

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 12 月 29 日 (29.12.2022)



(10) 国际公布号
WO 2022/266802 A1

(51) 国际专利分类号:
G16B 30/00 (2019.01) *G16B 50/30* (2019.01)

广东省深圳市南山区西丽大学城学苑大道
1068号, Guangdong 518055 (CN)。

(21) 国际申请号: PCT/CN2021/101247

(74) 代理人: 深圳中一联合知识产权代理有限公司 (SHENZHEN ZHONGYI UNION INTELLECTUAL PROPERTY AGENCY CO.,LTD.); 中国广东省深圳市福田区园岭街道深南中路1014号报春大厦9楼(5号信箱), Guangdong 518028 (CN)。

(22) 国际申请日: 2021 年 6 月 21 日 (21.06.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

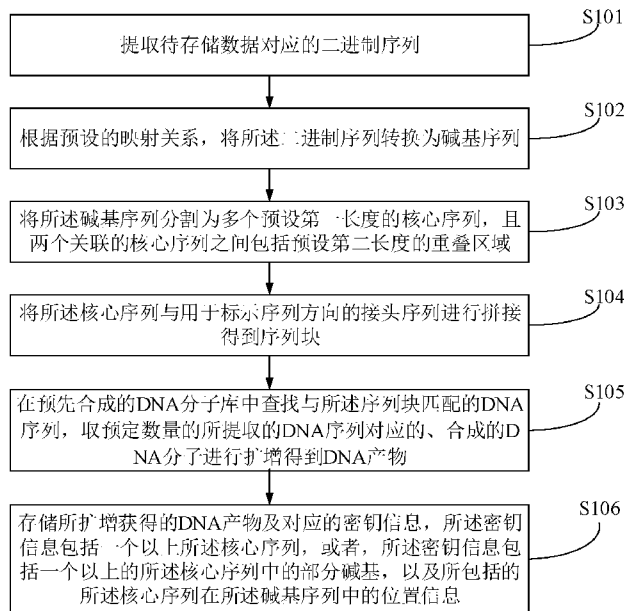
(71) 申请人: 深圳先进技术研究院 (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY) [CN/CN]; 中国广东省深圳市南山区西丽大学城学苑大道1068号, Guangdong 518055 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,

(72) 发明人: 戴俊彪(DAI, Junbiao); 中国广东省深圳市南山区西丽大学城学苑大道1068号, Guangdong 518055 (CN)。 黄小罗(HUANG, Xiaoluo); 中国

(54) Title: METHOD AND APPARATUS FOR PERFORMING DATA STORAGE BY USING DNA, AND STORAGE DEVICE

(54) 发明名称: 利用DNA进行数据存储的方法、装置及存储设备



- S101 Extract a binary sequence corresponding to data to be stored
S102 Convert the binary sequence into a base sequence according to a preset mapping relationship
S103 Segment the base sequence into a plurality of core sequences of a preset first length, wherein an overlapping region of a preset second length is comprised between two associated core sequences
S104 Splice the core sequences and a linker sequence which is used for marking a sequence direction, so as to obtain a sequence block
S105 Search, in a pre-synthesized DNA molecule library, for a DNA sequence matching the sequence block, and take a pre-determined number of synthesized DNA molecules which correspond to an extracted DNA sequence, and amplify same to obtain a DNA product
S106 Store the DNA product obtained by means of amplification, and corresponding key information, wherein the key information comprises one or more of the core sequences, or the key information comprises some bases in the one or more of the core sequences, and position information of the comprised core sequences in the base sequence

图1

(57) Abstract: The present application belongs to the field of data storage. Provided is a method for performing data storage by using DNA. The method comprises: extracting a binary sequence corresponding to data to be stored; converting the binary sequence into a base sequence according to a preset mapping relationship; segmenting the base sequence into a plurality of core sequences of a preset first length; splicing the core sequences and a linker sequence which is used for marking a sequence direction, so as to obtain a sequence block; searching, in a pre-synthesized DNA molecule library, for a DNA sequence matching the sequence block, and taking a pre-determined number of synthesized DNA molecules which correspond to an extracted DNA sequence, and amplifying same to



PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

obtain a DNA product; and storing the DNA product obtained by means of amplification, and corresponding key information. By means of the method in the present application, the cost of performing data storage by using DNA can be reduced, thereby facilitating an improvement in the wide application of a storage device based on a DNA medium.

(57) 摘要: 本申请属于数据存储领域, 提供了一种利用DNA进行数据存储的方法, 包括: 提取待存储数据对应的二进制序列; 根据预设的映射关系, 将所述二进制序列转换为碱基序列; 将所述碱基序列分割为多个预设第一长度的核心序列; 将所述核心序列与用于标示序列方向的接头序列进行拼接得到序列块; 在预先合成的DNA分子库中查找与所述序列块匹配的DNA序列, 取预定数量的所提提取的DNA序列对应的、合成的DNA分子进行扩增得到DNA产物; 存储所扩增获得的DNA产物及对应的密钥信息。本申请所述的方法能够降低利用DNA进行数据存储的成本, 有利于提升基于DNA介质的存储设备的广泛应用。

利用DNA进行数据存储的方法、装置及存储设备

技术领域

[0001] 本申请属于数据存储领域，尤其涉及利用DNA进行数据存储的方法、装置及存储设备。

背景技术

[0002] 随着互联网技术、大数据技术以及人工智能技术等科学技术的发展，全球数据呈现指数级增长。传统的硬盘、磁带或光盘等存储设备，由于维护成本高、空间占用大以及存储寿命短等原因，无法满足不断增长的海量数据存储的需求。

[0003] 使用DNA进行数据存储时，具有存储密度高、存储时间长及维护成本低的特点。并且，DNA作为生命遗传信息物质，可以插入到动植物微生物细胞中，通过生命体的复制，实现代代相传的永久保存。

[0004] 目前的基于DNA进行数据存储的方法中，通常需要根据不同的存储数据合成对应的不同的DNA碱基序列。存储的计算机比特位数和合成碱基的对应形式中，通常是1个碱基对应1到2位数据位。即存储的计算机二进制数据量，至少需要二进制数据量的二分之一的合成碱基量进行存储，并且一次体外合成的DNA分子只能用于一次特定数据的存储，合成成本较高，不利于DNA介质的存储技术的广泛应用。

发明概述

技术问题

[0005] 有鉴于此，本申请实施例提供了一种利用DNA进行数据存储的方法、装置及设备，以解决现有技术中利用DNA进行数据存储时，合成成本较高，不利于DNA介质的存储技术的广泛应用的问题。

问题的解决方案

技术解决方案

[0006] 本申请实施例的第一方面提供了一种利用DNA进行数据存储的方法，所述方法包括：

- [0007] 提取待存储数据对应的二进制序列；
- [0008] 根据预设的映射关系，将所述二进制序列转换为碱基序列；
- [0009] 将所述碱基序列分割为多个预设第一长度的核心序列，且两个关联的核心序列之间包括预设第二长度的重叠区域；
- [0010] 将所述核心序列与用于标示序列方向的接头序列进行拼接得到序列块；
- [0011] 在预先合成的DNA分子库中查找与所述序列块匹配的DNA序列，取预定数量的所述提取的DNA序列对应的、合成的DNA分子进行扩增得到DNA产物；
- [0012] 存储所扩增获得的DNA产物及对应的密钥信息，所述密钥信息包括一个以上所述核心序列，或者，所述密钥信息包括一个以上的所述核心序列中的部分碱基，以及所包括的所述核心序列在所述碱基序列中的位置信息。
- [0013] 结合第一方面，在第一方面的第一种可能实现方式中，所述接头序列包括左接头序列和右接头序列中的一项或者两项，当所述接头序列包括左接头序列和右接头序列时，所述左接头序列与所述右接头序列不同。
- [0014] 结合第一方面，在第一方面的第二种可能实现方式中，将所述二进制序列转换为碱基序列，包括：
- [0015] 对所述待存储数据对应的文件进行拆分，根据文件拆分的结果对所述碱基序列进行拆分，并为拆分的碱基序列分配对应的索引序列。
- [0016] 结合第一方面，在第一方面的第三种可能实现方式中，两个关联的核心序列之间包括预设第二长度的重叠区域，包括：
- [0017] 相邻两个核心序列之间包括预设第二长度的重叠区域；
- [0018] 或者，相邻的奇数位的核心序列之间包括预设第二长度的重叠区域，以及相邻的偶数位的核心序列之间包括预设第二长度的重叠区域；
- [0019] 或者，第 $M+i$ 位和第 $M+N+i$ 位的核心序列之间包括预设第二长度的重叠区域，其中， M 和 N 为预定的整数， i 为大于或等于0的整数变量。
- [0020] 结合第一方面，在第一方面的第四种可能实现方式中，将所述碱基序列分割为多个预设第一长度的核心序列，包括：
- [0021] 当划分得到的最后一个核心序列的长度小于第一长度时，通过预设的重复碱基补齐所述最后一个核心序列。

