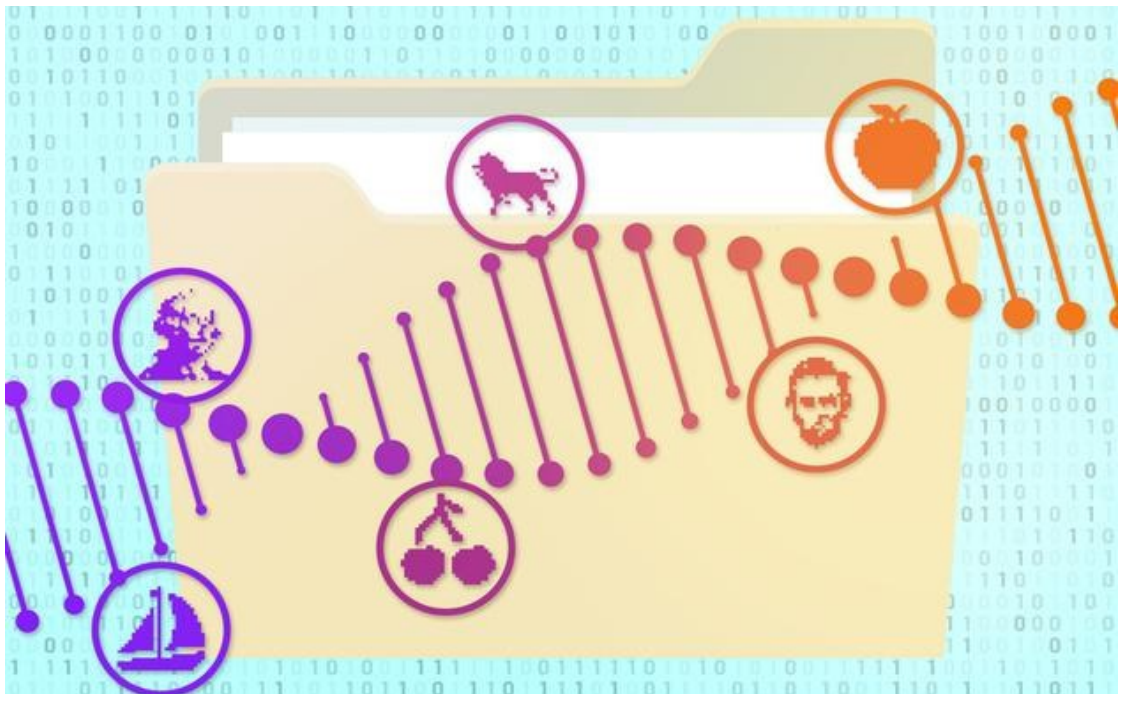
我們可以把所有數據存儲到DNA上嗎?



據國外媒體報道,在近期的一項新研究中,美國麻省理工學院的科學家開發了一種標記和檢索DNA數據文件的技術,這或許能讓DNA數據存儲成爲可能。

此時此刻,地球上大約有10萬億吉字節(GB)的數據量,而每一天,人類制造出來的電子郵件、照片、社交媒體動態和其他數字文件加起來,又有250萬吉字節的數據。這些數據中的大部分都存儲在名爲"艾字節(exabyte,簡稱EB)數據中心"的巨大設施中(1EB相當于10億GB),其規模可能有幾個足球場那麽大,建造和維護成本約爲10億美元。

許多科學家認爲,解決天量數據存儲問題的另一種辦法在于包含我們遺傳信息的生物大分子:脱氧核糖核酸(DNA)。從地球生命誕生至今,DNA已經進化到可以以極高的密度存儲大量信息,理論上一個裝滿DNA的咖啡杯就可以存儲世界上所有的數據。

我們需要新的解決方案,來存儲世界正不斷積累的大量數據,尤其是檔案數據,DNA的密度甚至是閃存的1000倍。另一個有趣的特性是,DNA聚合物一旦制造出來,它就不會再消耗任何能量。你可以把數據寫入DNA,然後永久存儲起來。

科學家已經證明,圖像和文本可以編碼爲DNA,但我們還需要一種從許多DNA片段混合物中挑選出所需文件的簡單方法。在新研究中,科學家展示了一種方法,能將每個數據文件封裝到一個6微米的二氧化硅球形"膠囊"中,并使用DNA短序列作爲標簽,以顯示其文件内容。

利用這種方法,研究人員從包含20張圖像的DNA文件中準確提取出了以DNA序列形式存儲的單個圖像。考慮到可以用到的標簽數量,這種方法最多能擴展到10^20個文件。

穩定的存儲介質

數字存儲系統將文本、照片和其他類型的信息都編碼爲一系列的0和1,同樣的信息也可以用構成遺傳密碼的4種核苷酸(A、T、G和C,即腺嘌呤、胸腺嘧啶、鳥嘌呤和胞嘧啶)編碼在DNA中。例如,G和C可以代表0,而A和T代表1。

