



人工智能生物学——生物学3.0

周祯^{1,2}, 闫超^{1,2}, 张辰宇^{1,2*}

1. 南京大学生命科学学院, 医药生物技术国家重点实验室, 南京 210023;

2. 南京大学人工智能生物医药技术研究院, 南京 210031

* 联系人, E-mail: cyzhang@nju.edu.cn

收稿日期: 2021-08-10; 接受日期: 2022-01-27; 网络版发表日期: 2022-02-23

摘要 现代生物学一方面取得了重大进展, 另一方面也面临着巨大的发展瓶颈, 而伴随着人工智能(artificial intelligence, AI)理论和技术的快速发展, AI与生物学深度融合的生物学V3.0——人工智能生物学(artificial intelligence biology, AIBIO)已经呼之欲出。人们将人工智能生物学定义为利用人工智能的原理和手段来研究生命系统基本规律的科学。其研究特点是: 动态整合多层面与多因素, 从而真正理解生命现象中的分子间相互作用与相互调控的规律, 解决生命科学中的重大基本问题。作为一个全新的生命科学学科, 人工智能生物学将全面提升生物学研究的高度, 革新生物学研究的现有范式, 拓展生物学研究的范围, 实现生命科学和医学科学关键领域的实质性突破。及时掌控并引领相关领域的研究, 对推动生命科学领域的基础研究、技术发展, 甚至对整个社会的进步都至关重要。

关键词 人工智能, 人工智能生物学, 机器学习, 深度学习

回顾生命科学的发展历史, 大致可以分为两个大的阶段。第一阶段为18世纪初至20世纪中叶的基于博物学研究方法的早期生物学研究(生物学V1.0)。这一时期的博物学家通过观察不同生物的形态和行为模式, 发现了地球上不同生命存在的形式和遗传演化的一般规律, 如达尔文的进化论^[1]和孟德尔的经典遗传学^[2]。第二阶段是1953年DNA结构发现^[3]至今的现代生物学研究阶段(生物学V2.0)。现代生物学研究的特点主要是利用物理和化学手段, 在分子、细胞和整体层面上分析生命体的分子组成及其相互作用规律。其中, 对于生命体的分子组成研究进展较大, 对单个蛋白、核酸、糖等构成生命体的基本物质的化学结构和生物学功能的了解相对充分, 但对于分子间的相互作

用的研究却进展缓慢。究其原因, 可能由于生命体系的调节和控制往往是多因素协同作用的结果, 各因素间的相互作用存在极端的复杂性。另外, 在不同物理尺度上, 从细胞器、细胞, 到组织、器官、系统, 再到个体、种群和整个生态系统, 存在着跨尺度的信息交互网络, 在生物代际之间和生命体与环境之间还存在着更加复杂的层阶调控, 研究时必须考虑时空维度上的动态变化。这些客观存在的极端复杂性导致以生物化学、分子生物学等学科为代表的现代生物学发展几十年后, 遇到了重大的瓶颈。一方面, 现代生物学研究虽然大量揭示了某些生物分子在特定情景下的可能作用方式, 但这种研究模式的基本假设是生命体系内的相互作用模式是线性单一的, 这与生命系统天然存在

引用格式: 周祯, 闫超, 张辰宇. 人工智能生物学——生物学3.0. 中国科学: 生命科学, 2022, 52: 291–300

Zhou Z, Yan C, Zhang C Y. Artificial intelligence biology—biology V3.0 (in Chinese). Sci Sin Vitae, 2022, 52: 291–300, doi: 10.1360/SSV-2021-0298

的非线性属性和超级复杂性相去甚远。另一方面, 现代生物学基于化学分析的基本实验技术手段和以人脑逻辑推理为基础的数据分析方式, 在很多情况下无法满足对生命体系超级复杂性研究的需要。特别是由于研究方法和研究手段不完全统一, 使得不同研究手段得到的数据并不具有高可比性。虽然随着组学和生物信息学的发展, 实现了相对大规模数据的分析, 但仍只能集中在一些特定的层次和体系, 并不能跨越生命科学研究的整个维度, 难以阐述生命现象的本质。例如, 现代生物学仍然不能解释基因序列和基因功能之间的因果关系。近年来, 尽管多组学技术为定量刻画生物系统提供了海量的精确数据, 但如何从海量数据中提炼普适性的生物学规律, 并最终形成可定量的理论生物学体系, 仍是生物学研究领域正在探索的前沿方向之一。

反观其他自然科学, 如物理学和化学, 都具有相对完整的基础理论体系作为主线, 指导相关分支学科的探索与发现。例如, 门捷列夫在1869年发明的元素周期律^[4], 揭示了物质世界的底层规律, 把看似不相关的元素统一起来, 组成了一个完整的逻辑自洽的理论体系。门捷列夫将当时已知的63种元素根据相对原子质量大小, 通过表的形式排列, 把相似化学性质的元素放在同一列, 形成了元素周期表的雏形。利用周期表, 门捷列夫成功地预测了当时尚未发现的元素, 例如镓、铟、锗等。随着科学的发展, 元素周期表中未知元素留下的空位逐渐被填满^[5]。元素周期律这一理论规律的发现成为化学科学史上的一座里程碑, 它使得人们能够理解元素之间的内在联系, 并为新元素的发现和元素性质的研究提供了巨大的帮助。