- [0022] 结合第一方面，在第一方面的第五种可能实现方式中，存储所扩增获得的DNA产物及对应的密钥信息，包括：
- [0023] 根据预设的碱基单元与位置信息的映射关系，将所述密钥信息中的位置信息转换为碱基单元，根据预设的组合方式，将密钥信息中的核心序列，以及所述碱基单元，以DNA形式进行存储；
- [0024] 或者，所述密钥信息中的核心序列，以及所包括的核心序列的位置信息通过计算机可读存储介质存储；
- [0025] 或者，所述的密钥信息通过DNA形式和计算机可读存储介质形式进行混合存储。
- [0026] 结合第一方面、第一方面的第一种可能实现方式、第一方面的第二种可能实现方式、第一方面的第三种可能实现方式、第一方面的第四种可能实现方式或第一方面的第五种可能实现方式，在第一方面的第六种可能实现方式中，所述密钥信息包括起始位置的核心序列和/或末尾位置的核心序列。
- [0027] 结合第一方面，在第一方面的第七种可能实现方式中，所述密钥信息还包括所述接头序列、拆分所述碱基序列得到的子碱基序列所对应的索引序列中的一项或者两项。
- [0028] 本申请实施例的第二方面提供了一种利用DNA进行数据存储的装置，所述装置包括：
- [0029] 二进制序列提取单元，用于提取待存储数据对应的二进制序列；
- [0030] 第一序列转换单元，用于根据预设的映射关系，将所述二进制序列转换为碱基序列；
- [0031] 碱基分割单元，用于将所述碱基序列分割为多个预设第一长度的核心序列，且两个关联的核心序列之间包括预设第二长度的重叠区域；
- [0032] 序列拼接单元，用于将所述核心序列与用于标示序列方向的接头序列进行拼接得到序列块；
- [0033] DNA分子提取单元，用于在预先合成的DNA分子库中查找与所述序列块匹配的DNA序列，取预定数量的所提取的DNA序列对应的、合成的DNA分子进行扩增得到DNA产物；

- [0034] DNA存储单元，用于存储所扩增获得的DNA产物及对应的密钥信息，所述密钥信息包括一个以上所述核心序列，或者，所述密钥信息包括一个以上的所述核心序列中的部分碱基，以及所包括的所述核心序列在所述碱基序列中的位置信息。
- [0035] 第三方面，本申请实施例提供了一种DNA介质存储数据的解码的方法，所述方法包括：
- [0036] 获取待解码的DNA序列及其密钥信息；
- [0037] 根据预设的接头序列，提取所述DNA序列中包括的核心序列；
- [0038] 根据所述密钥信息，结合所述核心序列之间的重叠区域，将所述核心序列组合生成碱基序列；
- [0039] 根据预设的映射关系，将所述碱基序列转换为二进制序列；
- [0040] 根据所转换的二进制数据生成数据文件。
- [0041] 结合第三方面，在第三方面的第一种可能实现方式中，根据预设的接头序列，提取所述DNA序列中包括的核心序列，包括：
- [0042] 根据预设的接头序列，对所述DNA序列进行切割得到核心序列；
- [0043] 根据所述接头序列的位置，确定所得到的核心序列的方向。
- [0044] 结合第三方面，在第三方面的第二种可能实现方式中，根据所述密钥信息，结合所述核心序列之间的重叠区域，将所述核心序列组合生成碱基序列，包括：
- [0045] 根据所述密钥信息确定一个以上的核心序列在碱基序列中的位置；
- [0046] 根据所述核心序列之间的重叠区域，确定所述核心序列之间的相对位置关系，根据确定了相对位置关系的核心序列生成所述碱基序列。
- [0047] 结合第三方面、第三方面的第一种可能实现方式或第三方面的第二种可能实现方式，在第三方面的第三种可能实现方式中，根据预设的接头序列，提取所述DNA序列中包括的核心序列，包括：
- [0048] 当提取的核心序列的末尾包括预设的重复碱基时，去除该核心序列的末尾所包括的所述重复碱基。
- [0049] 本申请实施例的第四方面提供了一种DNA介质存储数据的解码的装置，所述装置包括：

- [0050] DNA序列获取单元，用于获取待解码的DNA序列及其密钥信息；
- [0051] 核心序列提取单元，用于根据预设的接头序列，提取所述DNA序列中包括的核心序列；
- [0052] 序列组合单元，用于根据所述密钥信息，结合所述核心序列之间的重叠区域，将所述核心序列组合生成碱基序列；
- [0053] 第二序列转换单元，用于根据预设的映射关系，将所述碱基序列转换为二进制序列；
- [0054] 数据文件生成单元，用于根据所转换的二进制数据生成数据文件。
- [0055] 本申请实施例的第五方面提供了一种DNA分子库的生成方法，所述方法包括：
- [0056] 根据碱基组合生成核心序列，所述核心序列包括用于存储数据的碱基片段；
- [0057] 在所述核心序列上拼接预设的接头序列，得到所述核心序列对应的DNA序列，所述接头序列用于标识所述核心序列的方向；
- [0058] 根据所述DNA序列合成对应的DNA分子，根据所合成的DNA分子得到DNA分子库。
- [0059] 本申请实施例的第六方面提供了一种存储设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如第一方面任一项所述利用DNA进行数据存储的方法的步骤，或者执行所述计算机程序时实现如第二方面任一项所述DNA介质存储数据的解码的方法，或者执行所述计算机程序时实现如第五方面任一项所述的DNA分子库的生成方法。

发明的有益效果

有益效果

- [0060] 本申请实施例与现有技术相比存在的有益效果是：本申请通过将待存储数据对应的二进制序列转换为碱基序列，根据预设的第一长度对碱基序列进行分割，且分割得到的相邻的核心序列之间包括第二长度的重叠区域，结合密钥信息中包括的核心序列在碱基序列中的位置信息，从而便于解码时根据该重叠区域对核心序列进行组合；将核心序列与接头序列进行拼接得到序列块，从而便于解码时确定核心序列的方向；在单次的数据存储时，可以从DNA分子库中获取预定数量的少量合成的DNA分子，大大的减少一次数据存储所使用合成DNA分子数，

可以保证一次体外合成的DNA分子的多次调用，有效的降低了数据存储的DNA合成成本；同时，由于预先合成的DNA分子库可以反复调用，避免了针对于不同的二进制数据，需要合成对应数据至少二分之一碱基数量的情况，节省了总体的碱基合成数量，进一步降低了存储数据的DNA合成成本，有利于提升DNA介质的存储设备的广泛应用。

对附图的简要说明

附图说明

- [0061] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例或示范性技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。
- [0062] 图1是本申请实施例提供的利用DNA进行数据存储方法的实现流程示意图；
- [0063] 图2是本申请实施例提供的一种通过碱基序列生成核心序列的示意图；
- [0064] 图3是本申请实施例提供的又一碱基序列生成核心序列示意图；
- [0065] 图4是本申请实施例提供的一种通过分割得到的核心序列的示意图；
- [0066] 图5是本申请实施例提供的一种拼接后的序列块的示意图；
- [0067] 图6是本申请实施例提供的一种组合规则示意图；
- [0068] 图7是本申请实施例提供的一种DNA介质存储数据的解码的方法的实现流程示意图；
- [0069] 图8是本申请实施例提供的一种利用DNA进行数据存储的装置的示意图；
- [0070] 图9是本申请实施例提供的一种DNA介质存储数据的解码的装置的示意图；
- [0071] 图10是本申请实施例提供的存储设备的示意图。

发明实施例

本发明的实施方式

- [0072] 以下描述中，为了说明而不是为了限定，提出了诸如特定系统结构、技术之类的具体细节，以便透彻理解本申请实施例。然而，本领域的技术人员应当清楚，在没有这些具体细节的其它实施例中也可以实现本申请。在其它情况中，省略对众所周知的系统、装置、电路以及方法的详细说明，以免不必要的细节妨

