

我们可以把所有数据存储到DNA上吗？



许多科学家认为,解决海量数据存储问题的另一种办法在于包含我们遗传信息的生物大分子:脱氧核糖核酸(DNA)。从地球生命诞生至今,DNA已经进化到可以以极高的密度存储大量信息,理论上一个装满DNA的咖啡杯就可以存储世上所有的数据。

我们需要新的解决方案,来存储世界正不断积累的大量数据,尤其是档案数据,DNA的密度甚至是闪存的1000倍。另一个有趣的特性是,DNA聚合物一旦制造出来,它就不会再消耗任何能量。你可以把数据写入DNA,然后永久存储起来。

科学家已经证明,图像和文本可以编码为DNA,但我们还需要一种从许多DNA片段混合物中挑选出所需文件的简单方法。在新研究中,科学家展示了一种方法,能将每个数据文件封装到一个6微米的二氧化硅球形“胶囊”中,并使用DNA短序列作为标签,以显示其文件内容。

利用这种方法,研究人员从包含20张图像的DNA文件中准确提取出了以DNA序列形式存储的单个图像。考虑到可以用到的标签数量,这种方法最多能扩展到10²⁰个文件。

稳定的存储介质

数字存储系统将文本、照片和其他类型的信息都编码为一系列0和1,同样的信息也可以用构成遗传密码的4种核苷酸(A、T、G和C,即腺嘌呤、胸腺嘧啶、鸟嘌呤和胞嘧啶)编码在DNA中。例如,G和C可以代表0,而A和T代表1。

作为存储介质,DNA还具有其他几个特点。首先,它非常稳定,而且合成和测序都相当容易(但目前还十分昂贵)。其次,它具有非常高的存储密度——1个核苷酸相当于2个比特,大约为1立方纳米。因此,以DNA形式存储的数据完全可以放在我们的手掌中。

这种存储数据的新方法面临着诸多障碍,首先就是合成如此大量DNA需要耗费的成本。目前,写入1拍字节(100万GB)的数据需要花费1万亿美元。为了与磁带(通常用于存储档案数据)竞争,估计DNA合成的成本需要降低约6个数量级,这一目标可能会在10年或20年内实现,就像过去几十年来闪存存储信息的成本大幅下降一样。

除了成本之外,使用DNA存储数据的另一个主要瓶颈是,我们很难从所有文件中挑选出想要的文件。

假设写入DNA的技术已经很先进,可以实现在DNA中写入1艾字节或1泽字节(zettabyte,简称ZB,1ZB=1000EB)数据的成本效

益,会发生什么?你会有一大堆的DNA,也就是无数的文件、图像或电影和其他东西,但你需要在其中找到想要的某一张图片或某一部电影,这就像大海捞针。

目前,DNA文件通常使用PCR(聚合酶链式反应)方法来检索。每个DNA数据文件都包含一个与特定PCR引物结合的序列。为了读取某个特定的文件,需要将该引物添加到样品中,找到并放大所想要的序列。然而,这种方法的一个缺点是,引物与目标序列以外的DNA序列之间可能存在串扰,导致不必要的文件输出。此外,PCR的检索过程需要用到酶,最终会消耗库中的大部分DNA,这有点像在干草堆里找一根针,因为其他所有DNA都没有被放大,因此基本上它们都被扔掉了。

解决DNA文件检索难题

麻省理工学院的研究小组开发了一种新的检索技术,希望取代PCR方法。他们将每个DNA文件封装到一个微小的二氧化硅胶囊中,每个胶囊都贴上了由单链DNA组成的“条形码”,与文件内容相对应。为了证明这种方法的成本效益,研究人员将20个不同的图像编码到大约长度为3000个核苷酸的DNA片段中,这大致相当于100个字节(他们的研究还显示,这些胶

囊可以容纳高达1GB的DNA文件)。

据报道,在近期的一项新研究中,美国麻省理工学院的科学家开发了一种标记和检索DNA数据文件的技术,这或许能让DNA数据存储成为可能。

此时此刻,地球上大约有10万亿吉字节(GB)的数据量,而每一天,人类制造出来的电子邮件、照片、社交媒体动态和其他数字文件加起来,又有250万吉字节的数据。这些数据中的大部分都存储在名为“艾字节(exabyte,简称EB)数据中心”的巨大设施中(1EB相当于10亿GB),其规模可能有几个足球场那么大,建造和维护成本约为10亿美元。