和其他自然科学一样, 实现对生命系统的定量研究是生命科学向更高阶段发展的必由之路。生物学发展至今, 尚未形成一个完整的基础理论体系和与之配套的数学描述, 某些特定的生物学问题虽然已经有相应的数学模型存在, 但这些模型通常仅仅是唯象或者描述性的模型, 很难对实验结果进行预测或演绎, 其定性指导意义远远大于定量指导意义。本文认为, 随着人工智能(artificial intelligence, AI)理论和技术的快速发展, 人工智能与生物学的深度融合有可能解决生物学研究的定量问题, 催生“人工智能生物学”这一新的理论体系。

1 人工智能的特点能够帮助解决生物学研究的瓶颈问题

人工智能的本质是模拟人的思维方式来解决复杂问题(如无人驾驶系统就是一个典型的模拟人类思维过程的系统)。机器学习(machine learning, ML)是实现人工智能的核心方法, 是使用机器学习算法分析数据规律, 从而进行推断和预测。深度学习(deep learning, DL)是最近发展起来的一种机器学习算法, 深度学习源于对神经网络的研究, 目标是使计算机能模仿人的大脑网络进行工作^[6]。

最近两三年来, 人工智能领域飞速发展, 以谷歌的AlphaGo^[7]和AlphaGo Zero^[8]为代表的AI已经开始迅速颠覆从日常生活到科学研究等各个领域的传统认知。从目前人工智能在其他非生命科学领域的应用来看, 人工智能特别是深度学习所展现出来的几大特点对于解决目前生命科学研究中存在的根本性瓶颈问题有着极强的针对性。

(1) AI擅长解决“黑盒子”问题^[9], 而生物学研究的很多问题都处在黑盒子状态。以深度学习算法中的卷积神经网络(convolutional neural network, CNN)为例, 通过多个层级的堆叠, 卷积神经网络可以模拟黑盒子过程的多层级多因素调控, 只要输入和输出是已知的且可量化的, AI就可以通过已知数据的反复训练, 自发调整和优化卷积神经网络算法中各层级各变量之间的函数关系, 从而实现对黑盒子的过程模拟。训练集的数据量越大, 训练结果就越准确, 训练好的AI模型就可以根据新的输入来预测输出^[10]。当前的生物学研究在很多领域都还处于完全或部分黑盒子状态, 最典型的例子要属神经生物学和免疫学研究。从这个角度来说, 引入AI对于生物学的意义之一在于在没有完全搞清楚黑盒子内部运作机制的情况下, 也可以建立输入(基因和表达)和输出(生理和病理表型)的有效关联, 在医学实践上就可以实现对疾病的干预和治疗。

(2) AI能够以有限信息推测“图像”全貌。在图像识别领域, AI已经可以做到在学习海量照片资料后, 根据部分信息缺失的图片(信息丢失率可以高达90%以上)还原图片全貌^[11~14]。AI这种以有限信息推测全貌的能力在生物学研究的多个方面都可以起到重要作用。例如, 在结构生物学研究中, X射线晶体衍射照片的细节数据的缺失是非常常见的技术问题, 通过AI深度学习

就可以解决这个用传统方法无从下手的问题。再例如, 现有的二代测序技术本身会有大约1%~3%的错误率, 利用AI方法也可以对测序数据本身进行修正以达到去伪存真的目的。

(3) AI能够根据“低维”数据预测“高维”信息。从低维空间到高维空间的跨越是非线性的, 现有的生物学理论无法建立不同维度信息之间的直接因果关系。例如, 根据蛋白质的一维结构(氨基酸序列)很难推断其三维结构(空间构象); 同样的, 从DNA的一维结构(碱基序列)也无法预测其三维结构。解决这个问题经典的方法是从最底层的物理化学理论出发, 来计算原子间的可能结合方式, 然而计算生物大分子所需的算力远远超出了现有的机器能力; AI既可以模仿人的思维方式, 也可以超越人的思维方式, 采取完全不同的策略, 如模式识别(pattern recognition)来解决问题。通过海量数据的训练, AI甚至可以绕开人类总结出来的物理化学原理, 直接总结出对AI有意义的(但是人可能看不懂的)低维信息和高维信息之间的关联性和对应规律^[15,16], 从而实现从低维数据出发对高维信息的预测。最近横空出世的AlphaFold2实现了对人类基因组中98%蛋白的三维结构预测, 就是对AI这一能力的有力证明^[17]。

(4) AI擅长复杂系统的多因素分析, 从碎片化数据中找到规律。生命系统的复杂程度远高于一般的物理化学体系, 生命体所呈现出来的生理表型(正常状态)或病理表型(疾病状态)都是由内因(基因表达)和外因(环境因素)共同作用的结果, 而且表型的控制往往是多层次多因素的, 很少有单因素决定表型的现象存在。以疾病为例, 人类已经研究清楚的很多疾病是单因素导致的(因为其规律相对简单易于发现), 如血友病和色盲这种单基因遗传疾病; 但是人类的大多数疾病, 包括癌症、心血管疾病、神经退行性疾病这些真正有挑战性的疾病往往都是多因素导致的^[18-20]。当前生物学的研究范式还停留在试图寻找单个基因突变或表达差异对疾病的影响, 从而将其作为疾病检测的生物标志物或者药物干预的对象, 从源头上缺乏对于复杂系统多因素分析的有效手段^[21]。毋庸置疑, AI在复杂系统多因素分析方面的能力远远超过了人类。

