团体标准

T/CIE XXX-XXXX

|  |  |
| --- | --- |
|  |  |

生物数字融合 DNA存储系统参考架构

|  |
| --- |
| Bio-digital convergence —Reference architecture of DNA storage system  征求意见稿 |
|  |

202X-XX-XX发布

中国电子学会 发布

202X-XX-XX实施

目  次

前  言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由中国电子学会提出并归口。

本文件起草单位：。

本文件主要起草人：

生物数字融合 DNA存储系统参考架构

1. 范围

本文件规定了DNA存储系统的参考架构、主要功能和工作流程。

本文件适用于DNA存储系统的设计和研发。

1. 规范性引用文件

本文件没有规范性引用文件。

1. 术语和定义

DNA存储技术 DNA storage technology

用人工合成的脱氧核糖核酸(DNA)作为存储介质的数据存储技术。

DNA存储系统 DNA storage system

以DNA存储技术为核心，人工合成可以编码数字信息的DNA序列，并能进行稳定存储、读取和恢复原始信息的系统。

1. 参考架构

DNA存储系统主要包括写入、存储和读取三大组成部分，系统参考架构见图1。

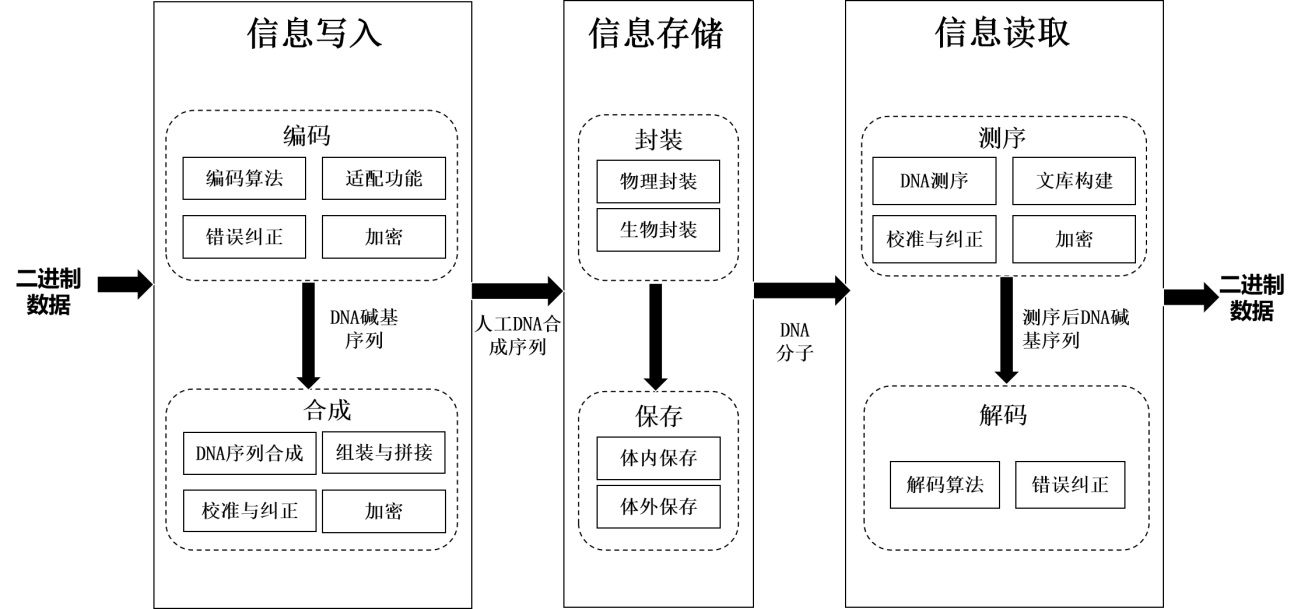


图1 DNA存储系统参考架构

1. 主要功能
   1. 信息写入

将二进制数据写入DNA存储媒体的过程被称为信息写入，信息写入主要包括编码和合成两大模块。

编码是将数字文件中的二进制数据提取出来，并按一定规则转换成为DNA碱基序列。

合成是通过化学原理或生物酶法原理，利用柱式合成、点阵芯片、半导体芯片等方式根据编码步骤获得的DNA序列进行从头人工合成。

* + 1. 转码映射

转码算法主要分为两类：

1. 考虑DNA分子序列的受限限制进行映射的算法，包括Church编码算法、Goldman编码算法、Grass编码算法、Blawat编码算法等；
2. 将两比特转化为一个碱基，然后增加筛选过滤步骤的编码算法，包括DNA Fountain喷泉码和Yin-Yang 双编码算法等。

不同类型算法性能不同、生成序列特性不同、隐私性及稳健性不同。待存储数据的编码算法选择可根据以下特征进行选择：

1. 数据特征，包括文件大小、字节频率等；
2. 算法性能特征，包括信息密度、转码速度等；
3. 序列特征，包括GC含量、单碱基重复长度等；
4. 隐私性特征，包括稳健性、破译难度等。
   * 1. 纠错编码

