基于合成DNA的数字信息储存和读取系统

李彦敏^{1†}, 钟云鹏^{1†}, 祁姗姗^{1†}, 盛付旭¹, 田净净¹, 朱沛煌¹, 杨平^{1*}, 蔡晓辉^{2*}

1. 苏州泓迅生物科技股份有限公司, 苏州 215123;

2. 上海生物信息技术研究中心, 上海 201203

† 同等贡献

* 联系人, E-mail: ping.yang@synbio-tech.com; xhcai@scbit.org

收稿日期: 2017-10-02; 接受日期: 2017-11-10; 网络版发表日期: 2018-01-16

江苏省基础研究计划(自然科学基金)青年基金项目(批准号: BK20170403)资助

在当今大数据时代, 数字信息的数量爆炸性的增长, 需要一种可靠的能够进行大规模数字数据存档的储存介质. 但是目前储存容量的增长并不能跟上数字信息量的增长, 而且常规储存介质如磁设备和光学设备的持久性有限. 自从1988年第一次实现使用脱氧核糖核酸(DNA)储存信息以来, DNA由于其高数据密度和长期稳定性(半衰期超过500年^[1]), 被认为是很有前途的数据储存介质. 已有多项研究提出了利用合成DNA进行信息储存的策略, 但是由于其采用三进制或二进制的数字编码方案, 信息提取的难度、费用和错误率相对较高, 限制了这些方法的应用^[2,3]. 因此, 本研究组开发了一种简单的新型编码方案, 利用DNA合成技术和二代测序技术(NGS)来存取数字信息.

首先将原始数据二进制流读出, 每二位转换成一四位二进制数字, 然后四进制数据的每一位就可以编码为一个核苷酸, 称之为“BitDNA”编码(即0为A, 1为T, 2为C, 3为G)(网络版附表1). 采用古文名作《论语》(网络版本1)进行了实验, 该文件大小43.1 KB, 其中的文本包含21505个汉字, 可以转换为长度176724 nt的核苷酸序列(网络版附表2). 由于现有单链DNA合成技术在合成序列长度上的限制, 将核苷酸序列分割为4017个字符串, 每个字符串长度为44个核苷酸. 与Goldman的工作类似, 将这些序列组织成为长度100 nt

的编码结构, 包括一个44 nt的核苷酸数据块, 两侧各两个8 nt的索引编码序列和两个20 nt的共同引物序列, 引物序列和索引编码序列可用于将来通过放大和测序方法进行数据提取(图1B). 每条链两侧连接的固定长度的引物序列可以保证只将感兴趣的信息通过聚合酶链式反应(PCR)提取出来, 而不是将整个DNA文库中的信息全部提取. 连接在引物内侧的成对索引编码在进行数据恢复时可指定数据块的位置, 因此能够实现数据的随机访问. 通过基于芯片的高通量DNA合成技术, 每个芯片可以合成12000条不同的寡核苷酸序列, 将设计的序列总计合成约微克量级的寡核苷酸文库.

为读取编码的DNA书本信息, 将整个DNA文库用PCR方法放大, 然后在Illumina HiSeq测序仪上使用二代测序技术测序(网络版附图1). 测序结果中只保留长度100 nt的具有完好的索引条码序列的读长. 合成序列中有一条序列没有完全恢复, 因为对应读出序列与原始序列相比有一个插入或删除的碱基. 与Goldman的报告相比, 其结果中有二个25 nt的片段没有读出, 由于其每个片段四重冗余设计, 表明至少有8条原始序列没有读出. 此处工作中, 所有数据块经计算机处理恢复, 结果在176724个碱基中有40 nt的错误(网络版附表2). 最终从测序结果中成功恢复了21504个字符的信

引用格式: 李彦敏, 钟云鹏, 祁姗姗, 等. 基于合成DNA的数字信息储存和读取系统. 中国科学: 生命科学, 2018, 48: 102–104, doi:10.1360/N052017-00249
英文版见: Zhong Y, Qi S, Sheng F, et al. A new digital information storing and reading system based on synthetic DNA. Sci China Life Sci, 2018, doi:10.1360/N052017-00249

图1 合成DNA是数字信息存储的可行介质

A: 数字信息储存系统示意图。二进制数字数据转化为四进制流后编码为DNA字母并分割编组为短片段, 然后根据设计结果合成DNA片段文库。文库经PCR放大及NGS技术测序读取序列, 经计算机解码恢复原始数据; B: 文本文件转化为DNA序列的编码方案: 内容为《论语》的文本文件中数字信息(黄色)首先转换为四进制数据(蓝色), 随后通过将每个四进制数字用对应核苷酸替代进一步转换为DNA序列(紫色)。在构建序列过程中加入了成对索引编码(橙色), 以及两侧引物序列(红色)。更多详细内容可参考附件

息(网络版本2), 其中信息错误率约0.12%, 但仅考虑测序错误时碱基错误率约0.001%。通过使用上述数据编码和表示方案, 得到的信息储存密度约123 GB g⁻¹。

本方法较近几项基于合成DNA的信息储存研究有几个优点。首先, 与二进制或三进制编码相比, 四进制“BitDNA”编码的信息压缩程度更高, 因此可以增加合成DNA中数字化信息的储存密度。其次, 以前的编码通常由于较长的数据信息会增加额外成本及较差的可靠性, 因此采用44个碱基长度的数据序列这种简单的编码格式。数据块的长度比Church和Goldman的研究中数据序列长度(分别为96和100 nt)短很多, 可以快速进行DNA文库的合成和测序。本研究组的设计方案中每KB的数据可以储存在约91条这样的序列中, 比Goldman的编码方案(每KB储存于302条序列)和XOR编码方案(每KB储存于404条序列)更为有效^[4]。此外, Goldman的编码方案中引入了重叠片段, 通过读取重

构序列中相邻区域序列来防止核苷酸丢失等错误, 增加了DNA文库的大小因而需要进行更多的处理。而此工作中DNA文库的设计没有采用重叠片段的方法, 会降低合成和测序的费用。在读取核苷酸序列时采用了成对索引编码序列以减少错误率。但是, 由于方案中数据随机访问功能以及较短的寡核苷酸序列长度, 数据储存密度与理论值仍有较大差距。另外值得一提的是, 这是第一次将中文信息转化保存到核苷酸中, 并从合成DNA中读取这些信息。

综上, 本研究组设计了一种利用DNA合成和二代测序技术的DNA信息储存系统。这一信息处理流程与本研究组前期的专利^[5]或Church和Goldman的基于二或三进制的系统相比有更为简单的编码方案。DNA合成和测序技术仍以指数速率不断发展, 基于合成DNA的信息储存和读取方法很有希望成为高数据密度和持久的信息存储方案。

参考文献

- 1 Allentoft M E, Collins M, Harker D, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. [Proc R Soc B-Biol Sci](#), 2012, 279: 4724–4733
- 2 Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA. *Science*, 2012, 337: 1628
- 3 Goldman N, Bertone P, Chen S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. [Nature](#), 2013, 494: 77–80
- 4 Bornholt J, Lopez R, Strauss K, et al. A DNA-Based Archival Storage System. Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, 2016, (ASPLOS'16): 637–649
- 5 杨平, 钟云鹏, 刁文一, 等. 人工合成DNA存储介质的信息存储读取方法. 中华人民共和国专利, ZL201510138778.3, 2015-03-27