随着人工智能理论和方法的日趋成熟, 后基因组学时代的生命科学和医学研究正站在历史的十字路口, 只有引入人工智能才能解决当前生物学研究中遇

到的重大瓶颈问题, 推动生物学研究进入一个全新的发展阶段。基于以上分析, 本文提出“人工智能生物学”这一全新的学科概念: 人工智能生物学是利用人工智能的原理和手段来研究生命系统基本规律的科学。其特点是: 动态整合多层面与多因素, 寻找生命现象中的基本规律, 并通过获得新的知识来解决生命科学的基本问题。人工智能生物学的优势在于其能够从海量的复杂数据中发现人类大脑无法分析、难以理解的数据结构, 捕捉和提取人类难以察觉的数据特征, 为解决生命科学的基础理论问题提供新的思路。本文认为, 人工智能对于生物学研究的影响将是颠覆性的, 将彻底改变现有的生物学研究思维和研究范式。

2 人工智能应用于生物学研究的发展历程

2.1 经典机器学习时期

人工智能与生物学研究早在生物信息学这门生命科学的分支学科诞生之初就紧密结合到一起了。自20世纪80年代起, 机器学习方法已经在生物信息学研究中得到广泛的应用。例如, 基于神经网络的蛋白质、RNA的二级结构预测^[22,23], 基于隐马尔可夫模型(Hidden Markov model, HMM)的基因识别、基因剪接位点、外显子及内含子预测^[24-26], 基于聚类算法的基因表达数据的挖掘^[27]等工作。

以HMM为例, 这是在生物信息学中应用最广泛的一种经典机器学习方法。1992年的Snowbird会议中, Haussler首次报道了使用HMM算法进行多序列比对的建模。之后Durbin等人^[28]利用HMM算法探索了许多生物信息学相关的问题, 并且为生物信息学建立了一套相对完整的概率和统计模型。

这一时期, 我国学者也作出了开拓性的贡献。例如, 中国科学院的陈润生教授早在1989年就尝试用人工神经网络预测蛋白质的二级结构, 这也是我国最早利用人工智能技术解决生物学问题的探索^[29]; 清华大学的孙之荣教授在1997年完善了用神经网络预测蛋白质二级结构的方法并在国际上产生了重要的影响^[30]; 孙之荣教授还对利用支持向量机(support vector machine, SVM)算法解决蛋白质二级结构预测和亚细胞定位问题作出了开创性贡献^[31,32]。

2003年, Gerstein团队^[33]在*Science*发表使用贝叶斯网络算法(Bayesian network)预测蛋白互作的论文之

后, 整个生物信息学领域逐步开始为经典机器学习算法建立统计学的基础, 从而增加模型的可解释性. 经典机器学习算法, 一般也可称为浅层学习(shallow learning)或者统计学习(statistical learning), 其中的各类算法, 包括SVM, HMM和Bayesian, 都可以看成只有单隐含层的神经网络. 它们虽然原理不同, 但都有非常好的统计学基础, 所以不能算是纯黑的黑盒子.

通过使用机器学习模型去拟合生物学数据, 在一定程度上帮助当时的研究者理解了一些生命活动中生物学规律. 但是, 由于生命模型的复杂性以及生物数据与传统数据的差异性, 传统的机器学习方法并不能在所有生物学问题上都有很好的应用, 因此, 2000年之后这方面的研究进展基本趋于平缓.

2.2 深度学习时期

近3年来, 随着深度学习在图像识别、自然语言处理等领域内的飞速发展^[34,35], 人工智能特别是深度学习算法重新在生物学研究领域内崭露头角^[36,37].

2015年, AI研究重镇之一的加拿大多伦多大学在*Nature Biotechnology*发表题为“Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning”的研究论文^[38], 这是最早利用深度学习研究方法研究生物学问题的报道. 得益于我国人工智能基础研究的实力和国际地位, 中国学者在这一方向上的探索和世界基本同步. 例如, 清华大学曾坚阳教授关于序列分析^[39,40]和北京大学来鲁华教授关于药物毒性预测的工作^[41]. 随着深度学习技术本身的发展, 无监督学习迅速成熟, 同济大学的刘琦教授使用无监督学习方法, 在CRISPR sgRNA的精准预测和单细胞测序数据分析等方面都做出了引领性的工作^[42,43].

2018年, 加州大学圣地亚哥分校的张康教授在*Cell*发表题为“Identifying medical diagnoses and treatable diseases by image-based deep learning”的研究论文^[44], 首开深度学习应用于生物医学图像处理之先河, 此后, 医学图像数据的处理、分析与整合开始广受关注并迅速商业化. 目前深度学习算法在图像分类与识别领域内的超强性能, 使得人工智能在生命科学和医学领域的应用几乎全部集中于代替病理医生阅读医学影像和代替内科医生进行疾病诊断(例如最近宣布失败的IBM公司的Watson医生项目)这两个方向上^[45,46]. 全世界仍然仅有少数实验室开始尝试利用深度学习的

方法进行生物学的基础研究. 人工智能如何与生命科学更加有机地结合依然需要进一步探究.