碍本申请的描述。

[0073] 为了说明本申请所述的技术方案，下面通过具体实施例来进行说明。

[0074] DNA的体外合成技术，通常是任意设定的序列，在不依赖模板的情况下，以修饰的A/T/C/G碱基化学分子单体原料为基础，在体外根据化学合成或者酶合成的方法，以每轮化学或酶反应逐个或逐几个碱基添加到上一个/几个碱基上的方式，通过多轮化学或酶反应，合成与设定序列一致的A/T/C/G组成的大分子DNA聚合物。其中由于每轮化学或酶反应，都需要花费支撑该轮化学或酶反应的试剂，耗材，人力及机械损耗成本，待合成DNA序列的碱基数量决定了大部分的大分子DNA聚合物的合成成本。市售的常规的DNA合成服务，通常也是以碱基单价乘以碱基数量，来给客户报价。比如50个碱基的常规单链DNA合成，商品公司可以报价0.3-0.6 元每个碱基，这样50个碱基的售价在15元至30元之间。在目前的DNA数据存储方式中，需要将至少二分之一的二进制数据量的碱基量进行合成，合成价格较高。而且，在单次的DNA合成操作中，根据目前的技术特点，由于每轮化学或酶反应需要固定的反应体积，一次合成至少预定数量的DNA分子。比如上述举例中，50个碱基的常规单链DNA合成，市售公司的合成量通常至少为 0.5 nmol (3×10^{14} 个分子)。然而，在DNA数据存储应用中，可以理解的是，仅需要预定数量的DNA分子的极少一部分（比如1个分子， 10^2 个分子， 10^3 个分子， 10^5 个分子等等）就能代表要存储的数据信息，这样用一次合成的DNA直接存储数据信息的方式，造成合成DNA分子原料的极大浪费。

[0075] 本申请提出了一种通过预先合成的DNA序列进行数据存储的方法。通过将存储数据的碱基序列划分为具有重叠区域的核心序列的片段，利用提前合成的可以多次调取的通用的DNA分子库对数据进行存储，通用性好，并且可以根据需要从预先合成的DNA分子库中提取与待存储数据对应的DNA序列，大大的减少了用于存储不同数据信息的DNA产物所需要合成的总碱基数量以及一次使用合成DNA分子数，降低数据存储成本，从而有利于DNA数据存储的广泛应用。下面结合附图具体进行说明。

[0076] 图1为本申请实施例提供的一种利用DNA进行数据存储的实现流程示意图，详述

如下：

[0077] 在S101中，提取待存储数据对应的二进制序列。

[0078] 其中，所述待存储数据可以包括图片、文本、程序、音频、视频等可以在计算机存在的数据信息中的一项或者多项。

[0079] 在获取待存储数据对应的二进制序列时，可以获取该待存储数据对应的编码信息，将对应的编码信息转换为二进制的编码信息，从而得到对应的二进制序列。比如，可以将文本信息中的文字转换为对应的ASCII（英文全称为American Standard Code for Information Interchange，中文全称为美国标准信息交换码）编码，UNICODE（英文全称为Universal Character Set，中文全称为通用字符集）编码，然后将编码信息转换为二进制序列。

[0080] 例如，提取文本“将美丽的冬季雕塑成一件艺术品”对应的二进制序列为“11100101101100001000011011100111101111101000111011100100101110001011110111100111100110101000010011100101100001101010110011100101101011011010001111010011001101110010101111001011010000110010001111001101000100010010000111001001011100010000000111001001011101110110110111010001000100110111010111001101001110010101111111001011001001110000001”。

[0081] 在S102中，根据预设的映射关系，将所述二进制序列转换为碱基序列。

[0082] 由于DNA中的碱基包括A、T、C、G四种类型的碱基，因此，该预设的映射关系，可以为二进制与四进制的映射关系。在一种可能的实现方式中，该映射关系可以如表1所示：

[] [表1]

二进制数值	00	01	11	10
碱基	A	T	C	G

[0083] 表1

[0084] 上表的映射关系为任意定义的一种，可以根据实际的使用习惯或要求，对二进制数值与碱基的映射关系进行定义。

[0085] 根据上表中的映射关系，可以将S101中的二进制序列转换为碱基序列为：“CG

TTGCAAGATGCGTCGCCGGACGCGTAGCGAGCCTCGTCGTGGGATACGTTGATGGGCACGTTGGCTGGA
CCGGTGTGCGTTTCGTTGGATGTATCGTGGAGAGTAACGTAGCGAGAAACGTAGCGCGCTGCGGAGAGT
GCGGCGTGGTCAGGCCCGTTGTACGAAT”。

[0086] 在可能的实现方式中，如果待存储数据较大时，可以将待存储数据的文件拆分为多个子文件，并可以通过索引序列来记录各个子文件所对应的子碱基序列的先后关系。在将待存储数据或文件拆分后，可以通过索引序列来记录所拆分得到的子文件。该索引序列也可以添加在用于扩增本方法所述核心序列块左接头或者右接头对应的引物上，通过PCR等聚合酶扩增的方式，进一步加在该子文件通过本方法存储的每一条核心序列的扩增用的DNA分子上。通过对同一子文件对应的核心序列块加上统一的索引标识，有利于将这些序列聚类到一起进行拼接，从而有利于子文件数据的解读。通过拆分和扩增的方式，便于对大的数据文件进行分段进行转换，从而有利于提升数据存储的准确性。

[0087] 其中，索引序列是指通过碱基构成的序列，该索引序列可用来表示子文件的位置信息。比如，索引序列可以定义为AAAA=1，AAAG=2，AAAT=3，AAAC=4，AATA=5，AACA=6等。

[0088] 在S103中，将所述碱基序列分割为多个预设第一长度的核心序列，且两个关联的核心序列之间包括预设第二长度的重叠区域。

[0089] 为了便于直接取用预先合成的DNA序列，需要将转换的碱基序列进行分割，得到预定的第一长度的核心序列。其中，第一长度可以4个碱基长度、5个碱基长度、6个碱基长度、7个碱基长度或8个碱基长度等。

[0090] 为了能够将分割后的核心序列正确的组合，对于分割得到的两个相邻的核心序列，包括第二预设长度的重叠区域的碱基。即两个核心序列在该重叠区域中的碱基相同。根据该重叠区域的碱基，可以在组合时查找存在关联的两个核心序列。

[0091] 其中，预先设定的关联关系所确定的存在关联的两个核心序列，可以为核心序列在碱基序列中直接相邻的两个核心序列，也可以为奇数位相邻的核心序列，或者偶数位相邻的核心序列。

[0092] 存在关联的两个核心序列为直接相邻的两个核心序列时，如图2所示的通过碱

基序列生成核心序列的示意图中，分别示出了重叠区域包括3个碱基交错重叠（a）、4个碱基交错重叠（b）、5个碱基交错重叠（c）、6个碱基交错重叠（d）和7个碱基交错重叠（e）的情形。其中，重叠区域的第二长度为核心序列的第一长度的一半。不局限于此，第一长度和第二长度还可以为其它比例关系。当然，重叠区域的碱基数量不局限于图2所示，还可以包括其它的重叠区域的碱基数量。

[0093] 当碱基序列分割得到的末尾的核心序列的长度小于预设的第一长度时，可以通过预设的碱基类型或预设的重复碱基进行补齐。比如，图2中的5碱基交错重叠（c）的分割过程中，在末尾的核心序列补齐碱基A。在7碱基交错重叠（e）的分割过程中，在末尾的核心序列补充碱基AAAA，从而使得补齐后的核心序列具有与其它核心序列的相同长度，便于对核心序列进行拼接和存储操作。

[0094] 存在关联的核心序列可以为奇数位相邻，或者偶数位相邻的核心序列。如图3示出的又一碱基序列生成核心序列的示意图中，分别示出了3个碱基奇偶交错重叠（f）、4个碱基奇偶交错重叠（g）以及3个连续的4碱基单元交错重叠（h）的示意图。其中，3个连续的4碱基单元交错重叠的示意图中，重叠区域为相邻的两个核心序列中的连续的8个碱基，且核心序列的长度为12个碱基。

[0095] 在可能的实现方式中，还可以设置第M+i位和第M+N+i位的核心序列包括预设第二长度的重叠区域，其中，M和N为预定的整数，i为大于或等于0的整数变量。通过更改参数M、N，可以实现不同形式的加密编码。

[0096] 将S102中的碱基序列按照图2中（b）方式进行核心序列分割时，可以得到图4所示的核心序列的表格示意图。其中，任意两个相邻的核心序列包括4个重叠区域的碱基。在对DNA序列进行解码时，可以根据DNA序列中的核心序列中的重叠区域的碱基，对核心序列进行两两拼接的组合操作。

[0097] 在S104中，将所述核心序列与用于标示序列方向的接头序列进行拼接得到序列块。

[0098] 本申请实施例中的接头序列，可以包括前接头序列和后接头序列，也可以为前接头序列或后接头序列中的任意一种。通过拼接接头序列得到的序列块，可以用于标示核心序列的前后方向。比如，通过前接头序列表示核心序列的前方或

左侧方向，通过后接头序列表示核心序列的后方或右侧方向。从而使得DNA序列在进行解码时，可以根据接头序列确定核心序列的方向，便于对核心序列进行正确的组合。

[0099] 图4为通过分割得到的核心序列示意图，可以根据预先设定的前接头序列“CGC CAGGGTTTTCCCAGTCACGAC”，以及预先设定的后接头序列“TCCTGTGTGAAATTGTTA TCCGCT”，对26个核心序列分别进行拼接，得到图5所示的拼接后的序列块的示意图。

