DOI: 10.13376/j.cbls/2023107
文章编号：1004-0374(2023)08-0977-07

睿 ● 观 ● 察

确定性思维模式在生命科学领域面临的挑战

吴家睿[1,2]
(1 中国科学院分子细胞科学卓越创新中心，上海 200031；2 上海交通大学安泰经济与管理学院，上海 200030)

面对充满未知的世界，人类需要通过科学来进行探索，进而获得确定性能力，让世间万物可认识、可解释、可控制。19 世纪法国著名科学家拉普拉斯 (Laplace P) 的一段名言或许最能反映出研究者对确定性之渴望：“我们可以把宇宙现在的状态视为其过去的果以及未来的因。如果一个智者能知道某一刻所有自然运动的力和所有自然构成的物件的位置，假如他也能够对这些数据进行分析，那宇宙里最大物体到最小粒子的运动都会包含在一条简单公式中。对于这智者来说没有事物会是含糊的，而未来只会像过去般出现在他面前”。

从拉普拉斯的这段话里，可以看到研究者对确定性之理解涉及两个层面。首先是本体论层面，即世界之本原是确定的，一切事物的存在和运行都是被决定了的，皆服从构成事物活动规律的因果关系。对确定性的第二层理解是属于认识论方面，即知识之本质是确定的，只要人们不断努力，就能提升知识的完备性，从“相对真理”递进为“绝对真理”，并揭示出事物之间确定的因果关系及其运行规律。

但是，这种对确定性的信念或许是研究者的一厢情愿。回望宇宙乃至生命的整个演化过程，展现出的其实是充满了偶然性的事件；而科学研究则始终存在着局限性乃至不确定性。美国著名物理学家费曼 (Feynman R) 在其《科学与宗教的关系》一文中提出自然界是不确定的 —— “一个未知的东西之所以为未知，首先是因为人们认识到它是未知的，然后才有所谓探索；这里面包含一个要求，要求人们不要去回答不能回答的宇宙秘密；这里面包含一种态度，即承认一切都是不确定的”。而在《科学的不确定性》一文中，费曼教授则进一步指出，在科学的本质中存在不确定性：“我们在科学研究中所说的一切，所得出的所有结论，都具有不确定性，因为它们只是结论。它们是关于会发生什么事情的猜测。你不可能知道会发生什么，因为你不可能进行最完备的实验”。

作为现代科学的一个重要组成部分，20 世纪初逐渐成型的生命科学是建立在确定性思维的基础之上；其中奥地利物理学家薛定谔 (Schrödinger E) 在 1944 年发表的《生命是什么》一书扮演了重要的角色。该书奠定了认识和研究生命的基本理论 —— 还原论，即生物体与非生命物体没有本质的区别，都要遵循严格的物理和化学的规律，生物体内一切活动或过程的发生发展都有着确定的因果关系；而生命科学研究的主要任务就是去揭示这种确定的因果关系。美国著名肿瘤生物学家温伯格 (Weinberg R) 对此有过一个很好的总结：“在 20 世纪，生物学从传统的描述性科学转变成为一门假设驱动的实验科学。与此紧密联系的是，还原论占据了统治地位，即对复杂生命系统的理解可以通过将其拆解为组成的零部件并逐个地拿出来进行研究”[1]。

还需要指出的是，在从 19 世纪以观察和描述为主的传统生物学范式转变为 20 世纪以实验和分析为主的现代生命科学范式的过程中，研究者不仅采用了还原论作为确定性思维模式的理论框架，而且发展出了高度简约化的实验规范作为确定性思维模式的逻辑工具，如孟德尔 (Mendel G) 的豌豆单性状实验确定了生物体的性状是由遗传因子所决定的。2023 年出版的一本新书《被争议的遗传：孟德尔与生物学未来之战》认为，20 世纪初叶在确立孟德尔遗传学的“基因决定论”过程中，尽管许多研究者的实验表明“没有单一遗传机制足以支持孟德尔遗传学”，但当时掌握了话语权的英国科学家贝特逊 (Bateson W) 及其支持者成功地压制了另一位英国科学家韦尔登 (Weldon R) 一派的不同观点 ——

