Parallel molecular data storage by printing epigenetic bits on DNA

2024.12.17

穆新宇, 邱天 and 许智威

Contents

1 背景

存储难题

全球数据领域的显著扩展对大规模数据存储提出了迫切挑战，并对更好的存储材料提出了紧迫需求。

> 全球每年产生数据总量达 1YB，其中非结构化数据比例超过 80%
>
> — Huawei

1.b.a 小复习

DNA（脱氧核糖核酸）

一种长链聚合物，组成单位称为核苷酸，而糖类与磷酸借由酯键相连，组成其长链骨架。

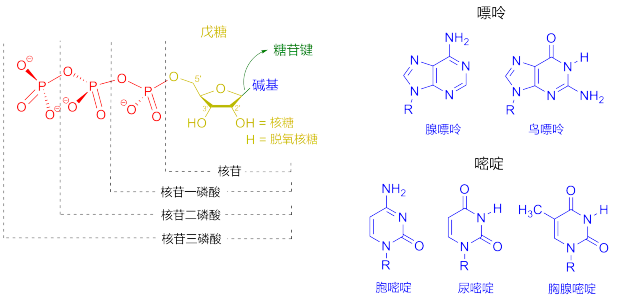

Figure 1: 核苷酸

DNA (ii)

转录（Transcription）

在 **RNA 聚合酶**的催化下，遗传信息由 DNA 复制到 RNA（尤其是 mRNA）的过程。

1. RNA 聚合酶与一种或多种通用转录因子一起结合 DNA 上的**启动子**
2. RNA 聚合酶催化聚合核糖核苷酸（与模板 DNA 链的脱氧核糖核苷酸互补）
3. 在 RNA 聚合酶的作用下形成 RNA 糖–磷酸骨架，进而形成 RNA 链

DNA (iii)

表观基因（Epigenetics）

在人类细胞中，除了 DNA 序列本身之外，还有一种叫做**表观基因**的机制，它能够在不改变 DNA 的基本顺序的前提下，稳定地记录和调控细胞的功能信息。

表观基因的实现：

- **DNA 甲基化**：DNA 分子上的某些胞嘧啶（C）碱基会被添加上甲基（CH_3）基团，这通常会抑制基因的表达。
- **组蛋白修饰**：组蛋白是与 DNA 结合形成染色质的蛋白质。组蛋白的修饰（如乙酰化、甲基化等）会改变染色质的结构，进而影响基因的表达。
- **非编码 RNA**：一些非编码 RNA（如 miRNA、lncRNA 等）也能通过与 DNA 或 RNA 的相互作用，调节基因的表达。

DNA (iv)

1.b.b DNA 作为存储介质的特点

- DNA 存储因其高存储密度和耐久性而引起了关注
- 当前 DNA 存储写入方法及其问题

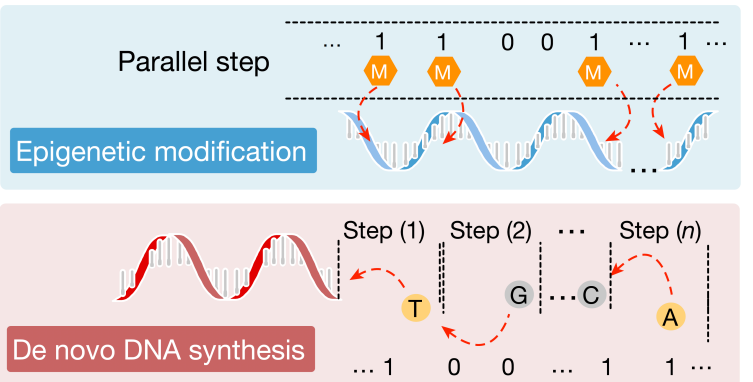

Figure 2: 表观基因修饰 vs. 从头 DNA 合成

DNA (v)

