中国科学家创新DNA存储算法

让敦煌壁画再“活”两万年

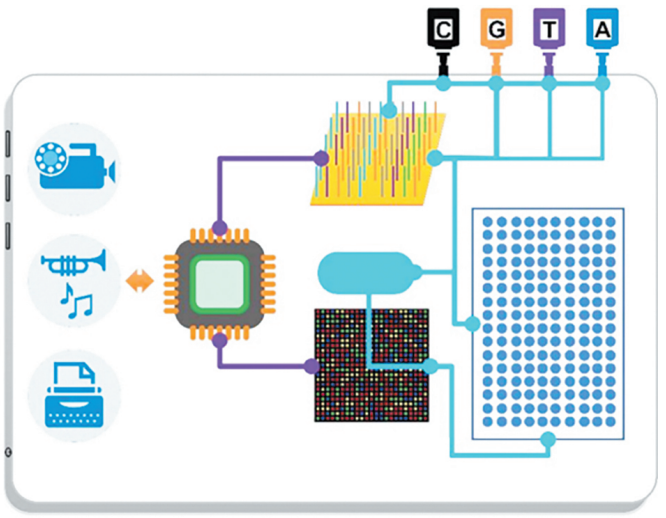
科幻大片《侏罗纪公园》里讲述了这样一个故事：科学家找到一块有史前蚊子的琥珀，从蚊子血中获得了恐龙的基因，从而让已灭绝了6000多万年的恐龙复活。

恐龙的生物信息存储在DNA中，若干年后被提取并还原出来。这听上去似乎有些道理，却也让人倒吸一口凉气。

最近，天津大学一项研究成果让人们离想象又近了一些。该校合成生物学团队将10幅精选敦煌壁画存入DNA中，并通过加速老化等实验，发现这些壁画信息在常温下可保存千年，在9.4℃下可保存两万年。

“如果在合适的温度等条件下，保存千万年也是可以的。”中国科学院院士、天津大学副校长元英进说。

DNA存储的敦煌壁画。



DNA存储技术概念图

小小的DNA却拥有惊人的存储容量

人类文明进化史，也是一部信息存储技术发展史。

从结绳记事、仓颉造字到磁带、硬盘等现代磁光电存储技术，数据存储帮助人类延续了思想，记录下灿烂文明。造纸与印刷术的发明，让人类能够存储的数据量在几百年内获得了大约5个数量级的提升。到了计算机时代，人类产生的数据呈爆发式增长。

“全世界都在建数据中心，而数据中心的能耗是惊人的。”元英进说。

人们一直在不断寻找更海量、更稳定、更安全的存储方式。大自然鬼斧神工的绝妙之处就在于此——最好的存储器或许就藏身于生命体之中。

自地球上出现生命以来，大自然一直用DNA来存储信息，至今已有30多亿年。人类的五官在脸上如何摆放，体内的蛋白怎样合成，眼睛是什么颜色……诸如此类纷繁复杂的人类基因组信息，都记录在比细胞还小得多的DNA上，一代代沿用至今。

不同于各种人造存储设备，DNA极其精巧却又如此经久耐用，它存储了亿万年来无数生物的遗传信息，造就生命繁衍、进化演化及生物多样性。

那么，假如把海量的信息，像存入U盘、硬盘一样，“写”到小小的DNA上，岂不是一举多得？事实上，当人类发现DNA的双螺旋结构后，美俄科学家就先后提出了用DNA存储数字信息的概念。

元英进解释说，DNA存储相较于磁、光、电等常规的信息存储介质有3个最显著的优势。其中最大的优势在于存储密度高。目前，天津大学研究团队将部分经典视频片段存储在DNA中，已实现了体积存储密度比普通硬盘高出6个数量级。

与此同时，存储的信息可用时间非常长。此次研究者将10幅敦煌壁画信息存储在DNA中，结合创新的算法，可以实现DNA分子在室温下保存超过千年，在9.4℃条件下保存两万年。

这样的长期保存需要的能耗却很低。元英进认为，DNA存储被视为一种极具潜力的存储技术，已经成为应对数据存储增长挑战的新机遇。

壁画“变身”DNA需要几步

DNA信息存储的原理共分两步——信息写入和信息读取。

这个过程实际上跨越了极难逾越的鸿沟：它打破了有机与无机的界限，连起生命和信息两大系统。

DNA是脱氧核糖核酸的缩写，含有“A”“T”“C”“G”四种碱基。如果用数字中的0、1、2、3分别代表一个碱基，就组成了一个四进制的存储方式，类似于计算机采用的0和1二进制代码。

通过编码转化，“碱基四进制”和“计算机二进制”就可以实现“对话”。天津大学合成生物学前沿科学中心博士生韩明哲解释说，壁画的数字图像本质上就是二进制的比特串，“我们通过编码将这些二进制的比特串，转化为四进制的ATGC碱基序列，再通过DNA合成技术将碱基序列写入DNA中，壁画的数据图像就‘变’为DNA了。”