[0100] 在S105中，在预先合成的DNA分子库中查找与所述序列块匹配的DNA序列，取预定数量的所提取的DNA序列对应的、合成的DNA分子进行扩增得到DNA产物。

[0101] 在本申请中，预先设定有合成的DNA分子库，在所述DNA分子库中存储有合成的DNA分子与DNA序列的对应关系。根据S104中通过拼接得到的序列块，可以在预先设定的DNA分子库中查找所述序列块对应的DNA序列，调取在DNA分子库中所查找获得的、合成的DNA分子中的少量分子（比如，1个分子， 10^2 个分子， 10^3 个分子， 10^5 个分子等），并对其进行扩增，可以保证一次体外合成DNA分子库中的DNA分子的多次调用，从而能够大大的减少一次数据存储所使用DNA分子数，从而有效的降低用于数据存储的DNA合成成本。

[0102] 其中，预先合成的DNA分子库中，可以包括不同的碱基长度的核心序列所构成的DNA分子库。比如，核心序列的长度可以包括2个碱基长度所构成的任意序列、3个碱基长度所构成的任意序列、4个碱基长度所构成的任意序列、5个碱基长度所构成的任意序列、6个碱基长度所构成的任意序列、7个碱基长度所构成的任意序列、8个碱基长度所构成的任意序列等等。其中，2个碱基长度所构成的任意序列的数量为 $4*4=16$ 个，3个碱基长度所构成的任意序列的数量为 $4*4*4=64$ 个，其它依此类推计算。

[0103] 根据预先设定的接头序列，将所设定的核心序列与所述接头序列进行拼接。比如，将左接头序列和右接头序列分别拼接在核心序列的左右侧。通过大量合成带有接头序列的DNA分子，得到DNA分子库。当需要使用DNA分子库时，只需要根据预先合成的DNA分子库中取一定量的、与序列块对应的DNA分子，根据所选择的DNA分子库中的DNA分子进行数据存储，并且所取的DNA分子库中的剩余的DNA

分子能够继续使用，有利于进一步降低DNA介质的数据存储成本。

[0104] 其中，对所述DNA分子进行扩增时，所采用的引物扩增方式可以包括等温扩增、PCR（中方全称为聚合酶链式反应）扩增等扩增方式。

[0105] 在S106中，存储所扩增获得的DNA产物序列及对应的密钥信息，所述密钥信息包括一个以上所述核心序列，或者，所述密钥信息包括一个以上的所述核心序列中的部分碱基，以及所包括的所述核心序列在所述碱基序列中的位置信息。

[0106] 其中，该密钥信息中包括一个以上的核心序列，或者一个以上的所述核心序列中的部分碱基，以及该核心序列在碱基序列中对应的位置信息，从而可以快速的确定该核心序列的准确位置。

[0107] 在可能的实现方式中，密钥信息包括起始位置的核心序列以及末尾位置的核心序列。从而可以根据起始位置的核心序列以及末尾位置的核心序列，根据重叠区域的碱基对中间位置的核心序列进行拼接。在可能的实现方式中，还可以包括若干中间位置的核心序列。当重叠区域的碱基数量越少时，可以增加核心序列的位置信息的数量。

[0108] 将密钥信息存储时，可以将其中的位置信息在碱基序列中的序号转换为碱基单元，将位置信息所转换的碱基单元，以及核心序列按照预定的组合方式，以DNA形式进行存储。比如，可以按照图6所示的组合规则进行组合，其中，位置序号1对应碱基单元AAAA，位置序号18对应碱基单元ACAA，位置序列100对应碱基单元CTGA。位置序号与碱基单元的对应关系可以根据预先设定的映射关系来确定。

[0109] 将密钥信息存储时，所述密钥信息中的位置信息及碱基单元序列也可通过计算机可读存储介质存储。或者，所述的密钥信息通过DNA形式和计算机可读存储介质形式进行混合存储。

[0110] 当密钥信息与DNA产物分开存放时，有利于进一步提升数据存储的安全性和保密性。

[0111] 在存储所扩增得到的DNA产物时，可以通过冻干的方式存储，或者也可以通过液体的形态进行存放。存放的温度可以为-20度或-80度等。或者，扩增得到的DNA产物可以存放在离心管、冻存管中，或者也可以通过蜡滴的方式保存扩增得

到的DNA产物。

[0112] 其中，在重叠区域为4个碱基，核心序列为8个碱基时，可以经过计算模拟确定密钥信息。比如密钥信息所记录的核心序列的位置信息包括起始位置的核心序列、末尾位置的核心序列和中心位置的核心序列。比如，S102所确定的碱基序列为：

[0113] “CGTTGCAAGATGCGTCGCCGACGCGTAGCGAGCCTCGTCGTGGGATACGTTGATGGGCACGTTGGCTGGACCGGTGTGCGTTTCGTTGGATGTATCGTGGAGAGTAACGTAGCGAGAAACGTAGCGCGCTGCGGAGAGTGCGGCGTGGTCAGGCCCGTTGTACGAAT”。

[0114] 与该碱基序列对应的密钥信息可以包括该碱基序列的第一位开始的4碱基CGTT，从第83位开始的4碱基TTCG，以及从165位开始的4碱基GAAT，记录为“1=CGTT；83=TTCG；168=GAAT”

[0115] 在可能的实现方式中，该密钥信息还可以包括接头序列和/或拆分所述碱基序列得到的子碱基序列所对应的索引序列中的一项或者两项。

[0116] 图7为本申请实施例提供的一种DNA介质存储数据的解码的方法的实现流程示意图，该方法包括：

[0117] 在S701中，获取待解码的DNA序列及其密钥信息。

[0118] 测序读取的方式包括任意可以读取DNA产物的方式，比如二代测序，三代测序等等，获取DNA产物所对应的待解码的DNA序列。

[0119] 其中，待解码的DNA序列及其密钥信息，即为图1所示的数据存储方法所得到的扩增后的DNA序列及密钥信息。

[0120] 其中，密钥信息可以包括起始位置的核心序列和/或末尾位置的核心序列。通过确定起始位置的核心序列和/或末尾位置的核心序列，可以根据所确定的核心序列进行快速的核心序列的组合。

[0121] 在可能的实现方式中，密钥信息还可以包括接头序列、拆分所述碱基序列得到的子碱基序列所对应的索引序列中的一项或者两项。通过接头序列可以对DNA序列中的接头信息进行分割，从而得到DNA序列中所包括的核心序列，并且通过接头信息可以区分核心序列的方向，便于准确的进行核心序列的组合操作。

[0122] 通过索引序列可以便于确定所组合的多个碱基子序列的顺序，从而根据所确定

的顺序，得到准确的碱基序列。

[0123] 在S702中，根据预设的接头序列，提取所述DNA序列中包括的核心序列。

[0124] 其中，预设的接头序列可以预先设定，可在解码时直接调取使用。或者，对于不同的数据存储方式，可以选用不同的接头序列，并通过DNA方式存储在密钥信息中。在解码时可通过解析密钥信息，提取所设定的接头信息，从而提高DNA数据存储的安全性。

[0125] 该接头序列可以包括左接头序列和/或右接头序列，通过一个或者两个接头序列，可以用于标示核心序列的方向，便于解码得到正确方向的核心序列。

[0126] 比如，对于图5所示的DNA序列对应的序列块，可以根据预先设定的前接头序列“CGCCAGGGTTTTCCCAGTCACGAC”，以及预先设定的后接头序列“TCCTGTGTGAAAT TGTATCCGCT”，对26个序列块进行切割，得到图4所示的核心序列。

[0127] 在S703中，根据所述密钥信息，结合所述核心序列之间的重叠区域，将所述核心序列组合生成碱基序列。

[0128] 本申请实施例中使用该密钥信息的解密过程，与使用密钥信息进行编码存储过程对应。

[0129] 比如，与图1所示的数据存储方式对应的，根据密钥信息中包括的一个以上的核心序列的位置信息，准确确定需要组合的核心序列的初始位置，根据该初始位置对其它核心序列进行组合，从而确定其它核心序列在碱基序列中的位置。

[0130] 根据核心序列之间的重叠区域，包括相邻的序列之间的重叠区域，或者奇数位相邻的重叠区域和偶数位相邻的重叠区域，对核心序列进行组合，得到拼接后的碱基序列。