基金项目：中国科学院先导专项“多维大数据驱动的中国人群精准健康研究”(XDB38020000)；上海市科委“科技创新行动计划”软科学研究项目(22692114600)

环境和发育等多种因素对遗传也有影响[2]。孟德尔主义的确立，不仅使得遗传复杂性乃至生命复杂性之认识和研究受到忽视，而且强化了确定性思维在整个生命科学领域的主导地位。这种确定性思维模式甚至延伸到了医学领域——1949 年，美国化学家鲍林 (Paulin L) 在实验室用电泳方法证明，“镰刀型细胞贫血症”源自红细胞内血红蛋白分子异常；鲍林进而提出了全新的疾病观——“分子病”。这一观点随后发展成为一种广为流行的说法：“一个基因一种疾病”。

随着生命科学的发展，尤其是世纪之交“人类基因组计划”的实施，人们进入了一个“后基因组时代”。研究者逐渐发现了生命复杂系统的诸多不确定性，认识到了确定论思维模式的局限性。笔者将从本体论和认识论两个层面总结和分析一下生命科学领域的确定论思维模式当前面临的挑战。

1 生命复杂系统的挑战

在基于确定性思维的研究者眼里，生命是一个由众多基因和蛋白质等零部件拼装而成的简单“机器”；这些生物大分子的空间三维结构决定其相应的生物学功能，进而决定在个体层面的各种生理活动；一旦某个生物分子“零件”的结构出现了异常，通常就会导致生物体产生相应的病理活动。然而，今天的研究进展给出的却是一幅非常复杂的生命“画像”。

1.1 无尽的生物大分子变异

多细胞生物的个体是由数量众多的体细胞组成，如一个成年人个体大约有 30 万亿个体细胞。研究者过去认为，个体内所有体细胞都源自同一个受精卵，通过一次次的细胞分裂方式扩增而来；因此，这些体细胞内的基因组序列都是高度一致的。但是，随着基因组测序能力的提升，尤其是单细胞基因测序技术的出现，研究者看到的则是一个完全不同的图景。

不久前有研究指出，在人类胚胎早期发育过程中，这些胚胎细胞内部广泛发生着各种染色体结构变异，不仅在大多数卵裂期胚胎上发现了具有非整倍体的细胞，而且在随后的分裂球上许多细胞的基因组内也可看到各种大片段 DNA 缺失或扩增[3]。通过单细胞测序技术对人脑部额皮质的神经细胞基因组分析发现，在 13%~41% 的神经细胞内，至少有 100 万碱基大小的基因拷贝数变异 (copy number variant, CNV) 是新产生的[4]。此外，研究者利用诱导干细胞技术分析人体皮肤细胞的基因组，发现近 30% 的成纤维细胞的基因组内具有源于体细胞的 CNV (somatic CNV)[5]。

细胞分裂的核心任务是进行 DNA 复制并把复制后的两份拷贝分配给两个子代细胞。研究者过去认为 DNA 复制是一个“高保真”过程，复制过程中很少出现碱基配对错误，如果偶尔出现一点微小的复制错误，机体还备有若干种错误修复方法来进行修正。但一项基于生物学大数据的分析指出，在 DNA 复制过程中会随机地产生碱基突变 (复制突变) 并传递给子代细胞，在每次基因组复制过程中平均会产生 3 个复制突变；这种复制突变是不可避免的，因此细胞分裂的次数越多，复制突变就越多[6]。该项研究还通过对 69 个国家肿瘤发病率的分析提出，人类肿瘤中三分之二的突变来源于复制突变[6]。还要指出的是，环境对体细胞基因组也能造成随机突变，如紫外线照射能够引起正常人体皮肤的上皮细胞基因组发生突变，大约每 100 万碱基平均出现 2~6 个突变[7]。