此前，该团队成功在酿酒酵母中合成了一条额外的人工染色体，并在上面存储了两张图片及一段视频信息，将其称之为“酵母CD”。随着酵母的不断繁殖扩增，数字信息也随之廉价且稳定地复制。

“我们传代培养酵母到100代，依然可以完美地恢复出原始数据。”元英进说，假如脑洞更大一点，将信息存储到一棵树中，随着树生长千百年，人类的子孙后代都可以随时从这棵树中读取到千百年前存储的信息。

这一次，这支年轻团队的创新之处在于，能实现更恶劣条件下可靠读取信息。韩明哲说，存了壁画信息的DNA，本质上其实跟天然的DNA没有什么不同，同样也存在长时间存放而产生的断裂和降解等问题，影响信息存储的长期可靠性，这也成为亟待解决的关键科学问题。

于是，他们设计了基于德布莱英图理论的序列重建算法来解决DNA断裂等问题，可以从严重降解的DNA样本中，恢复原始的信息。

为了验证数据的长期可靠性，团队制备了一个没有任何特殊保护的DNA水溶液样本，随后在70℃的温度下加速样本断裂、降解长达十周。韩明哲说：“这个过程使得DNA片段80%以上都发生了断裂错误，模拟了DNA在自然环境下千年万年的降解情形。”

随后，团队依靠设计的序列重建算法，依然可以准确组装并解码96.4%以上的片段，再通过一种编码方式解决了少量片段丢失的问题，使原始的敦煌壁画图片能够完美恢复。

DNA存储走向实用化还有多远

尽管DNA存储还不被大众所熟知，但它正在努力走出实验室，“距离实用化并不遥远。”元英进说，惊人的数据存储需求是新技术走向市场的最大推动力。

据国际数据公司估计，到2025年全球数据总量将达到175ZB（1ZB为十万亿亿字节）。到2024年，全球将有30%的数字业务进行DNA存储试验。然而从目前来看，DNA存储想要大规模应用，尤其是在中国实用化还需要突破几个关键瓶颈。

团队分析了当前DNA信息存储面临的主要挑战。信息存储成本高、信息读写速度慢，以及无法高效对接现有信息系统是三大主要限制因素。

根据测算，目前DNA存储写入成本相当于20世纪80年代内存的存储成本，而要达到当前数据存储成本还需要降低7至8个数量级。

“DNA信息存储成本在未来有很大的下降潜力。”韩明哲认为，今后可以从优化合成反应、改良芯片结构、替换廉价耗材、优化试剂分配量等方面着手，大幅降低合成成本。

与此同时，由于信息存储领域市场规模巨大，随着半导体器件、微纳加工在DNA信息存储领域的应用，该领域的巨大投入将对DNA合成技术产生重大影响，DNA合成技术与装备快速迭代升级，也有望使成本快速下降。

DNA信息存储的读取依赖测序技术，与磁、光、电等存储相比，读取速度较慢。目前DNA测序仪的读取速度与硬盘相比，还存在3至4个数量级的差距——现有电、磁存储技术通常每秒可读取几十到几百兆字节数据。此外，DNA存储的标准尚待建立，面临与现有数字存储系统兼容的问题。

“DNA信息存储是一个新兴的、多学科深度交叉融合的研究方向。”元英进认为，DNA存储在未来极有可能成为庞大冷数据存储的主要存储介质。

所谓冷数据，就如同档案馆的历史资料，需要把海量信息保存好，但平时又很少去使用。因为这些数据需要长期存储、耗能又大，而电子存储设备的寿命往往只有十年到几十年，并需要不断更新迭代，难以满足冷数据存储的需要。

DNA存储走向实用化仍面临很多挑战。元英进认为，眼下的突破可能还只是冰山一角，“技术进步需要十年磨一剑的耐心，还需要一点运气。”

据《科技日报》、上观新闻

相关链接

仅200公斤DNA就能装下整个世界大数据

1994年，比尔·盖茨坐在33万纸上，向全世界发布，我们现在有了“光盘”。一张光盘能够记录的内容，也就是33万张纸，这在当时是非常具有震撼性的广告效应。

在数据大爆炸的时代，2020年全世界的数据量是44个ZB，即440万亿亿字节。要把这些数据存下来，光耗电量就是一个长江三峡大坝所产生的电量。

中国科学院院士、上海交大化工学院院长樊春海表示，数据还在不断增长，到2025年预计达175个ZB，而且里面80%至90%是很少被调用的冷数据。

不仅存储耗能巨大，传输也越来越成为问题。1969年阿波罗登月计划时，存储介质还是纸，就是叠起来1人多高的这么多数据。到2019年，人类第一次拍下了黑洞的照片，而把图片信息传回来，数据量达5个PB，相当于1万块硬盘，足有半吨重。

樊春海以一根测核酸的试管比方，推算出一两也就是50克DNA，就可以存下1000万块硬盘的信息。“通过DNA存储，可以把数据电子存储的容量极限提升7个数量级。”他说，这样一来，全世界440万亿亿的字节，只要200公斤DNA就可以存下来了。