囊可以容纳高达1GB的DNA文件)。

研究中的每个文件都有相应的条形码标签,如“猫”或“飞机”等。当研究人员想要提取一个特定的图像时,他们会取出一个DNA样本,加入与目标标签相对应的引物。例如,老虎的图像对应的标签是“猫”“橘色”和“野生”,而家猫的图像对应“猫”“橘色”和“家养”。

这些引物用荧光或磁性颗粒标记,便于从样本中提取并识别匹配片段。通过这种方法,研究人员可以将需要的文件移出来,剩下的DNA则完整地放回去,继续存储数据。他们的检索过程允许布尔逻辑语句,如“总统和18世纪”会生成“乔治·华盛顿”的结果,这很类似谷歌的图像检索。

在目前的概念验证阶段,搜索速度是每秒1000字节(1KB)。文件系统的搜索速度是由每个胶囊的数据量大小决定的,而目前限制数据量大小的因素就是在DNA上写入100兆字节(MB)数据所需的高昂成本,以及可以并行使用的分类器的数量。如果DNA合成变得足够便宜,就能够用这种方法将每个文件存储的数据量最大化。

研究人员所使用的条形码——单链DNA序列——取自哈佛医学院遗传学和医学教授史蒂芬·

埃利奇开发的序列库,其中包含了10万个序列。如果给每个文件贴上两个这样的标签,就可以唯一地标记100亿(10¹⁰)个不同的文件;如果每个文件上有4个标签,就可以唯一地标记10²⁰个文件。

在DNA中写入、复制、读取,以及用DNA进行低能耗的档案数据存储方面,我们取得了快速进步,但这也使得从巨大的数据库(10²¹字节,泽字节规模)中精确检索数据文件变得极为困难,这项新研究引人注目的地方在于,它使用一个完全独立的DNA外层解决了这个问题,扩展了DNA的不同属性(杂交而非测序),而且使用的是现有的仪器和化学试剂。

科学家设想这种DNA封装技术可以用于存储“冷”数据,即保存在档案中但不经常访问的数据。目前,研究实验室已经成立了一家名为Cache DNA的初创公司,正在开发DNA的长期存储技术,既可以用于长期的DNA数据存储,也能用于短期的临床和其他现有的DNA样品存储。

虽然我们可能还需要一段时间才能将DNA作为数据存储介质,但目前在Covid-19检测、人类基因组测序和其他基因组学领域中,对于DNA和RNA样品的低成本和大规模存储的解决方案都有很紧迫的需求。

宇宙可以自己学习吗？

更加稳定的能量状态。

遵循这个想法,研究小组开发了一个可能的框架。根据这个框架,宇宙可以利用矩阵数学方法(一种按阵列排列的数学方法)、神经网络和其他机器学习原理,进行学习。简而言之,他们想知道,宇宙是否是一台可以学习的计算机。

研究的另一位作者、加拿大滑铁卢圆周理论物理研究所的物理学家李·斯莫林说:“我们尝试改变叙述的方式,就像生物学家达尔文须改变叙述,从而加深研究对象的理解一样。”

达尔文式的宇宙

和飞蛾进化出更好的伪装类似,自训宇宙也可能进化到更高的状态——意味着,这可能是一个更加稳定的能量状态。根据研究人员开发的数学框架,这个系统只能前进,每一次迭代都会创造出一个比原先更好、或更稳定的宇宙。我们今天测量到的物理常数,仅在当下有效,而在过去可能是不同的数值。

研究小组发现,某些量子引力和量子场论——即所谓的规范理论,旨在在爱因斯坦的狭义相对论和描述亚原子粒子的量子力学之

间架起桥梁的一组理论,可以被映射或解读为矩阵数学语言,创造出一个机器学习系统的模型。这种联系表明,在机器学习系统的每一次迭代或循环中,其结果可能是宇宙的物理定律。

研究小组在论文中提到,这个学习框架,代表了通向自训宇宙概念的第一“小步”。但是,随着工作的不断推进,团队或许可以创建一个完整的宇宙模型,为了解我们的现实宇宙打开新的大门。