随着表观遗传学 (Epigenetics) 的提出和发展，研究者已经认识到，不仅基因组的核酸序列负责遗传信息的传递，而且那些响应内外环境变化的核酸序列或染色质上的各种化学修饰也参与了遗传活动。重要的是，表观遗传修饰不仅用来控制体细胞的基因表达，而且也参与了对子代的遗传控制。例如，不久前的一项研究揭示，在斑马鱼受精卵的发育过程中，来自父本染色体的 DNA 甲基化图谱保留不变，直至囊胚期才被消除重建；而来自母本染色体的 DNA 甲基化图谱则在胚胎发育初期就很快被清除，然后在这些母本染色体上依照父本 DNA 甲基化图谱进行重建；这些父本染色体的 DNA 甲基化图谱在胚胎早期发育中发挥了作用[8]。2023 年初，以色列研究者在《Nature》杂志发表了一项研究，通过分析来自 205 个人体样本的 39 种类型细胞的全基因组 DNA 甲基化图谱，发现个体之间同种细胞类型的 DNA 甲基化模式高度保守，但不同种类的细胞则具有不同的 DNA 甲基化图谱[9]。

从生命科学的“中心法则”来看，通过 DNA 复制进行代际间遗传信息的传递只是生命利用 DNA 的目的之一，生命还需要利用 DNA 来指导 RNA 的合成，进而通过 RNA 指导蛋白质的合成。研究者发现，尽管在转录过程中 RNA 分子是按照碱基配对原则进行合成，但在这些新产生的 RNA 分子之上往往有着不同程度的碱基编辑——RNA

editing。为此，解析不同组织基因表达的研究联合体 ——The Genotype-Tissue Expression (GTEx) Consortium—— 系统地分析了近 9 000 个人体样本的 RNA editing，发现在这些组织样本上广泛存在着 Adenosine-to-inosine (A-to-I) RNA editing；这种 RNA editing 程度在不同组织中有明显的差别，且 RNA editing 程度在编码基因区域要超过非编码的重复序列区域[10]。此外，研究者还发现，通过 DNA 转录产生的各种 RNA 分子也同样被进行广泛的化学修饰。研究者很早就认识到，在 tRNA 分子上有着广泛的化学修饰，如真核细胞中每一个 tRNA 分子平均有 13 个化学修饰。近年来，mRNA 上的化学修饰成了研究的热点；例如，根据对生物医学领域最大的文献数据库 PubMed 的分析，关于 mRNA 序列上腺嘌呤第 6 位的甲基化 ——m^6A 之研究，从 1990 年代至今，相关的研究论文已近 6 千篇。研究者为此专门提出了一个新的学科分支 ——RNA epigenetics。

显然，蛋白质水平的序列信息也与 RNA 水平的序列信息一样，不会完全被 DNA 序列所决定。笔者曾在一篇综述文章里讨论过，在细胞里存在着独立于基因组单核苷酸多态性 (SNP) 的蛋白质单氨基酸多态性 (SAP)[11]。另外一项对人体结直肠癌样本基因组 SNP 与蛋白质组 SAAV (single amino acid variant) 的比较研究也发现，在肿瘤细胞中近 1/4 的 SAAV 没有对应的 SNP[12]。需要指出的是，蛋白质的翻译后修饰更远离基因组的控制，而且这些化学修饰发挥着重要的生物学功能。例如，组成染色质的组蛋白通过多种化学修饰参与基因表达的调控，在细胞内控制各种代谢反应的蛋白酶的活性则往往通过乙酰化修饰进行调节，而细胞周期的运行或细胞信号转导则离不开蛋白激酶催化的磷酸化修饰。可以说，蛋白质的翻译后修饰是蛋白质功能调控的关键手段，其化学修饰类型估计超过 400 种且复杂多变，如组蛋白 H3 尾部有 20 多个氨基酸残基可以被修饰，其中个别氨基酸残基可进行 10 多种化学修饰。