[0131] 比如，对于图4所示的核心序列示意图，通过核心序列之间的重叠区域，结合预先设定的密钥信息，可以得到待解密的DNA序列所对应的碱基序列为“CGTTGC AAGATGCGTCGCCGACGCGTAGCGAGCCTCGTCGTGGGATACGTTGATGGGCACGTTGGCTGGACCGG TGTGCGTTTCGTTGGATGTATCGTGGAGAGTAACGTAGCGAGAAACGTAGCGCGCTGCGGAGAGTGCGG CGTGGTCAGGCCCGTTGTACGAAT”。

[0132] 其中，密钥信息中的核心序列以及所包括的核心序列的位置信息中，可以包括第一个位置的核心序列和末尾位置的核心序列。在可能的实现场景中，如果DNA

产物对应的DNA序列中的DNA分子类型较少，可以直接根据重叠区域进行DNA序列的拼接得到碱基序列。

[0133] 在S704中，根据预设的映射关系，将所述碱基序列转换为二进制序列。

[0134] 可以预先设定的与S102的数据存储时匹配的映射关系，将碱基序列转换为二进制序列，比如，可以将S703所示的碱基序列转换为二进制序列为：“111001011011000010000110111001111011111010001110111001001011100010111101111001111001101010000100111001011000011010101100111001011010110110100011111010011001101110010101111001011010000110010001111001101000100010010000111100100101110001000000011100100101110111011011011010001000100110111010111001101001110010101111111001011001001110000001”。

[0135] 在S705中，根据所转换的二进制数据生成数据文件。

[0136] 根据所生成的二进制数据，结合预先所设定的编码规则，可以将二进制数据文件转换为对应的数据文件，包括如图片、文本、程序、音频、视频等文件。

[0137] 应理解，上述实施例中各步骤的实现，可以通过人为计算实现，也可以通过计算机程序实现。并且上述各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

[0138] 综上，本申请实施例中利用DNA进行数据存储的方法，具有以下作用：

[0139] 1、在单次数据存储时，可以从预先在体外合成的DNA分子库中获取预定数量的少量的合成的DNA分子，从而能够大大的减少一次数据存储所使用合成DNA分子数，可以保证一次体外合成的DNA分子的多次调用，有效的降低了用于数据存储的DNA合成成本。

[0140] 2、预先合成的DNA分子库可以反复调用，避免了针对不同的二进制数据，需要合成对应数据至少二分之一碱基数量的情况，节省了总体的碱基合成数量，进一步降低了存储数据的合成成本。

[0141] 3、由于本申请通过密钥信息对DNA产物中存储的数据进行解读，可以有效的提升数据存储的安全性。

[0142] 图8为本申请实施例提供的一种利用DNA进行数据存储的装置，该装置包括：

- [0143] 二进制序列提取单元801，用于提取待存储数据对应的二进制序列；
- [0144] 第一序列转换单元802，用于根据预设的映射关系，将所述二进制序列转换为碱基序列；
- [0145] 碱基分割单元803，用于将所述碱基序列分割为多个预设第一长度的核心序列，且两个关联的核心序列之间包括预设第二长度的重叠区域；
- [0146] 序列拼接单元804，用于将所述核心序列与用于标示序列方向的接头序列进行拼接得到序列块；
- [0147] DNA分子提取单元805，用于在预先合成的DNA分子库中查找与所述序列块匹配的DNA序列，取预定数量的所提提取的DNA序列对应的、合成的DNA分子进行扩增得到DNA产物；
- [0148] DNA存储单元806，用于存储所扩增获得的DNA产物序列及对应的密钥信息，所述密钥信息包括一个以上所述核心序列，或者，所述密钥信息包括一个以上的所述核心序列中的部分碱基，以及所包括的所述核心序列在所述碱基序列中的位置信息。
- [0149] 该利用DNA进行数据存储的装置，与图1所示的利用DNA进行数据存储的方法对应。
- [0150] 图9为本申请实施例提供的一种DNA介质存储数据的解码的装置的示意图，该装置包括：
- [0151] DNA序列获取单元901，用于获取待解码的DNA序列及其密钥信息；
- [0152] 核心序列提取单元902，用于根据预设的接头序列，提取所述DNA序列中包括的核心序列；
- [0153] 序列组合单元903，用于根据所述密钥信息，结合所述核心序列之间的重叠区域，将所述核心序列组合生成碱基序列；
- [0154] 第二序列转换单元904，用于根据预设的映射关系，将所述碱基序列转换为二进制序列；
- [0155] 数据文件生成单元905，用于根据所转换的二进制数据生成数据文件。
- [0156] 图9所示的DNA介质存储数据的解码的装置，与图7所示的DNA介质存储数据的解码的方法对应。

- [0157] 图10是本申请一实施例提供的存储设备的示意图。如图10所示，该实施例的存储设备10包括：处理器100、存储器101以及存储在所述存储器101中并可在所述处理器100上运行的计算机程序102，例如利用DNA进行数据存储或解码程序。所述处理器100执行所述计算机程序102时实现上述各个利用DNA进行数据存储或解码方法实施例中的步骤。或者，所述处理器100执行所述计算机程序102时实现上述各装置实施例中各模块/单元的功能。
- [0158] 示例性的，所述计算机程序102可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器101中，并由所述处理器100执行，以完成本申请。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述所述计算机程序102在所述存储设备10中的执行过程。
- [0159] 所述存储设备可包括，但不仅限于，处理器100、存储器101。本领域技术人员可以理解，图10仅仅是存储设备10的示例，并不构成对存储设备10的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如所述存储设备还可以包括输入输出设备、网络接入设备、总线等。
- [0160] 所称处理器100可以是中央处理单元(Central Processing Unit, CPU)，还可以是其他通用处理器、数字信号处理器(Digital Signal Processor, DSP)、专用集成电路(Application Specific Integrated Circuit, ASIC)、现成可编程门阵列(Field-Programmable Gate Array, FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。
- [0161] 所述存储器101可以是所述存储设备10的内部存储单元，例如存储设备10的硬盘或内存。所述存储器101也可以是所述存储设备10的外部存储设备，例如所述存储设备10上配备的插接式硬盘，智能存储卡(Smart Media Card, SMC)，安全数字(Secure Digital, SD)卡，闪存卡(Flash Card)等。进一步地，所述存储器101还可以既包括所述存储设备10的内部存储单元也包括外部存储设备。所述存储器101用于存储所述计算机程序以及所述存储设备所需的其他程序和数据。所述存储器101还可以用于暂时地存储已经输出或者将要输出的数据。

- [0162] 所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将所述装置的内部结构划分成不同的功能单元或模块，以完成以上描述的全部或者部分功能。实施例中的各功能单元、模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中，上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。另外，各功能单元、模块的具体名称也只是为了便于相互区分，并不用于限制本申请的保护范围。上述系统中单元、模块的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。
- [0163] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述或记载的部分，可以参见其它实施例的相关描述。
- [0164] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。
- [0165] 在本申请所提供的实施例中，应该理解到，所揭露的装置/终端设备和方法，可以通过其它的方式实现。例如，以上所描述的装置/终端设备实施例仅仅是示意性的，例如，所述模块或单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通讯连接可以是通过一些接口，装置或单元的间接耦合或通讯连接，可以是电性，机械或其它的形式。
- [0166] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0167] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0168] 所述集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一计算机可读存储介质中，该计算机程序在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机程序包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读介质不包括是电载波信号和电信信号。

[0169] 以上所述实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围，均应包含在本申请的保护范围之内。

权利要求书

- [权利要求 1] 一种利用DNA进行数据存储的方法，其特征在于，所述方法包括：
- 提取待存储数据对应的二进制序列；
- 根据预设的映射关系，将所述二进制序列转换为碱基序列；
- 将所述碱基序列分割为多个预设第一长度的核心序列，且两个关联的核心序列之间包括预设第二长度的重叠区域；
- 将所述核心序列与用于标示序列方向的接头序列进行拼接得到序列块；
- 在预先合成的DNA分子库中查找与所述序列块匹配的DNA序列，取预定数量的所提取的DNA序列对应的、合成的DNA分子进行扩增得到DNA产物；
- 存储所扩增获得的DNA产物及对应的密钥信息，所述密钥信息包括一个以上所述核心序列，或者，所述密钥信息包括一个以上的所述核心序列中的部分碱基，以及所包括的所述核心序列在所述碱基序列中的位置信息。
- [权利要求 2] 根据权利要求1所述的方法，其特征在于，所述接头序列包括左接头序列和右接头序列中的一项或者两项，当所述接头序列包括左接头序列和右接头序列时，所述左接头序列与所述右接头序列不同。
- [权利要求 3] 根据权利要求1所述的方法，其特征在于，将所述二进制序列转换为碱基序列，包括：
- 对所述待存储数据对应的文件进行拆分，根据文件拆分的结果对所述碱基序列进行拆分，并为拆分的碱基序列分配对应的索引序列。
- [权利要求 4] 根据权利要求1所述的方法，其特征在于，两个关联的核心序列之间包括预设第二长度的重叠区域，包括：
- 相邻两个核心序列之间包括预设第二长度的重叠区域；
- 或者，相邻的奇数位的核心序列之间包括预设第二长度的重叠区域，以及相邻的偶数位的核心序列之间包括预设第二长度的重叠区域；
- 或者，第M+i位和第M+N+i位的核心序列之间包括预设第二长度的重叠