采用合适的技术或算法添加额外的校验数据，对原始数据进行数学变换使之能够应对DNA存储过程中出现的错误。

* + 1. 生化适配

进行序列筛选以适应DNA分子的生化特性，如：无均聚物、平衡的GC含量、无复杂二级结构、无生物活性序列等。

* + 1. 数据加密

为保障DNA存储中的加密与数据安全，可根据存储数据的隐私性要求，对于编码环节进行加密，实现DNA编码高效、隐私。

* 1. 合成
     1. DNA序列合成

DNA合成环节的主要功能是完成DNA序列合成。DNA的合成按原理一般分为化学法合成和生物法合成。

化学法合成是主流成熟方法，包括一代合成（柱式）和二代合成（芯片）。生物法合成即三代合成（生物酶）。

* + 1. 序列拼接

受限于高通量合成DNA序列的长度限制，为适应高通量长片段测序技术的要求，将合成的短DNA片段拼接为长序列的过程。

* 1. 信息贮存
     1. 封装

封装分为物理封装和生物封装。物理封装即通过低温密封、矿化、无机物密封、固体胶囊等形式进行封装，生物封装即将DNA分子通过不同形式存储在活细胞内。

* + 1. 保存

保存分为体外保存和体内保存。当体外保存DNA分子时，可采用干粉、溶液、纳米颗粒等形式保存。当体内保存DNA分子时，可采用活细胞内的存储介质，如染色体、质粒等。

* 1. 信息读取
     1. 概述

将保存在不同存储媒体中的DNA序列进行测序获得DNA片段的碱基排列顺序，并将碱基序列转换为二进制数据从而得到存储的文本文档、图片和声音文件等数据的过程称为信息读取。信息读取过程主要包括检索、测序和解码。

* + 1. 检索

通过探针、PCR等方式抽取目标DNA以实现随机索引、重复获取以及快速预览功能。

* + 1. 测序

测序是获得利用多重PCR分子等方式获取的编码DNA分子片段的碱基序列，即该片段中腺嘌呤A、胸腺嘧啶T、胞嘧啶C、鸟嘌呤G的排列顺序。

DNA测序技术主要包括Sanger测序法、高通量测序技术和单分子测序技术。

高通量测序技术主要基于可逆末端终止法、连接测序法、焦磷酸测序法和联合探针锚定聚合测序法。

单分子测序技术主要基于单分子荧光测序法、单分子纳米孔测序法和单分子晶体管测序法。

目前技术最为成熟的是高通量测序技术，已经能够实现高通量、高准确性和低重复序列率的性能。

* + 1. 解码

解码算法的类型与选择与编码算法相同，见5.1.1。

解码环节的错误纠正是确保信息读取正确率的最后一环，为了纠正DNA存储过程中出现的插入、删除、替换等多种类型的错误，开发与编码算法适配的纠错方法。

1. DNA存储系统工作流程

DNA存储系统工作流程一般可分为4个主要步骤，包括编码、合成、贮存、测序和解码：

1. 编码：将数字文件中的数据提取出来，并按一定规则转换成为DNA碱基序列（包括但不限于自然碱基）；
2. 合成：通过化学原理或生物酶法原理，利用柱式合成、点阵芯片、半导体芯片等方式根据编码步骤获得的DNA序列进行人工合成；
3. 贮存：通过物理封装或生物封装方式，将携带信息的DNA分子保存于体外或体内；
4. 测序：通过利用如多重PCR、生物素-亲和素序列特异性磁珠捕获等方式，完成全部或部分编码DNA分子的获取，将获取的DNA分子通过Sanger测序、高通量测序、单分子测序等方式进行DNA分子序列的测定；
5. 解码：对测序获得的DNA序列进行信息分析，根据编码规则对DNA碱基序列进行解码并得到原始数字文件的二进制数据，最终实现信息恢复。

DNA存储系统工作流程见图2。

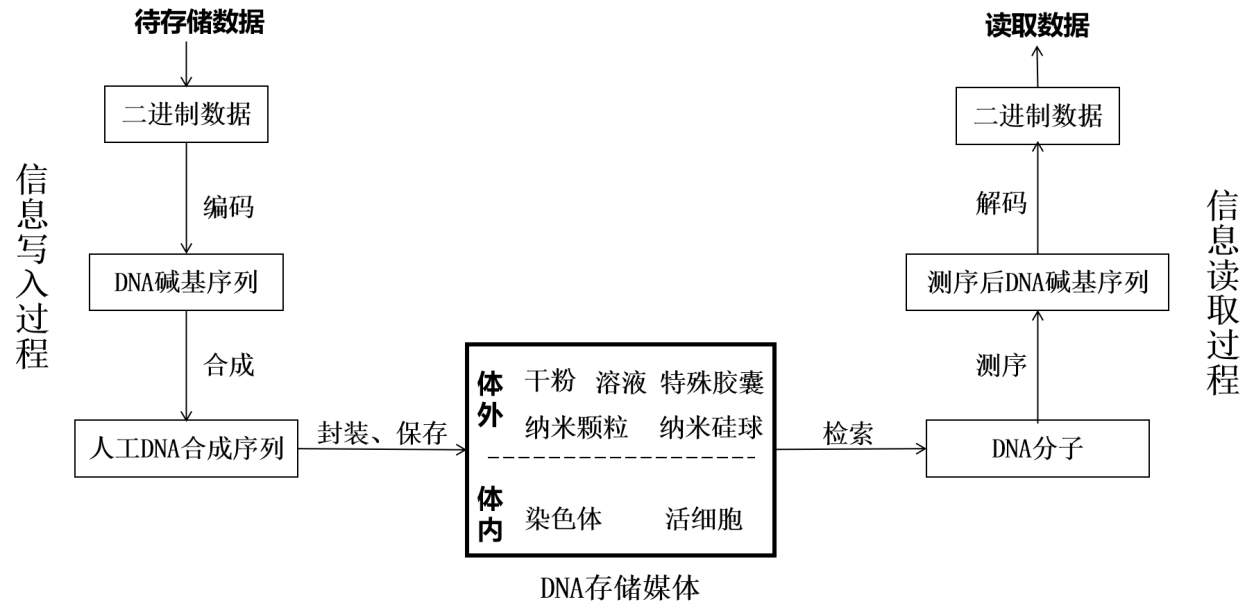


图2 DNA存储系统工作流程

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_