综上所述，生物体内各种类型的生物大分子具有大量形形色色的变异，而且这些变异之间的关系往往表现出相对独立性。首先，在一个个体内众多体细胞之间的基因组具有各种碱基序列变异和长长短短的 DNA 片段差异，以及在碱基序列上广泛存在的化学修饰。其次，RNA editing 和 RNA epigenetics 的存在清楚地表明，生命在转录水平上广泛存在着与基因组 DNA 序列不一致的变异。再次，虽然蛋白质的合成是在基因组的控制之下，但依然存在着各种的氨基酸变异，尤其是其功能的实施基本是在不直接涉及基因组序列信息的化学修饰调节下进行。所有生物大分子的变异有一个共同的目的，就是让机体能够很好地响应体内外环境的变化，满足生理活动的需要。

需要强调的是，这些生物大分子之间的变异与机体的表型或临床表现之间并不是线性关系。人们一般认为基因缺失往往会导致机体的异常表现。不久前一项关于 3 千多人全外显子序列的研究发现，虽然每个个体拥有平均 1.6 个相当于隐性致死突变的功能缺失变异 (recessive-lethal-equivalent loss-of-function)，但这些基因功能缺失变异和临床表现之间并没有明显的相关性[13]。随后的另一项研究也支持了这个结论 —— 研究者通过对 6 万多人的全外显子序列分析，发现了 3 千多个几乎全部缺失或部分缺失蛋白质编码序列的突变基因，但 72% 的突变基因并没有表现出目前已知的人类疾病表型[14]。让情况更为复杂的是，过去人们认为不会影响表型的变异今天却发现并非如此。由于 DNA 的蛋白质编码区存在密码子的“简并性”，有 1/4 至 1/3 的碱基点突变是不会改变蛋白质氨基酸序列的，它们被称为同义突变。因为同义突变不改变蛋白质序列，所以研究者认为这类突变对生物体无害或损害程度很低，属于不会改变生物适应度的中性或近中性突变。最近研究者通过对芽殖酵母基因组中有代表性的 21 个基因之突变体进行分析，发现至少 75% 的同义突变显著损害适应度，且损害幅度超过 0.1%[15]。这一发现挑战了长达半个多世纪的关于同义突变是中性或近中性的观点。

1.2　“万物互联”的相互作用网络

基于还原论的生命科学倾向于从碎片化的角度看待生命，即生物体的生理或病理活动通常是建立在个别基因或蛋白质的结构和功能的基础之上。但是，越来越多的研究工作，尤其是生命组学研究工作表明，生命是一个“万物互联”的复杂系统，每项生命活动都离不开团体合作。例如，传统的基因调控图景通常是具有明晰的指向；从“基因决定论”的观点来看，机体的每个性状都对应着特定的基因，如豌豆的形状或颜色等简单性状由单个基因决定，而身高或血压等复杂性状则由多个基因决定。但研究者之后的分析表明，这种简单的基因与性状之决定论关系只能解释个别基因与简单性状或孟德尔遗

传病之关系，远远不能解释基因组与复杂性状或复杂疾病的关系，在这二者间存在着一个巨大的空白，即“遗传度缺失”(missing heritability)[16]。

2017 年，美国斯坦福大学的研究者提出了一个新的模型 ——“全基因模型”(omnigenic model) 来解释基因组与复杂性状或复杂疾病之关系：基因组里不仅有直接作用于某个特定性状的“核心基因”(core genes)，而且存在着数量更大的与核心基因有相互作用的“外围基因”(peripheral gene)；尽管单个外围基因相比核心基因而言对复杂性状只起到微小的作用，但由于这些外围基因的总数远远超过核心基因，因此来自众多外围基因的微小遗传贡献对复杂性状的调控作用之总和就超过了核心基因；该文作者把这种现象称为“网络基因多效性”(network pleiotropy)[17]。也就是说，涉及复杂性状或复杂疾病的遗传度在整个基因组中广泛传播，基因组内每一个基因通过基因相互作用网络或多或少都对个体的各种复杂性状或疾病有所影响。

