaughlin JA, Beal J, Grunberg R, et al. The Synthetic Biology Open Language (SBOL) version 3: simplified data exchange for bioengineering. Front Bioeng Biotechnol, 2020, 8: 1009
[67] Angermueller C, Dohan D, Belanger D, et al. Model-based reinforcement learning for biological sequence design. ICLR, 2020, Section 3: 1-23
[68] Karniadakis GE, Kevrekidis IG, Lu L, et al. Physics-informed machine learning. Nat Rev Physics, 2021, 3: 422-40
[69] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 2016, 3: 160018
[70] Yuan S, Chan HCS, Hu Z. Implementing WebGL and HTML5 in macromolecular visualization and modern computer-aided drug design. Trends Biotechnol, 2017, 35: 559-71
[71] Weis JW, Jacobson JM. Learning on knowledge graph dynamics provides an early warning of impactful research. Nat Biotechnol, 2021, 39: 1300-7