区域，其中，M和N为预定的整数，i为大于或等于0的整数变量。

- [权利要求 5] 根据权利要求1所述的方法，其特征在于，将所述碱基序列分割为多个预设第一长度的核心序列，包括：
- 当划分得到的最后一个核心序列的长度小于第一长度时，通过预设的重复碱基补齐所述最后一个核心序列。
- [权利要求 6] 根据权利要求1所述的方法，其特征在于，存储所扩增获得的DNA产物及对应的密钥信息，包括：
- 根据预设的碱基单元与位置信息的映射关系，将所述密钥信息中的位置信息转换为碱基单元，根据预设的组合方式，将密钥信息中的核心序列，以及所述碱基单元，以DNA形式进行存储；
- 或者，所述密钥信息中的核心序列，以及所包括的核心序列的位置信息通过计算机可读存储介质存储；
- 或者，所述的密钥信息通过DNA形式和计算机可读存储介质形式进行混合存储。
- [权利要求 7] 根据权利要求1-6任一项所述的方法，其特征在于，所述密钥信息包括起始位置的核心序列和/或末尾位置的核心序列。
- [权利要求 8] 根据权利要求7所述的方法，其特征在于，所述密钥信息还包括所述接头序列、拆分所述碱基序列得到的子碱基序列所对应的索引序列中的一项或者两项。
- [权利要求 9] 一种利用DNA进行数据存储的装置，其特征在于，所述装置包括：
- 二进制序列提取单元，用于提取待存储数据对应的二进制序列；
- 第一序列转换单元，用于根据预设的映射关系，将所述二进制序列转换为碱基序列；
- 碱基分割单元，用于将所述碱基序列分割为多个预设第一长度的核心序列，且两个关联的核心序列之间包括预设第二长度的重叠区域；
- 序列拼接单元，用于将所述核心序列与用于标示序列方向的接头序列进行拼接得到序列块；
- DNA分子提取单元，用于在预先合成的DNA分子库中查找与所述序列块

匹配的DNA序列，取预定数量的所提取的DNA序列对应的、合成的DNA分子进行扩增得到DNA产物；

DNA存储单元，用于存储所扩增获得的DNA产物及对应的密钥信息，所述密钥信息包括一个以上所述核心序列，或者，所述密钥信息包括一个以上所述核心序列中的部分碱基，以及所包括的所述核心序列在所述碱基序列中的位置信息。

- [权利要求 10] 一种DNA介质存储数据的解码方法，其特征在于，所述方法包括：
- 获取待解码的DNA序列及其密钥信息；
 - 根据预设的接头序列，提取所述DNA序列中包括的核心序列；
 - 根据所述密钥信息，结合所述核心序列之间的重叠区域，将所述核心序列组合生成碱基序列；
 - 根据预设的映射关系，将所述碱基序列转换为二进制序列；
 - 根据所转换的二进制数据生成数据文件。
- [权利要求 11] 根据权利要求10所述的方法，其特征在于，根据预设的接头序列，提取所述DNA序列中包括的核心序列，包括：
- 根据预设的接头序列，对所述DNA序列进行切割得到核心序列；
 - 根据所述接头序列的位置，确定所得到的核心序列的方向。
- [权利要求 12] 根据权利要求10所述的方法，其特征在于，根据所述密钥信息，结合所述核心序列之间的重叠区域，将所述核心序列组合生成碱基序列，包括：
- 根据所述密钥信息确定一个以上的核心序列在碱基序列中的位置；
 - 根据所述核心序列之间的重叠区域，确定所述核心序列之间的相对位置关系，根据确定了相对位置关系的核心序列生成所述碱基序列。
- [权利要求 13] 根据权利要求10-12任一项所述的方法，其特征在于，根据预设的接头序列，提取所述DNA序列中包括的核心序列，包括：
- 当提取的核心序列的末尾包括预设的重复碱基时，去除该核心序列的末尾所包括的所述重复碱基。
- [权利要求 14] 一种DNA介质存储数据的解码的装置，其特征在于，所述装置包括：

DNA序列获取单元，用于获取待解码的DNA序列及其密钥信息；

核心序列提取单元，用于根据预设的接头序列，提取所述DNA序列中包括的核心序列；

序列组合单元，用于根据所述密钥信息，结合所述核心序列之间的重叠区域，将所述核心序列组合生成碱基序列；

第二序列转换单元，用于根据预设的映射关系，将所述碱基序列转换为二进制序列；

数据文件生成单元，用于根据所转换的二进制数据生成数据文件。

[权利要求 15] 一种DNA分子库的生成方法，其特征在于，所述方法包括：
根据碱基组合生成核心序列，所述核心序列包括用于存储数据的碱基片段；
在所述核心序列上拼接预设的接头序列，得到所述核心序列对应的DNA序列，所述接头序列用于标识所述核心序列的方向；
根据所述DNA序列合成对应的DNA分子，根据所合成的DNA分子得到DNA分子库。

[权利要求 16] 一种存储设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至8任一项所述利用DNA进行数据存储的方法的步骤，或者执行所述计算机程序时实现如权利要求10-13任一项所述DNA介质存储数据的解码的方法，或者权利要求15所述的DNA分子库的生成方法。