蛋白质相互作用网络在生物体中的作用已经得到了广泛的认可。例如，最有名的抑癌因子 P53，自 1979 年被发现至今已经有超过 11 万 6 千多篇关于它的研究论文。P53 通常被定义为一个重要的转录因子，但后来发现该基因还有许多非转录的功能。P53 之所以能够发挥诸多不同的功能，就在于它能够与不同的蛋白质互作而形成不同的相互作用网络。笔者 2012 年的一项研究工作发现，P53 作为经典的肿瘤抑制因子，如果在特定的条件下与某些蛋白质发生相互作用，竟然可以具有促进肿瘤生长的作用 [18]。不久前，研究人员通过规模化的蛋白质相互作用技术检测了肿瘤细胞中数以百万计的蛋白质相互作用，发现突变往往会改变蛋白质之间的相互作用，进而形成新的蛋白质相互作用网络 [19]。因此，研究者不仅要关注肿瘤细胞内突变蛋白自身的功能改变，而且还要考虑突变蛋白与其他蛋白之间新产生的相互作用，以及基于新的蛋白质相互作用网络的功能变异或新功能。

一项对不同物种的蛋白质数量与相互作用网络大小之关系分析发现，虽然研究中统计到的人类蛋白质种类只比果蝇的略微多一点，但是前者的蛋白质相互作用网络比后者的大一倍多 [20]。也就是说，不同物种之间复杂程度的差别与蛋白质相互作用网络的大小高度相关，越是复杂的生命，其蛋白质相互作用就越广泛。还需要指出的是，生物体内的“网络”特征并不仅仅停留在分子层面，而同样存在于细胞、组织、器官等各个层面，且在这些不同层级之间也有着广泛的相互作用。显然，要真正认识个体，尤其具有不同层级的多细胞生物体，就不能只研究生物分子网络，还要研究体内其他层级的网络以及它们之间的关系。2019 年，美国国立卫生研究院 (NIH) 启动了一项名为“人类生物分子图谱计划”(Human Biomolecular Atlas Program, HuBMAP) 的国际合作项目，计划在 7 年的时间里发展各种先进技术，针对各种正常的人体组织开展细胞水平和分子水平的研究，从而建立一个涵盖不同尺度的整合组织图谱 [21]。

随着生命活动的“网络化”现象之普及，过去的“通路”(Pathways) 概念今天已经让位给“网络”(Network) 概念，如“代谢调控通路”和“信号转导通路”转变成了“代谢调控网络”和“信号转导网络”。更重要的是，生物体内广泛存在的“网络”表明，生命活动并不是按照线性通路中那种基于上下游模式的因果关系进行。机体内的“网络”就好比一个复杂的上海地铁网络，通过众多的“站点”把不同的路线联接起来，信息的流动不再是单向的，可以有多条路径；控制往往伴随着各种正反馈或负反馈；输入和输出既可以是单点的，也可以是多点的。例如，传统观点认为肿瘤细胞需要消耗大量的葡萄糖为其供能；但不久前的研究发现，肿瘤细胞会选择谷氨酰胺和脂肪酸当作主要的营养物质来源；最近的一项研究指出，在营养匮乏的环境下，肿瘤细胞能够采用溶酶体将胞外蛋白质分解为氨基酸并作为能量来源加以利用 [22]。

2 生命科学大数据的挑战

还原论指导下的生命科学实验有两个特点，一是高度简约，尽可能少的实验变量和尽可能明确的实验目标；二是偏重定性，以发现新基因或解析蛋白质结构和功能为主要任务。这些特点使得经典生命科学研究属于“小科学”和“小数据”领域。而世纪之交的人类基因组计划则推动生命科学进入了“大科学”和“大数据”领域，进而引出了与经典生命科学不同的研究范式，即数据驱动的数据密集型生命科学研究新范式。这种新范式对追求确定性知识的现代实验生物学带来了新的挑战。