3 人工智能生物学的任务

综合上述人工智能的特点及其在生物学研究中的应用的发展历程, 本文认为, 未来AI和生物学基础研究的结合应该在以下三大方向有所突破.

3.1 AI与分子生物学

生命活动的本质是生物分子间的相互作用与相互调控. 希望通过对作为生命活动物质基础的蛋白质及核酸大分子的结构和功能的探究, 能更好地理解它们之间相互作用和相互调控的生物学规律. 结构生物学家已经在这方面做出了巨大的贡献, 例如, 1953年DNA双螺旋结构^[3]的发现直接导致了“DNA-RNA-蛋白质”中心法则的揭示^[47], 并直接催生了分子生物学这门学科. 目前结构生物学的主要研究手段包括X射线晶体衍射技术、核磁共振技术和冷冻电子显微镜技术, 然而这些技术最大的共同瓶颈在于, 它们所解析的分子结构都无法真正反应生物大分子在生理条件下的真实状态, 而只是给出了在特定人工条件下的生物大分子的结构最优解(最低能量状态). 生物大分子结构的最基本特征之一是其高度动态性, 而这种动态性往往来源于生物大分子结构中的可变区域/柔性区域, 因此可以认为生理状态下的生物大分子结构是四维的, 而使用现有的方法只能做到对三维静态结构的解析. 如果结构生物学在底层理论和分析方法上没有突破, 那么建立在生物大分子结构信息基础之上的生物学功能研究和基于生物大分子结构的药物设计都可能成为无根之木、无源之水. 因而, 人工智能的引入为探究生物大分子结构的跨越式发展提供了理论可能性.

尽管由Google和DeepMind联合开发的AlphaFold系列算法已经在蛋白质结构预测领域取得了优异的效果^[17,48], AlphaFold2更是预测了人类蛋白组98%的蛋白质结构^[17], 但生物大分子的研究还有很多问题尚未解决. 本文认为可以利用人工智能的原理和方法从以下四个角度展开研究工作.

(1) 缺失结构信息的修复. 现有的几种结构生物学技术在采集结构信息的过程中都不可避免地会出现信息缺失的情况, 因此现有的结构数据库存在严重的数

据完整性问题. 可以通过深度学习算法(如卷积神经网络)使用现有的结构数据来训练AI, 实现缺失信号的修复, 提高信噪比, 从根本上解决结构数据的完整性问题.

(2) “高维”动态结构的预测. 通过AI深度学习可建立低维结构和高维结构之间的关联, 实现蛋白质和核酸动态结构的预测. 例如, 可以利用冷冻电子显微镜技术或核磁共振技术分析蛋白质结构中柔性区域的动态结构, 并使用所获得的实验数据训练AI模型, 以实现从蛋白质的三维结构出发预测其“四维结构”.

(3) 生物分子相互作用研究. 生命过程的本质在分子层面最终要归因于生物分子之间的结合与相互作用, 包括蛋白-蛋白相互作用(细胞内信号转导)、核酸-蛋白相互作用(基因转录调节)、核酸-核酸相互作用(蛋白质翻译调节)、化学小分子-蛋白相互作用(药物作用于靶点)等, 而这些决定生物学功能的分子间结合/相互作用又以生物大分子的结构为基础, 因此结构生物学不仅要研究单个生物大分子的空间结构, 更重要的是要研究不同生物分子间的结合与相互作用. 结构生物学现有的理论和方法指导单个生物大分子的结构研究已然捉襟见肘, 对生物大分子间的结合与相互作用研究的复杂度更是远远超出了现有研究范式的能力. 通过引入AI, 可以从已有的分子间结合的结构数据出发, 建立专门适用于研究分子间结合的机器学习方法, 系统地解析与预测分子间的结合与相互作用.

(4) 破译RNA结构. 1953年DNA双螺旋结构的破译已经成为生物学发展史上的里程碑事件, 然而对于同样由四种碱基组成的RNA的空间结构至今没有理论和实践的突破, 唯一的例外是转运RNA的倒三叶草结构. 究其原因, 一方面长期以来RNA的结构研究并不被重视, 人们一直认为, RNA作为遗传信息表达过程的中间载体其生物学功能主要来自于碱基序列而非空间结构; 另一方面, 和双链DNA的相对规则的双螺旋结构不同, RNA分子往往以单链形式存在, 长短不一, 柔性大, 化学稳定性差, 因而现有的结构生物学方法很难适用于RNA分子. 冷冻电子显微镜和人工智能技术的发展使人们看到了破解RNA结构的曙光. 例如, microRNA(miRNA)是一类能够在细胞内和细胞外稳定存在的具有重要生物学功能的非编码RNA分子, 其长度相对固定(22~24个碱基), 功能相对单一(通过结合信使RNA调节基因表达), 因此是利用人工智能进行

RNA结构研究的理想突破口. 通过开展miRNA的大规模人工合成和结构数据采集工作, 在此基础上使用人工智能方法来分析各种碱基排列组合和miRNA三级结构的对应关系, 有望提出RNA结构的新理论, 实现从miRNA一级结构到三级结构的精准预测.