1.b.c epi - bit

文章提出了 epi - bit——一种基于表观基因的并行、可编程、稳定、可扩展的 DNA 数据存取模式，并尝试在大规模数据存取上进行了实验和验证。

2 系统设计

写入流程

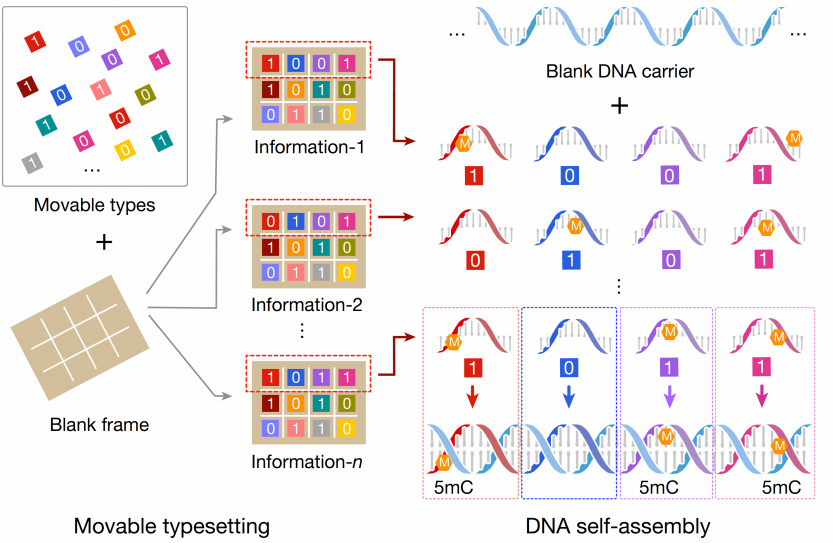

Figure 3: DNA 版本的活字印刷

写入流程 (ii)

类比活字印刷，文章提出的基于表观基因的并行 DNA 写入系统的写入过程大致如下：

1. 制作一些单链 DNA 模板（白纸），将与这些单链 DNA 互补的另一条链打散，形成预制可移动类型集合（Premade DNA movable types，活字印刷中的活字），再将这些可移动类型与模板单链 DNA 混合，发生碱基互补配对

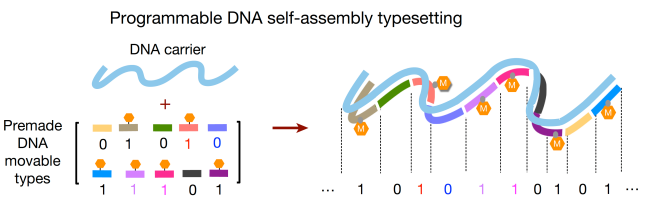

Figure 4: 预制一些 DNA 小片段，与单链 DNA 混合

写入流程 (iii)

2. 利用 DNM1（一种酶）催化 CH_3的挂载，这个过程是**并行的**，这也是这篇文章的精华之处

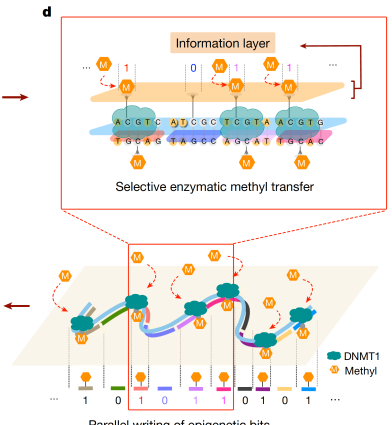

Figure 5: 预制一些 DNA 小片段，与单链 DNA 混合

写入流程 (iv)

2.a.a 一些可能的疑问

- 为什么预制的 DNA 片段集合中有多个 0 和 1?
- DNM1 催化 CH_3挂载时，CH_3是从哪里来的?

单 bit 写入验证

在验证单 bit 的写入是否正确时，文章采用了一些生物和化学方法，这里就不再给出。

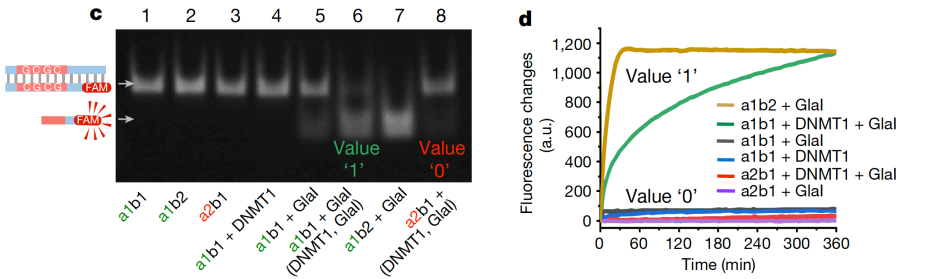

Figure 6: 分子信标与荧光测定法

3 总结

主要成就

与传统方法相比，epi - bit 存储框架具有显著优势。它采用并行 “印刷” 模式，利用预制的 DNA 可移动类型和甲基转移酶 DNMT1，如同活字印刷术一样，将 epi - bit 信息快速、高效地写入 DNA 模板，极大地提高了数据写入速度，降低了成本。此外，该框架成功存储了约 275,000 比特的数据，证明了其在大规模数据存储方面的可行性和有效性，为未来数据存储技术的发展提供了新方向。

进一步改进

- **优化序列设计**：进一步优化 DNA 序列设计，确保可移动类型与 DNA 模板之间的相互作用更加精准、稳定，减少非特异性结合和错误组装的可能性。
- **提高 DNM1 催化性能**：优化 DNM1 的性能，使其能够在不同条件下稳定地维持 DNA 甲基化状态。
- **增加存储密度**：研究并引入更多种类的 DNA 修饰方式，除了 5 - 甲基胞嘧啶（5mC）外，探索其他碱基修饰（如 N6 - 甲基腺嘌呤、5 - 羟甲基胞嘧啶）等。