2.1 生命科学大数据的相关性和开放性

现代实验生物学的关注点是揭示生命活动的规律或作用机制，即探寻生物分子之间或事件之间确定的因果关系。尽管寻找生物学机制的研究者在实

验中往往不去区分事件发生的“充分条件”和“必要条件”，但是他们在大多数情况下获得的实验结果实际上只是事件发生的充分条件，并非事件发生的必要条件，更不是满足让一个真正的决定论事件发生所需要的“充分必要条件”。一个根本的原因在于研究者不可能穷尽所有的实验条件，正如费曼所说：“你不可能知道会发生什么，因为你不可能进行最完备的实验”。

从最近报道的两个实验结果可以看到，研究过程的局限性确实很难避免。一项神经科学实验发现，接触小鼠的男性研究人员和女性研究人员会分别让小鼠大脑对氯胺酮产生不同的响应，导致小鼠行为方式出现差异[23]。另外一项肿瘤细胞研究揭示，如果将患者的肿瘤样本从体内取出进行培养，这些肿瘤细胞就从体内缺氧环境下进入到富氧的细胞培养环境下，造成了肿瘤细胞的基因表达和信号通路等发生明显的改变，生长速度变得更快，并更加耐药[24]。如果把科学实验视为猜谜游戏，那么它就是人与自然之间的“无限博弈”。西内克 (Sinek S) 在他的新书《无限游戏》(The Infinite Game) 里写到：在无限游戏中，玩家可以变更，规则不断变化，没有所谓的胜利可言。

基于基因组学和蛋白质组学等组学的数据密集型研究新范式具有全局性的特点，它有力地克服了传统生命科学实验视野过小的短板。例如，在寻找调控血压的遗传因子方面，传统生命科学通常是利用人群小样本开展研究，在数十年里总共发现了274 个遗传位点；2018 年，《Nature Genetics》杂志发表了一项利用 100 万欧洲人样本进行遗传相关性研究的文章，报道了影响血压的 535 个新位点[25]。2022 年，《Science》杂志发表了一项世界上目前规模最大的肿瘤人群全基因组测序工作，从 12 000 多名癌症患者细胞的基因组序列中发现了海量的变异，包括近 3 亿个单碱基置换 (substitutions)、260 多万个双碱基置换 (double substitutions)、1 亿 5 千多万个插入或缺失 (indels) 和近 200 万个重排 (rearrangements)[26]。

现代生命实验科学主要目的是探寻事物之间的因果关系，但从大数据中获得的生命科学知识基本上是相关性的而非因果性的。关于相关性和因果性的详细讨论不是这篇小文能够胜任的。笔者在此只是从生物体是一个“万物互联”网络的观点提出，生命科学实验得到的都是相关性知识而非因果性知识 —— 因为研究者不能控制任何一个生理或病理活动涉及的所有变量。从某种意义上说，因果关系也属于一种特殊的相关关系，即两个事件之间具有唯一变量关系的相关性，能够满足事件发生的“充分必要条件”。

数据密集型研究范式还有一个重要的特点 —— 开放性。传统的生命科学实验是某个科学假设驱动的，为了回答或解决具体的科学问题而进行研究，通常是在已有的知识基础之上并在相应的理论框架里开展，所以研究者的视域和思维往往是受限的。另一方面，由于数据驱动的研究不依赖于假设，因而研究者不仅可以避免现存理论的限制以及对“实验事实”的主观性选择和判断，而且还可以利用各种算法对获得的大数据进行分析，进而能够发现全新的现象或者事物之间隐藏着的内在联系。例如，在一项对瑞典近 170 万名患者的医学档案分析工作中，发现了阑尾与帕金森症发病具有相关性，因为早期切除阑尾的个体患帕金森症的风险明显降低了[27]。