3.2 AI与精准医学和整合生理学

后基因组时代所提出精准医学是一种新型医学概念与医疗模式, 其本质是使用基因组、蛋白质组、代谢组、表观组等组学技术, 对大样本人群与特定疾病类型进行生物标记物的分析、鉴定和验证, 从而精确找到疾病的病因和治疗靶点, 并对同一种疾病的不同状态和不同阶段进行精确分类, 最终实现对患者进行个性化精准治疗的目的. 精准医学能够成为现实是建立在组学技术革命所带来的生物大数据的爆炸性增长基础上的. 随着测序成本的不断下降, 人类已经积累了海量的生物学大数据. 理论上讲, 现在已经可以实现对地球上所有人类个体的各个层次的组学数据的收集, 这意味着人类已经读到控制生命和疾病的“密码本”, 然而目前的生物统计学和生物信息学方法却并不足以帮助人们破译这些密码, 已经挖掘出来的能够指导医学实践的有用信息不到万一, 因为现有的分析方法无法做到多层次多因素的关联性分析, 引入人工智能的方法是解决目前精准医学所遇瓶颈的必由之路.

精准医学的研究对象是疾病, 疾病的表型从微观到宏观依次可以分为分子表型(分子水平)、病理表型(细胞和器官水平)和临床表型(人体水平)几个层次, 各种组学测序所实现的是对分子表型的分析, 而病理表型和临床表型数据主要从医疗实践中获得. 精准医学的核心问题是如何建立分子表型、病理表型和临床表型之间的关联性和因果关系. 通过引进人工智能方法来分析现有的组学数据, 通过多层次多因素关联性分析找到分子表型和病理表型、临床表型之间精准的因果关系, 有望实现仅通过分子表型就可以回答疾病相关的几个关键问题: (i) 一个健康人得这种病的概率有多大(预防/检测); (ii) 病人为什么会得这种病(诊断); (iii) 病人的预后如何(分型); (iv) 病人的最佳个体化治疗靶点是什么(治疗).

3.3 面向人工智能生物学问题的人工智能方法

有别于其他自然学科的数据, 生物学数据有其独

特的特点. 以转录组数据为例, 其主要特点是: (i) 大数据量; (ii) 高维数、小样本; (iii) 非线性; (iv) 高噪音; (v) 数据分布不均衡^[49]. 通过检索美国国立生物技术信息中心(National Center for Biotechnology Information, NCBI)的GEO(Gene Expression Omnibus)数据库和欧洲生物信息研究所(EMBL-European Bioinformatics Institute, EMBL-EBI)的ArrayExpress数据库. 截至2021年7月31日, GEO收录了15万多套数据, 来自22442个平台, 包含455万多个样本, 涉及到4000多个物种, 其中人类的数据就有69312套, 来自5882个平台, 包含将近235万个样本; ArrayExpress数据库收录了74622套数据, 包含255万多个样本, 数据量达60.69 TB. 进一步分析这两个数据库显示, 对于每个样本来说, 大多数样本特征维数超过3万(已知的基因、转录本数目远超这个数值); 但是因为组织样本很难获得、检测费用较高等原因, 一个数据集的样本数往往不会很大, 如GEO包含人类数据有69312套, 包含近455万个样本, 平均每套数据只有约65个样本, 即对每个基因来说, 其维数仅为约65. 由这种数据建立的预测模型非常容易陷入“过拟合(overfitting)”的问题. 大多数生物学数据本质上都是非线性的, 传统的线性模型很难完美拟合这些数据. 另外, 生物学数据产生过程的每一步, 无论是样本制备, 还是实验操作, 无论是机器, 还是实验人员, 都会引入噪声, 这也增加了理解数据的难度. 传统的数据分析方法总是希望数据集里的数据分布是均衡的, 如果数据分布不均衡, 则易于产生分析上的困难或错误, 而生物学实验中, 不同种类的样本获得难度可能相差很大, 造成某类容易获得样本(如癌症组织)数量过多, 而某类不易获得的样本(如癌旁正常组织)又过少. 同时很多数据还存在复杂性和不确定性, 更是给使用人工智能模型去拟合这些数据带来了巨大的挑战.

目前的人工智能算法, 尤其是深度学习算法在处理低维度(如声音、图像及视频数据)和简单结构数据上具有很强的优势, 大多数算法也是针对这些数据类型及其数据特点进行开发和优化. 若直接将已有的人工智能算法嵌套进生物学数据中, 很大程度上不能保证模型的鲁棒性, 更无从挖掘背后的生物学基本规律. 因此, 开发面向生物学问题的人工智能方法显得尤为重要.

现有的人工智能算法中, 监督学习仍然是一个主

要的研究方向. 这类算法需要数据本身具有已知标记(如针对离散数据的类别、针对连续数据的监督数值). 在实际的生物数据中, 很多生物学样本本身的标记并不单一, 存在多标记的情况, 一些临床样本还可能存在标记不够确定, 甚至没有标记的情况(即弱标记或无标记). 在图像识别等领域, 研究者可以通过人工标记明确样本标记情况, 但是生物学数据的不确定性使得这一过程很难完成, 因而也限制了人工智能算法在生命科学中的应用.