作爲存儲介質,DNA還具有其他幾個特點。首先,它非常穩定,而且合成和測序都相當容易(但目前還十分昂貴)。其次,它具有非常高的存儲密度——1個核苷酸相當于2個比特,大約爲1立方納米。因此,以DNA形式存儲的數據完全可以放在我們的手掌中。

這種存儲數據的新方法面臨着諸多障礙,首先就是合成如此大量DNA需要耗費的成本。目前,寫入1拍字節(100萬GB)的數據需要花費1萬億美元。爲了與磁帶(通常用于存儲檔案數據)競争,估計DNA合成的成本需要降低約6個數量級,這一目標可能會在10年或20年内實現,就像過去幾十年來閃存存儲信息的成本大幅下降一樣。

除了成本之外,使用DNA存儲數據的另一個主要瓶頸是,我們很難從所有文件中挑選出想要的文件。

假設寫入DNA的技術已經很先進,可以實現在DNA中寫入1艾字節或1澤字節(zettabyte,簡稱ZB,1ZB=1000EB)數據的成本效益,會發生什麽?你會有一大堆的DNA,也就是無數的文件、圖像或電影和其他東西,但你需要在其中找到想要的某一張圖片或某一部電影,這就像大海撈針。

目前,DNA文件通常使用PCR(聚合酶鏈式反應)方法來檢索。每個DNA數據文件都包含一個與特定PCR引物結合的序列。爲了讀取某個特定的文件,需要將該引物添加到樣品中,找到并放大所想要的序列。然而,這種方法的一個缺點是,引物與目標序列以外的DNA序列之間可能存在串擾,導致不必要的文件輸出。此外,PCR的檢索過程需要用到酶,最終會消耗庫中的大部分DNA,這有點像在幹草堆裏找一根針,因爲其他所有DNA都没有被放大,因此基本上它們都被扔掉了。

解決DNA文件檢索難題

麻省理工學院的研究小組開發了一種新的檢索技術,希望取代PCR方法。他們將每個DNA文件封裝到一個微小的二氧化硅膠囊中,每個膠囊都貼上了由單鏈DNA組成的"條形碼",與文件内容相對應。爲了證明這種方法的成本效益,研究人員將20個不同的圖像編碼到大約長度爲3000個核苷酸的DNA片段中,這大致相當于100個字節(他們的研究還顯示,這些膠囊可以容納高達1GB的DNA文件)。

研究中的每個文件都有相應的條形碼標簽,如"猫"或"飛機"等。當研究人員想要提取一個特定的圖像時,他們會取出一個DNA樣本,加入與目標標簽相對應的引物。例如,老虎的圖像對應的標簽是"猫""橘色"和"野生",而家猫的圖像對應"猫""橘色"和"家養"。

這些引物用熒光或磁性顆粒標記,便于從樣本中提取并識别匹配片段。通過這種方法,研究人員可以將需要的文件移出來,剩下的DNA則完整地放回去,繼續存儲數據。他們的檢索過程允許布爾邏輯語句,如"總統和18世紀"會生成"喬治·華盛頓"的結果,這很類似谷歌的圖像檢索。

在目前的概念驗證階段,搜索速度是每秒1000字節(1KB)。文件系統的搜索速度是由每個膠囊的數據量大小決定的,而目前限制數據量大小的因素就是在DNA上寫入100兆字節(MB)數據所需的高昂成本,以及可以并行使用的分類器的數量。如果DNA合成變得足够便宜,就能够用這種方法將每個文件存儲的數據量最大化。

研究人員所使用的條形碼——單鏈DNA序列——取自哈佛醫學院遺傳學和醫學教授史蒂芬·埃利奇開發的序列庫,其中包含了10萬個序列。如果給每個文件貼上兩個這樣的標簽,就可以唯一地標記100億(10^10)個不同的文件;如果每個文件上有4個標簽,就可以唯一地標記10^20個文件。

在DNA中寫入、復制、讀取,以及用DNA進行低能耗的檔案數據存儲方面,我們取得了快速進步,但這也使得從巨大的數據庫(10^21字節,澤字節規模)中精確檢索數據文件變得極爲困難,這項新研究引人注目的地方在于,它使用一個完全獨立的DNA外層解決了這個問題,擴展了DNA的不同屬性(雜交而非測序),而且使用的是現有的儀器和化學試劑。

科學家設想這種DNA封裝技術可以用于存儲"冷"數據,即保存在檔案中但不經常訪問的數據。目前,研究實驗室已經成立了一家名爲Cache DNA的初創公司,正在開發DNA的長期存儲技術,既可以用于長期的DNA數據存儲,也能用于短期的臨床和其他現有的DNA樣品存儲。

雖然我們可能還需要一段時間才能將DNA作爲數據存儲介質,但目前在Covid-19檢測、人類基因組測序和其他基因組學領域中,對于DNA和RNA樣品的低成本和大規模存儲的解決方案都有很緊迫的需求。