2.2　生命科学大数据的不完备性

生命科学实验范式既然把发现确定的因果关系或规律作为主要目标，因此尽可能地追求结果的完备性通常就成为了评价实验质量的基本要求。这一特点在学术期刊的“同行评议”(peer review) 中有着最为突出的表现。近年来，随着实验技术的进步，期刊评审专家对研究结果完备性的要求也在不断提高，甚至近乎“鸡蛋里面挑骨头”，以至于被广大研究者所“吐槽”;《Science Signaling》主编亚法 (Yaffe MB) 在一篇批评同行评审的社论中指出：“作为编辑，我们需要确保审稿人不要提出过分的要求，不要通过‘移动球门’(move the goalposts) 的方式对新提交的修改稿件进行第二轮或第三轮的评审”[28]。

与之相反的是，数据驱动的生命科学研究新范式并不追求结果的完备性，它采用的是一种全新的工作模式 ——“迭代”(iterate)，即每一次研究工作获得的成果都不是完备的，需要未来研究者在已有版本的基础上不断地进行完善而产生新的版本。人类基因组计划就是生命科学大数据“迭代”的典型：2001 年 2 月，《Nature》发表了人类基因组测序“草图”，它仅仅覆盖了人类基因组 90% 的核酸序列；2004 年 10 月，《Nature》再次发表了人类基因组测序论文，给出了常染色质区域内大约 99% 的核酸序列的测定结果；2020 年 9 月，《Nature》发表了人类第一条染色体没有测序“缺口”的完整核酸序列；2022 年 4 月，以“人类基因组完整序列”为

标题的研究论文在《Science》上发表，研究人员终于时隔基因组草图发表22年之后完整地测定了人类基因组全部碱基序列，而且比2004年发表的版本增加了近2亿个碱基对和近2千个新基因[29]。然而，人类基因组数据的“迭代”远没有结束。2023年5月，《Nature》发表了题为“人类泛基因组参考草图”的论文，通过对采自不同种族的47个人体基因组的分析，为目前的参考人类基因组数据库又添加了1亿多个碱基对[30]。不久前启动的“人类细胞图谱计划”(Human Cell Atlas)要完成30万亿个左右人体细胞的分析，显然更是一个需要无数次“迭代”的生命大科学计划。

近年来，这种认可实验结果不完备性的理念已经拓展到了生命科学整个领域，主要表现为预印本(Preprint)模式的出现和普及。预印本模式是指研究者把未经评审的学术论文直接发布到一个网络开放平台上，供广大用户免费访问和使用。目前在生命科学界最有影响的预印本平台是美国冷泉港实验室在2013年建立的“bioRxiv”(https://www.biorxiv.org/)。该平台在2017年被《Science》杂志评为2017年度十大“科技突破”新闻之一。不同于正规科学期刊那种对实验结果完备性的追求，在预印本平台发表的论文可以是初步的或阶段性的成果，甚至只是阴性结果;作者在预印本文章发表之后可以继续以“迭代”的方式进行更新，即通过修改和完善后作为新版本发布；这些更新后的不同版本都被保留在预印本平台；一个预印本文章的不同版本甚至可以被分别引用。也就是说，在生命科学大数据的“迭代”影响下，经典生命科学也正在接受科学实验的不完备性观念。《Nature》杂志于2023年2月正式推出了一种全新的论文发表模式——注册报告(Registered Reports)，即研究者可以把基于科学假设的实验设计方案以注册的方式向该杂志投稿；如果这份注册报告中提出的科学问题及实验方法通过了同行评议，那么杂志承诺，研究者按照注册的实验方案获得的结果无论是阳性的还是阴性的，都会进行发表;为此，杂志编辑部在题为“自然杂志欢迎注册报告”的社论中强调：“注册报告有助于激励研究，而不管结果如何”[31]。

生命科学领域引入“迭代”模式以及认可实验结果的不完备，表明正在接受“非确定性”的科研观——科学实验是有限的，认知能力是有限的。更重要的是表明研究者认识到，生命复杂系统是开放的，有无限的变异，有无限的连接。把人的有限认知能力和自然的无限可能性结合起来考虑，这就意味着研究者需要放弃确定性思维。笔者在2019年5月纪念p53发现40周年会议的致辞中这样说道：“知识就好像是位于无边无际的自然界‘未知海洋’中小小的‘岛屿’，随着‘知识岛屿’的扩增，相应的‘未知水域’同样也在增长。研究的挑战在于此，研究的乐趣也在于此”。