对于人工智能生物学的终极目标而言, 监督学习是远远不够的, 发展非监督学习可能是更有前途的方向. 以AlphaGo和AlphaGo Zero的区别为例, 前者的训练集是人类已有的围棋棋谱(监督学习), 而后者并没有学习人类围棋的经验, 仅从“第一性原理”出发(对围棋来说就是两眼活, 占的面积大), 通过机器智能的自我博弈, 就能找出问题的最优解和人类没有认识到的客观规律. 生命的“终极第一性原理(生命和生命分子对地球上声、光、电、磁、引力及pH、离子的理化因素的反应及共进化适应)”要远比围棋复杂, 目前还未能提炼出来. 但是否可以利用AI的特性, 代入生命研究的各个层面上的“第一性原理”, 整合各个层面的实验数据, 先解决各个层面的问题, 逐渐向“终极第一性原理”逼近?

最后, 深度学习算法尽管在实际应用中有着很强的性能, 但是算法本身是一个“黑箱模型”, 即算法构建者无法解释模型中成千上万个参数的意义, 因此对人们理解模型、通过模型挖掘生物学基本规律造成了巨大的困难. 目前人工智能学家研究的重点仍在算法开发上, 努力去提升算法及模型的鲁棒性, 然而, 如何解释模型、理解模型、挖掘模型背后的规律却令很多科学家望而却步^[9]. 目前, 通过人工智能算法挖掘生理现象及疾病背后因素的工作仍然停留在传统的机器学习模型上^[50,51], 而深度学习算法在生命科学和医学领域的应用几乎全部集中于图像识别及可以转化成图像识别的问题上.

综上, 为了更好地将人工智能与生命科学研究相结合, 真正形成人工智能生物学(图1), 从而实现生命活动各因素间因果关系的有效挖掘、探究生命活动背后的基本生物学规律这一根本目标, 建立新的面向人工智能生物学问题的人工智能方法势在必行.

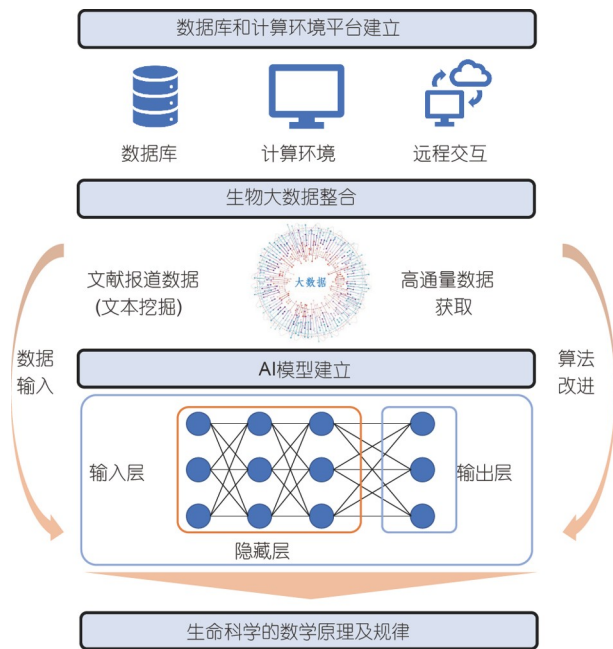


图 1 人工智能生物研究思路

Figure 1 Schematic of proposed research workflow of AI Biology

4 人工智能生物学和其他相关学科的区别与联系

生物学的发展历史上,应用理论和计算方法研究生物学的尝试层出不穷,并已经发展出生物信息学、计算生物学、生物统计学、理论生物学、系统生物学等学科。人工智能生物学和这些学科之间既有联系也有区别(表1)。比如,生物信息学/计算生物学主要是利用统计学方法,以及自行发展出来的序列分析方法从各种组学数据中找出规律性。经典生物信息学中的序列分析方法最大的问题是普适性差,针对每一个问题需要发展一套新的算法,运算速度慢,准确性也有待提高。生物统计学是一门利用统计学原理主要解决生物学研究中数值统计问题的学科,简单和复杂的模型都有,普适性比较好,但准确性比较低。生物信息学和

生物统计学虽然在一定程度上已获得了一些重要的研究成果,但仍然可以认为仅是利用计算机的计算能力帮助人处理大数据,而非直接从大数据分析中获得一般规律。它们所使用的基本算法基础依然主要是人类思考问题的模式而非机器学习的模式,无法从根本上跨越人类目前认知生命系统的限制。理论生物学主要是用物理和化学的思想、方法来探索生物学问题,模型非常复杂、普适性比较差、准确性极低。系统生物学虽然设定了从整体、系统层面上分析生命现象的一般规律这一目标,但缺乏具体的可执行的研究策略和技术手段,尚未形成体系。人工智能生物学所应用的AI理论与方法既可以模拟人类研究的特点,也可以完全摆脱一般人类的研究思考范式,而完全采用机器智能对生物学大数据进行基于深度学习和其他人工智能方法的分析。人工智能生物学在技术上不仅可实现系统生物学所设定的目标,而且可望获得某些根本性的研究突破,彻底阐明生命科学领域大量悬而未决的基本问题。然而,人工智能生物学也有其自身的缺点,最大的缺点在于虽然机器学习算法与物理和化学方法、序列分析方法、统计学方法相比较,其准确性和通用性都有显著的提升,但是缺乏可解释性,因此上文提到的面向人工智能生物学的人工智能方法的研究极其重要。