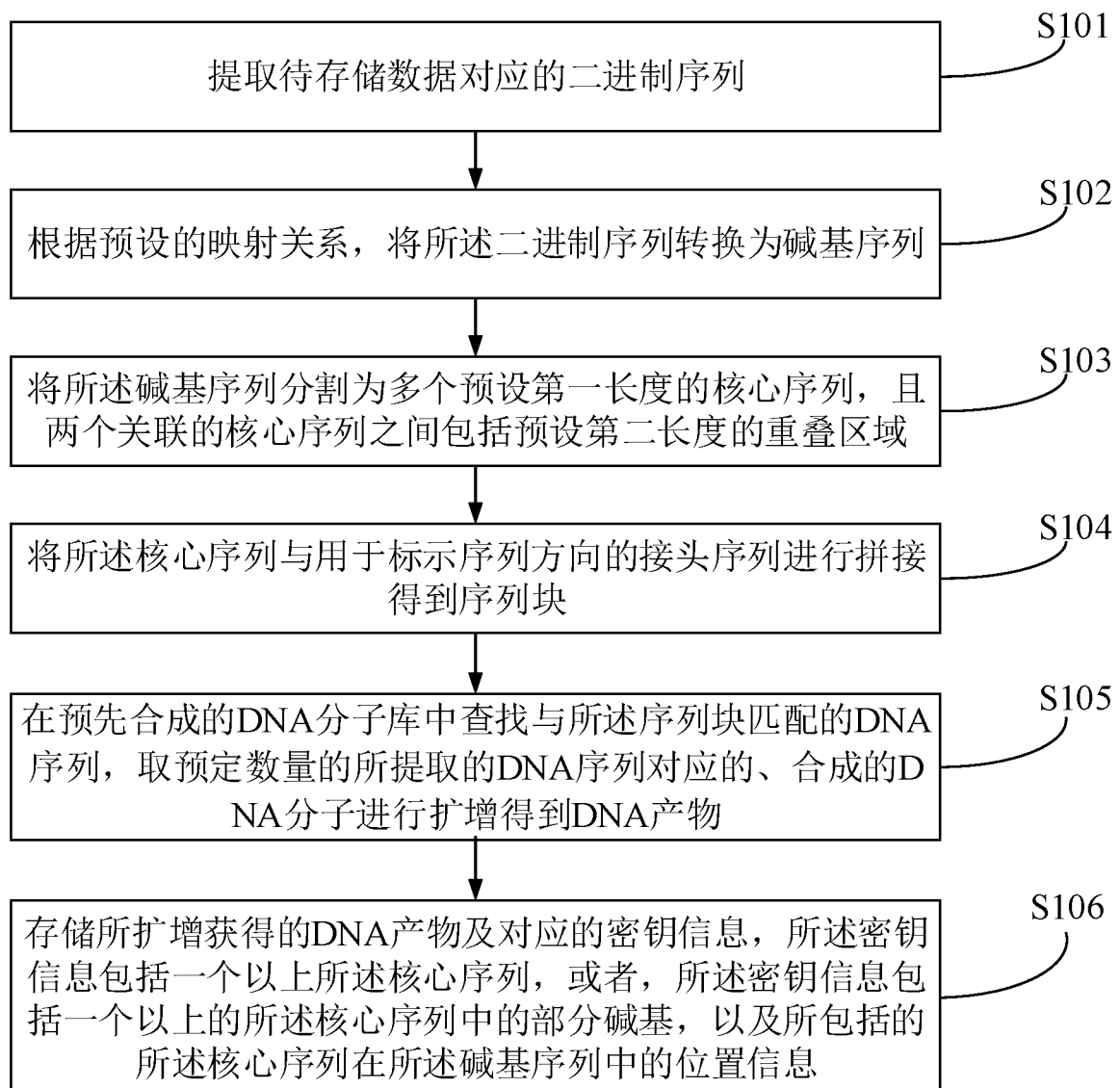


图 1

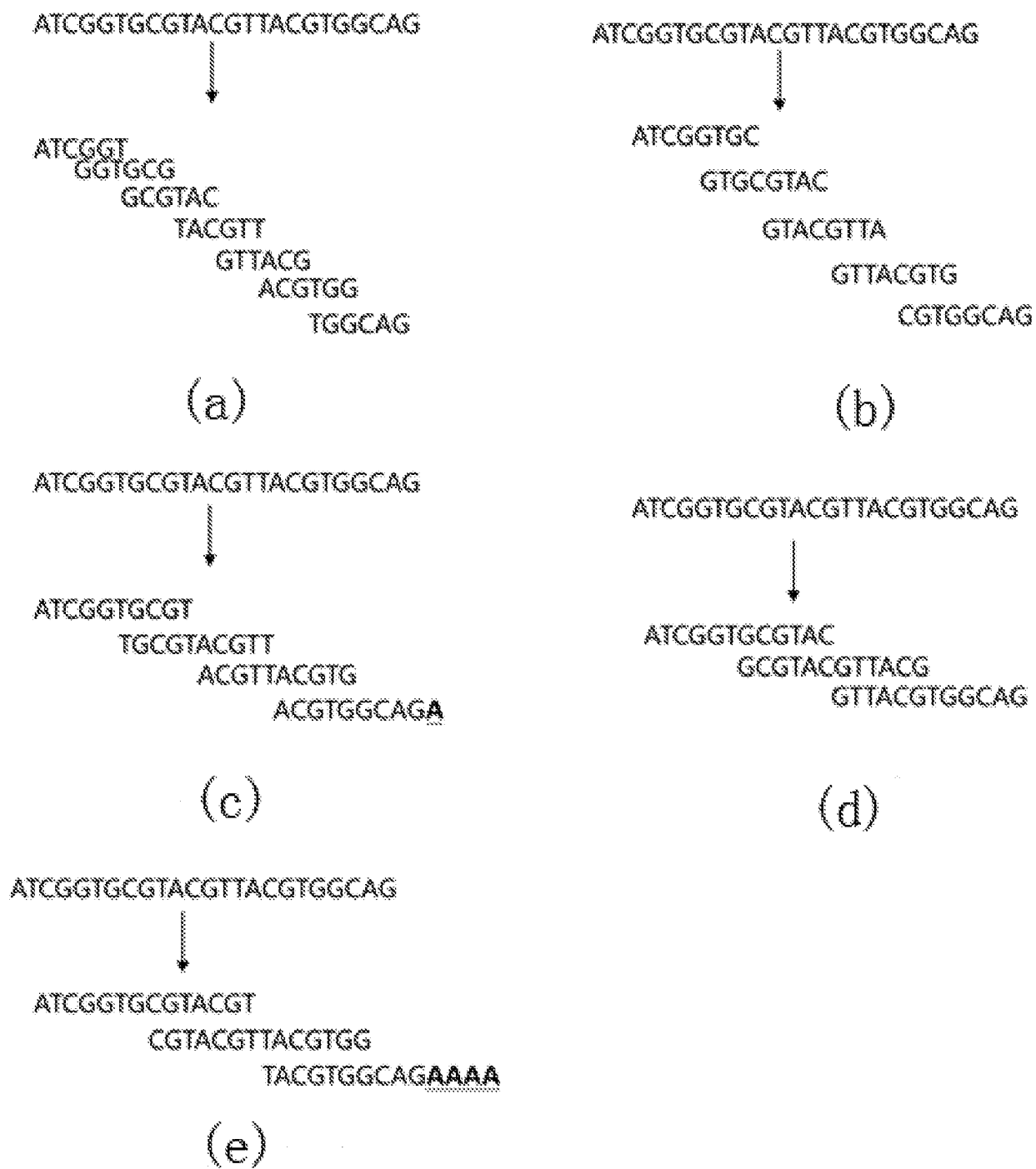


图 2

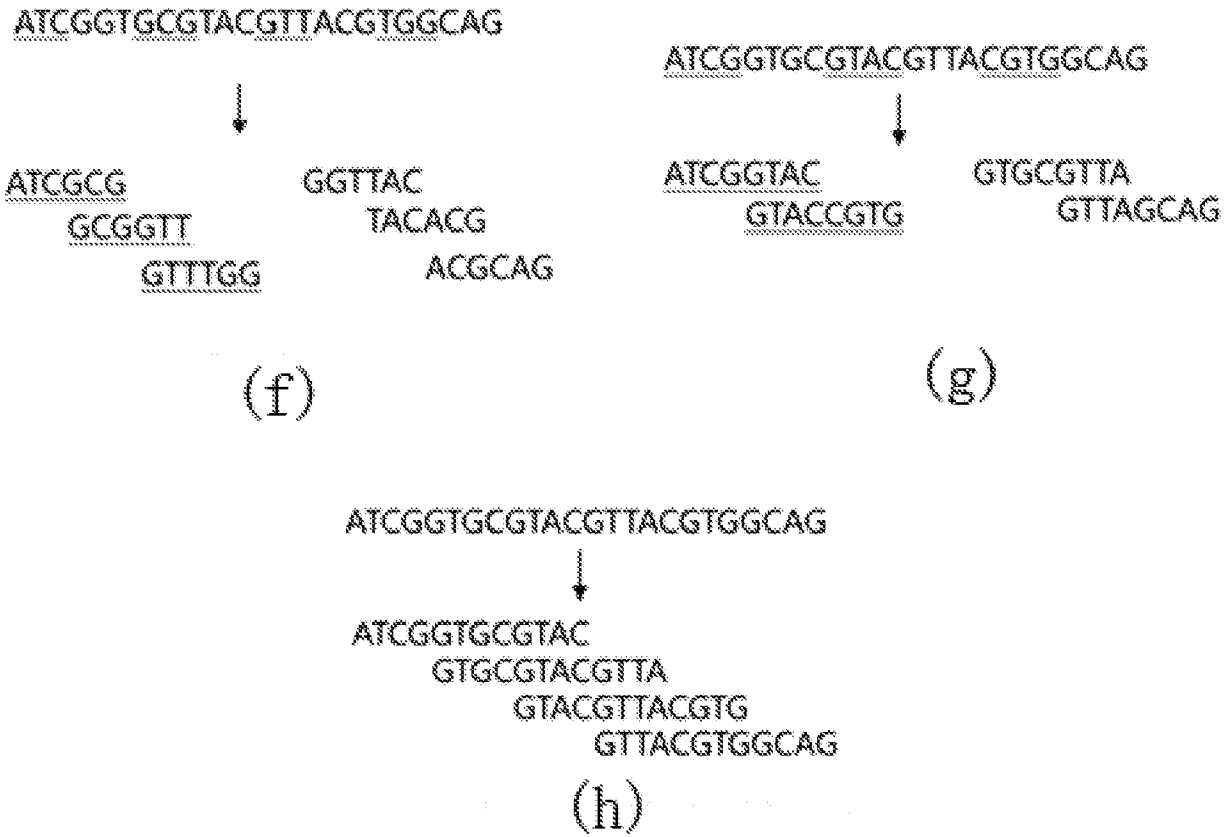


图 3

CGTTGCAA	GCAAGATG	GATGCGTC	CGTCGCCG	GCCGGACG
GACGCGTA	CGTAGCGA	GCGAGCCT	GCCTCGTC	CGTCGTGG
GTGGGATA	GATACGTT	CGTTGATG	GATGGGCA	GGCACGTT
CGTTGGCT	GGCTGGAC	GGACCGGT	CGGTGTGC	GTGCGTTT
GTTTCGTT	CGTTGGAT	GGATGTAT	GTATCGTG	CGTGGAGA
GAGAGTAA	GTAACGTA	CGTAGCGA	GCGAGAAA	GAAACGTA
CGTAGCGC	GCGCGCTG	GCTGCGGA	CGGAGAGT	GAGTGCGG
GCGGCGTG	CGTGGTCA	GTCAGGCC	GGCCCGTT	CGTTGTAC
GTACGAAT				