[参 考 文 献]

[1] Weinberg R. Point: hypotheses first. Nature, 2010, 464: 678

[2] Hall BK. How a scholarly spat shaped genetic research. Nature, 2023, 619: 690-1

[3] Vanneste E, Voet T, Caignec C, et al. Chromosome instability is common in human cleavage-stage embryos. Nat Med, 2009, 15: 577-83

[4] McConnell MJ, Lindberg MR, Brennand KJ, et al. Mosaic copy number variation in human neurons. Science, 2013, 342: 632-7

[5] Abyzov A, Mariani J, Palejev D, et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature, 2012, 492: 438-42

[6] Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science, 2017, 355: 1330-4

[7] Martincorena I, Roshan A, Gerstung M, et al. High burden and pervasive positive selection of somatic mutations innormal human skin. Science, 2015, 348: 880-6

[8] Jiang L, Zhang J, Wang J, et al. Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. Cell, 2013, 153: 773-84

[9] Loyfer N, Magenheim J, Peretz A, et al. A DNA methylation atlas of normal human cell types. Nature, 2023, 613: 355-64

[10] Tan MH, Li Q, Shanmugam R, et al. Dynamic landscape and regulation of RNA editing in mammals. Nature, 2017, 550: 249-54

[11] Wu JR, Zeng R. Molecular basis for population variation: from SNPs to SAPs. FEBS Lett, 2012, 586: 2841-5

[12] Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. Nature, 2014, 513: 382-7

[13] Narasimhan VM, Karen A, Hunt KA, et al. Health and population effects of raregene knockouts in adult humans with related parents. Science, 2016, 352: 474-7

[14] Lek M, Konrad J, Karczewski KJ, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature, 2016, 536: 285-91

[15] Shen X, Song S, Li C, Zhang J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature, 2022, 606: 725-31

[16] Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature, 2009, 461: 747-53

[17] Boyle E, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell, 2017,

169: 1177-86
[18] Song W, Wang J, Yang Y, et al. Rewiring drug-activated p53-regulatory network from suppressing to promoting tumorigenesis. J Mol Cell Biol, 2012, 4: 197-206
[19] Mo X, Niu Q, Ivanov AA, et al. Systematic discovery of mutation-directed neo-protein-protein interactions in cancer. Cell, 2022, 185: 1974-85
[20] Stumpf MPH, Thorne T, de Silva E, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A, 2008, 105: 6959-64
[21] HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature, 2019, 574: 187-92
[22] Pechincha C, Groessl S, Kalis R, et al. Lysosomal enzyme trafficking factor LYSET enables nutritional usage of extracellular proteins. Science, 2022, 378: eabn5637
[23] Georgiou P, Zanos P, Mou TC, et al. Experimenters' sex modulates mouse behaviors and neural responses to ketamine via corticotropin releasing factor. Nat Neurosci, 2022, 25: 1191-200
[24] Kumar B, Adebayo AK, Prasad M, et al. Tumor collection/processing under physioxia uncovers highly relevant signaling networks and drug sensitivity. Sci Adv, 2022, 8: eabh3375
[25] Evangelou E, Warren HR, Mosen-Ansorena D, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet, 2018, 50: 1412-25
[26] Degasperi A, Zou X, Amarante TD, et al. Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. Science, 2022, 376: eabl9283
[27] Killinger BA, Madaj Z, Sikora JW, et al. The vermiform appendix impacts the risk of developing Parkinson's disease. Sci Transl Med, 2018, 10: aar5280
[28] Yaffe MB. Re-reviewing peer review. Sci Signal, 2009, 2: eg11
[29] Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. Science, 2022, 376: 44-53
[30] Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. Nature, 2023, 617: 312-24
[31] Nature welcomes Registered Reports. Nature, 2023, 614: 594