坎宁汉说:“一个令人兴奋的前景是,你可以利用其中的一个模型,或许还可以提炼出一些新的事物。”可能是发现一种新型黑洞,或描绘一个尚未被解释的物理系统,比如暗能量。

然而,也并非所有研究人员都对这个新的想法感到期待。纽约大学的哲学教授蒂姆·莫德林没有参与这项研究。但是莫德林认为,对于这一概念,不仅没有证据可予以支持,相反却有许多证据与之相悖,比如某些已经测得的物理定律,在今日与大爆炸之后的瞬间,并无不同。此外,如果宇宙定律是不断进化的,那么莫德林认为必须有一个更庞大的不变的定律,可以支配这种变化,但这又否定了自学

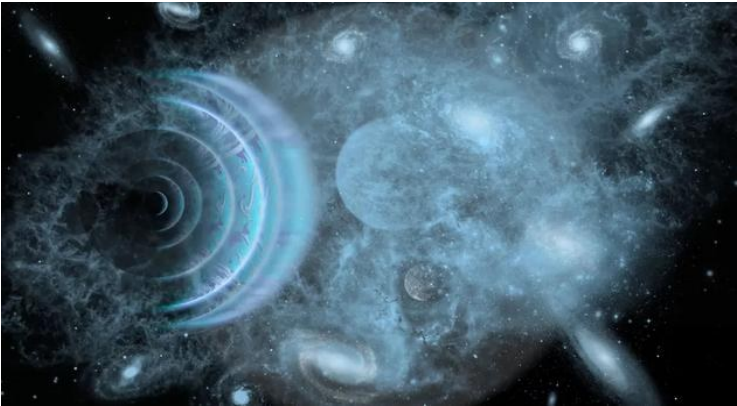
系统的概念。

莫德林说:“我们在观察基本定律时——比如薛定谔方程或广义相对论,它们看起来压根不是随机的。你可以用非常严谨的数学方式把它们表述出来,且没有太多可调整的参数。”

澳大利亚昆士兰大学的哲学家彼得·W·埃文斯也没有参与这项研究。起初,埃文斯也对这项新研究持怀疑态度;但他赞同花一些时间用非正统的方式解决诸如“为什么宇宙是这个样子?”等根本性问题。他认为,这些方式,即便本身没有结果,但或许可以引发意料之外的想法,而这些意料之外的东西或许可以为我们了解宇宙打开新思路。

这项新研究背后的研究人员也承认,他们的工作目前仍处于初步阶段,并非是要作为最终理论,而更多的是以新的方式思考问题的开端。最终,尽管该论文没有就哪种模型可以用来描述我们的宇宙得出任何结论,但它也确实提出了宇宙可以学习的可能性。

坎宁汉说:“到目前为止,我们仍有许多悬而未决的问题,当然我们也没办法证明任何事情。但是,我们的真正目标是抛砖引玉,开启一场新讨论。”



宇宙或许可以自己学习如何进化成一个更好、更稳定的宇宙。近日,一组科学家提出了这个十分新颖的想法。他们称,他们正在重新构想宇宙,正如达尔文改变我们对自然世界的看法一样。

这个颇具争议的想法试图用数学框架来描述物理学中的各种定律——比如量子场论和量子引力等,来解释这些物理定律为什么是我们看起来的样子。结果则是一个类似于机器学习程序的系统。

科学家已经发现无数的物理定律,和具有固定值的量,来定义宇宙。从电子的质量到引力,宇宙中有许多特定的常数,由于其精确性和看似无规律的值,而让一些人觉得似乎有些随意。

研究作者威廉·坎宁汉是量子计算创业公司Agnostiq的物理学家与

软件负责人。他说:“当前,基本物理学的目标之一是,不仅要理解物理定律,还要弄清楚这些物理定律为什么偏偏就是这样。没有明显的理由可以让一组定律优先于另一组定律。”

自学系统

为了解答这个疑问,研究小组想知道我们今天所见到的宇宙模式,是不是只是宇宙曾经的一种模式?也许,我们今天所认识的种种定律,只是许多迭代中的一个。也许,宇宙正在持续地进化。

为了得到一个能够进化的宇宙,研究人员提出了一个叫做“自训宇宙”的概念,意思就是一个可以自学的宇宙。在这种情况下,这个学习的过程,碰巧与机器学习算法的工作方式类似,即某一阶段的反馈会影响到下一阶段的发展,其目的是达到一个