

AI给细胞生物学研究开“外挂”

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/42204.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

AI给细胞生物学研究开“外挂”。



图片来源：AI生成

显微镜下看一眼，只能知道细胞长什么样，但要真正弄清一个细胞在做什么，得看它“打开”了哪些基因。如今，全球科学家已经积累了数亿个细胞的基因数据，量太大，人脑已经处理不过来了。

为此，科学家开发了两个人工智能（AI）模型来帮忙：第一个叫通用细胞嵌入（UCE），可把不同细胞放进同一个坐标系里；第二个叫“TranscriptFormer”，直接学习了12个物种1.12亿个细胞的基因表达数据。这两个模型放在一起建立起“通用空间”，让跨物种比较细胞变得容易得多，也为人类疾病研究和细胞治疗带来了新线索。

这一相关研究成果发表在《自然》和《科学》杂志上。项目的共同主要作者、美国斯坦福生物工程教授斯蒂芬·奎克说，他希望这两篇论文能代表一个全新领域的开始，而后面，还有十年工作等着全球生物学家去做。

搞懂细胞“打开”了哪些基因

每个细胞都带着同一套基因组，但不会把所有基因都打开来用。

胰腺里的 β 细胞会打开胰岛素相关基因，负责制造和释放激素；B细胞会打开抗体相关基因，用来对抗疾病；皮肤里的某类细胞，则可能打开和毛发、色素有关的基因。正因为打开的基因不一样，细胞才分成不同种类。健康细胞和生病细胞不一样，不同物种的细胞也各有各的表达模式。

这几年，科学家构建了不少单细胞图谱，把数万个基因、数亿个细胞的数据存在了数据库里。听起来很壮观，但它也带来一个现实问题：这么多信息，怎么同时看懂？

奎克表示：“人类大脑不可能做到。而AI模型就是来帮我们做这件事的。”他和团队计划用算法处理复杂数据，相当于给研究者们多了一双“外挂”眼睛。

先“建立坐标”再“完形填空”

两个模型分工明确。

UCE先来建立坐标，它能把来自不同实验室、不同组织，甚至不同物种的细胞，都放到同一个抽象的数学空间里。奎克形容说那不是个物理空间，而是一个可以衡量“相似还是不同”的坐标系。地球上所有生物的细胞都可以落到这个空间里，彼此离得近，说明关系近；离得远，说明差异大。这样一来，原先零散拼不起来的图谱，就有了统一的底图。

第二步“TranscriptFormer”上场做完形填空。它的训练方式很像ChatGPT。ChatGPT作为语言模型，会反复看一段缺了词的文本，不断猜缺什么词直到猜准；“TranscriptFormer”也一样，只不过它看的不是词，而是基因表达值。它反复接触大量“缺失某些基因数据”的细胞，学着推断空缺的部分，慢慢就掌握了细胞内部基因协同工作的规律。

模型的训练数据涵盖多达12个物种，除人类和酵母外，还有小鼠、兔子、鸡、斑马鱼、果蝇、非洲爪蛙、导致疟疾的寄生虫、海胆、海绵，以及一种叫秀丽隐杆线虫的小蠕虫。从单细胞酵母到人类，演化跨度非常大。也正因为数据够杂，模型才可能学到跨物种通用的细胞“语言”。

从最简单的谜题，到想象中的细胞

这种模型有什么具体用处？研究团队先拿最简单的海绵做例子。

海绵没有神经元，但科学家一直好奇：它体内哪种细胞会和神经元的“祖先”最为接近？模型给出的答案是，一种叫choanocyte的细胞，和线虫、青蛙的神经元在表达空间上很接近。这可帮助研究者们理解神经元是怎么演化出来的，也重新认识了choanocyte在海绵里的作用。

有意思的是，海绵里还有一种名叫neuroid的细胞，名字听起来很像神经元，但模型发现它的基因表达更像青蛙的腺体，那么，也许它真正参与的其实是消化。

而对于模型没训练过的新物种，只要给出基因表达数据，它也能大致认出有哪些细胞类型。团队还发现，它能成功区分健康细胞和病变细胞。

更远一点想，未来也许可用这个模型“想象”一些现在还不存在的功能性细胞，反过来指导细胞疗法的设计。当然，奎克也强调，这不是让AI直接替代生物学家，恰恰是给有经验的研究者提供一个强大的发现工具，帮他们思考更复杂的问题。

美国国防高级研究计划局、美国国家科学基金会，以及亚马逊公司、英国葛兰素史克公司等多家机构也参与了这一重大研究。对细胞生物学来说，这两篇论文的真正意义，或许不只是又多了一个AI工具，而是让细胞研究开始拥有一个可以跨物种、跨疾病反复使用的共同基座。

作者：张梦然 来源：科技日报

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发