5 总结与展望

生命活动的本质是生物分子间的相互作用与相互调控。生命科学通过研究组成、影响和调节生命现象的各个要素在特定时空中的相互作用,从而发现和理解生命的一般规律。生命科学研究取得的重大进展往往依赖于物理和化学等学科的技术手段的发展。通过技术革新,生命科学研究不断向微观层面深入,人们对生命活动的物质基础也有了较深的理解。尽管如此,现有的生命科学研究仍停留在分子层面上探讨

表 1 人工智能生物学与相关学科的特点比较

Table 1 Comparison of the characteristics between AI Biology and related disciplines

特点	人工智能生物学	生物信息学	生物统计学	理论生物学	系统生物学
复杂性	高	高低均有	高低均有	高	
普适性	高低均有	低	高	低	未形成体系
准确性	高	中	低	低	

生物的分子组成与变化及其与宏观表型之间的相关性和线性因果关系。由于生命体系的调节和控制往往是多因素协调作用的结果, 这些因素之间存在一对多、多对一以及多对多的关系, 这种极端复杂性导致当前生物学研究范式下得到的相关性结果并不足以在各个层面上阐明生命系统因果关系的基本规律。近年来, 随着自动化高通量实验技术的飞速进步带来的生物学数据的爆发式增长, 以及基于大数据和深度学习的现代人工智能技术的发展, 人工智能和生命科学研究的深度结合已成为可能, 并将推动生命科学的研究范式从“湿实验”和“干实验”各自为政向计算和理论驱动的“干湿结合”方向转变, 进入一个生命科学研究的新时代。

科技进步并不是线性和匀速的, 在某个特定的历史时期, 会呈现出跨越式的大发展。以物理学为例, 20世纪初, 在牛顿力学为代表的经典物理学(物理学V2.0)的发展成熟的基础上, 出人意料地产生了以量子力学和相对论力学为代表的现代物理学(物理学V3.0), 对人类社会的科学技术、社会经济和文明进程等几乎所有领域都产生了物理学V2.0时期无法预计的巨大影响。一百年后的当下, 伴随着人工智能理论和技术的极速发展, 生物学V3.0——人工智能生物学已经呼之欲出, 人工智能正在或者即将显著地改变传统生物学的研究范式, 并通过获得新的知识来解决生命现象中的基本问题, 势必会为人类带来再一次的科技冲击和社会进步。

参考文献

- 1 Darwin C. On The Origin of Species, 1859. London: Routledge, 2004
- 2 Mendel G. Experiments in plant hybridization (1865). Verhandlungen des naturforschenden Vereins Brünn, accessed on 1 January 2013. Available from: URL: www.mendelweb.org/Mendel.html
- 3 Watson J D, Crick F H C. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 1953, 171: 737–738
- 4 Mendeleev D, Moseley H. The periodic law. Chemists, 2014, 97
- 5 Soddy F. The Radio-elements and The Periodic Law. Cambridge: Harvard University Press, 2013
- 6 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521: 436–444
- 7 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, 529: 484–489
- 8 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*, 2017, 550: 354–359
- 9 Castelvocchi D. Can we open the black box of AI? *Nature*, 2016, 538: 20–23
- 10 Bengio Y. Learning Deep Architectures for AI. Hanover: Now Publishers Inc., 2009
- 11 Tai Y W, Liu S, Brown M S, et al. Super resolution using edge prior and single image detail synthesis. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco. 2010. New York: IEEE, 2010. 2400–2407
- 12 Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution. In: Fleet D, Pajdla T, Schiele B, et al., eds. Computer Vision—ECCV 2014. ECCV 2014. Lecture Notes in Computer Science. Cham: Springer, 2014. 184–199
- 13 Zhang H, Hu Z, Luo C, et al. Semantic image inpainting with progressive generative networks. 2018 ACM Multimedia Conference. Seoul. 2018
- 14 Elharrouss O, Almaadeed N, Al-Maadeed S, et al. Image inpainting: a review. *Neural Process Lett*, 2020, 51: 2007–2028
- 15 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in Neural Information Processing Systems 27 (NIPS 2014). Montreal. 2014
- 16 Kingma D P, Welling M. Auto-encoding variational bayes. arXiv, 2013, 1312.6114
- 17 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596: 583–589
- 18 Ritchie M D, Hahn L W, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 2001, 69: 138–147
- 19 Wang J Y, Wen L L, Huang Y N, et al. Dual effects of antioxidants in neurodegeneration: direct neuroprotection against oxidative stress and indirect protection via suppression of gliamediated inflammation. *Curr Pharm Des*, 2006, 12: 3521–3533
- 20 Haskell W L, Berra K, Arias E, et al. Multifactor cardiovascular disease risk reduction in medically underserved, high-risk patients. *Am J Cardiol*, 2006, 98: 1472–1479
- 21 Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. *Genome Biol*, 2011, 12: R60

- 22 Qian N, Sejnowski T J. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol*, 1988, 202: 865–884
- 23 Takefuji Y, Lin C W, Lee K C. A parallel algorithm for estimating the secondary structure in ribonucleic acids. *Biol Cybern*, 1990, 63: 337–340
- 24 Haussler D K D, Eeckman M G R F H. A generalized Hidden Markov model for the recognition of human genes in DNA. *Proceedings International Conference on Intelligent Systems for Molecular Biology*. St Louis, 1996. 134–142
- 25 Reese M G, Eeckman F H, Kulp D, et al. Improved splice site detection in genie. *J Comput Biol*, 1997, 4: 311–323
- 26 Burge C B, Karlin S. Finding the genes in genomic DNA. *Curr Opin Struct Biol*, 1998, 8: 346–354
- 27 Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol*, 1999, 6: 281–297
- 28 Durbin R, Eddy S, Krogh A, et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press, 1998
- 29 Wang H J, Chen R S, Ni X S, et al. Neural network methods for predicting the secondary structure of proteins (in Chinese). *Acta Biophys Sin*, 1989, 5: 422–427 [王化军, 陈润生, 倪向善, 等. 预测蛋白质二级结构的人工神经网络方法. *生物物理学报*, 1989, 5: 422–427]
- 30 Sun Z, Rao X, Peng L, et al. Prediction of protein supersecondary structures based on the artificial neural network method. *Protein Eng Des Sel*, 1997, 10: 763–769
- 31 Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol*, 2001, 308: 397–407
- 32 Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 2001, 17: 721–728
- 33 Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 2003, 302: 449–453
- 34 Brock D C. Learning from artificial intelligence's previous awakenings: the history of expert systems. *AI Mag*, 2018, 39: 3–15
- 35 Haenlein M, Kaplan A. A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. *CA Manage Rev*, 2019, 61: 5–14
- 36 Ching T, Himmelstein D S, Beaulieu-Jones B K, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*, 2018, 15: 20170387
- 37 Webb S. Deep learning for biology. *Nature*, 2018, 554: 555–557
- 38 Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 2015, 33: 831–838
- 39 Liang M, Li Z, Chen T, et al. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE ACM Trans Comput Biol Bioinf*, 2014, 12: 928–937
- 40 Zhang S, Zhou J, Hu H, et al. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res*, 2016, 44: e32
- 41 Xu Y, Dai Z, Chen F, et al. Deep learning for drug-induced liver injury. *J Chem Inf Model*, 2015, 55: 2085–2093
- 42 Chuai G, Ma H, Yan J, et al. Deepcrispr: optimized crispr guide rna design by deep learning. *Genome Biol*, 2018, 19: 1–8
- 43 Duan B, Zhu C, Chuai G, et al. Learning for single-cell assignment. *Sci Adv*, 2020, 6: eabd0855
- 44 Kermany D S, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 2018, 172: 1122–1131.e9
- 45 Munir K, Elahi H, Ayub A, et al. Cancer diagnosis using deep learning: a bibliographic review. *Cancers*, 2019, 11: 1235
- 46 Razzak M I, Naz S, Zaib A. Deep learning for medical image processing: overview, challenges and the future. In: Dey N, Ashour A, Borra S, eds. *Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics*. Cham: Springer, 2018. 323–350
- 47 Crick F. Central dogma of molecular biology. *Nature*, 1970, 227: 561–563
- 48 Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 2020, 577: 706–710
- 49 Cui Q H. Analysis of DNA microarray data (in Chinese). Dissertation for Doctoral Degree. Institute of Automation, Chinese Academy of Science, 2005 [崔庆华. 基因芯片数据分析. 博士学位论文. 北京: 中国科学院自动化研究所, 2005]
- 50 Liang L, Rasmussen M L H, Piening B, et al. Metabolic dynamics and prediction of gestational age and time to delivery in pregnant women. *Cell*, 2020, 181: 1680–1692.e15
- 51 Shen B, Yi X, Sun Y, et al. Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell*, 2020, 182: 59–72.e15

Artificial intelligence biology—biology V3.0

ZHOU Zhen^{1,2}, YAN Chao^{1,2} & ZHANG ChenYu^{1,2}

1 State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210023, China;

2 Institute of Artificial Intelligence Biomedicine, Nanjing University, Nanjing 210031, China

Despite significant progress, modern biology research is facing a huge bottleneck in its development due to lack of revolutionary theoretical frameworks. The rapid development of artificial intelligence (AI) based on big data and deep learning is spurring “Biology V3.0”—Artificial intelligence biology (AI Biology, AIBIO). We define AI Biology as the science of using the principles and means of AI to study the basic laws of living systems. AI Biology is characterized by its ability to integrate multi-dimensions and multi-factors involved in the dynamic process of intermolecular interactions and regulations, so as to truly understand the basic principles in life sciences. As a new life science discipline, AI Biology has the potential to fundamentally change the existing paradigm of biological research, to expand both the connotation and the complexity of biological research, to answer some of the fundamental questions in biological science, and to achieve substantial breakthroughs in key areas of life science and medical science.

artificial intelligence, artificial intelligence biology, machine learning, deep learning

doi: [10.1360/SSV-2021-0298](https://doi.org/10.1360/SSV-2021-0298)