图 4

CGCCAGGGTTTTCCCAGTCACGAC <u>CGTTGCAAT</u> CCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGAC <u>GCAAGAT</u> GTCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGATGCGTCTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTCGCCGTCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGCCGACGTCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGACGCGTATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTAGCGATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGCGAGCCTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGCCTCGTCTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTCGTGGTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGTGGGATATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGATACGTTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTTGATGTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGATGGGCATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGGCACGTTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTTGGCTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGGCTGGACTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGGACCGTTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGGTGTGCTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGTGCGTTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGTTTCGTTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTTGGATTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGGATGTATTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGTATCGTGCTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTGGAGATCCTGTGTGAAATTGTTATCCGCT

CGCCAGGGTTTTCCCAGTCACGACGAGAGTAATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGTAACGTATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTAGCGATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGCGAGAAATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGAAACGTATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTAGCGCTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGCGCGCTGTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGCTGCGGATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGGAGAGTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGAGTGCGGTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGCGGCGTGTCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTGGTCATCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGTCAGGCCTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGGCCCGTTTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACCGTTGTACTCCTGTGTGAAATTGTTATCCGCT
CGCCAGGGTTTTCCCAGTCACGACGTACGAATTCCTGTGTGAAATTGTTATCCGCT

图 5

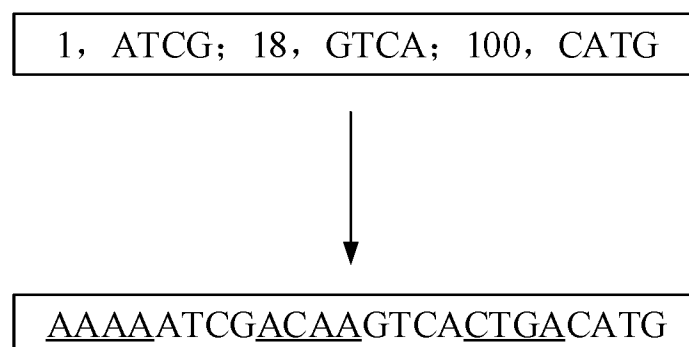


图 6

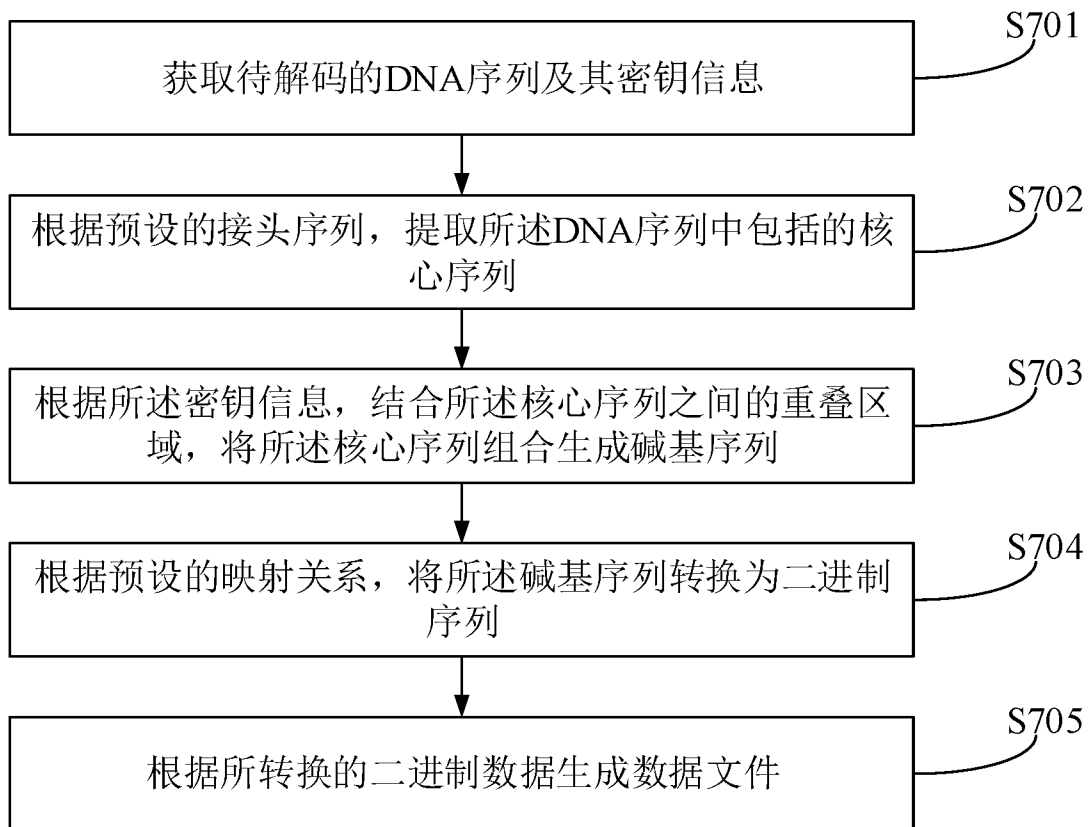


图 7

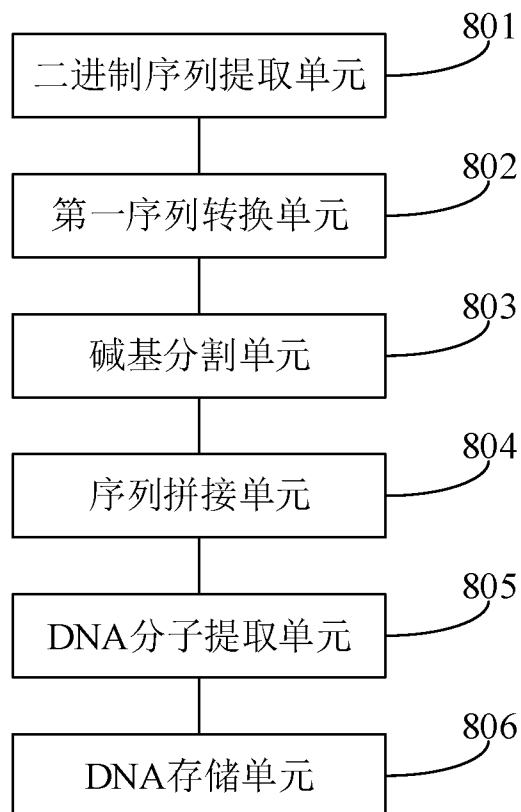


图 8

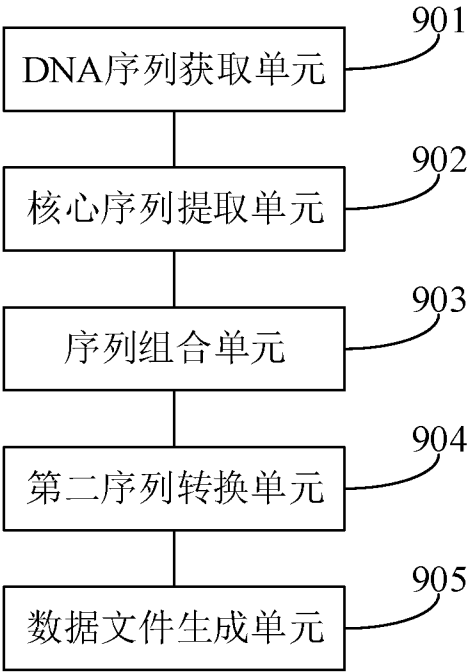


图 9

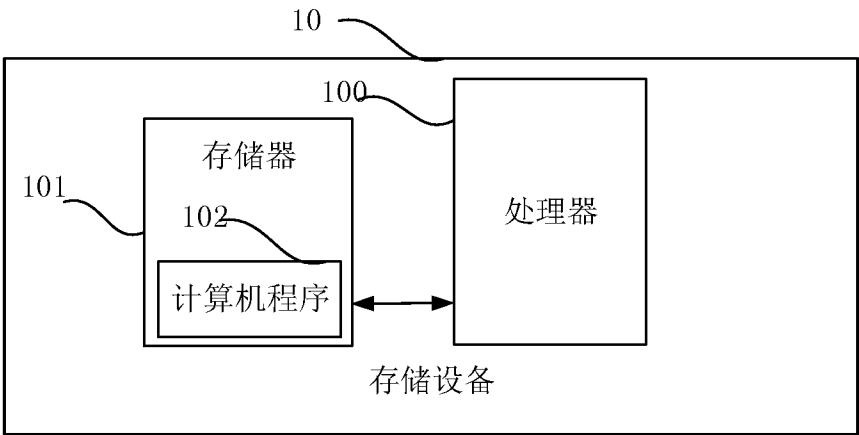


图 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/101247

A. CLASSIFICATION OF SUBJECT MATTER

G16B 30/00(2019.01)i; G16B 50/30(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16B G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, SIPOABS, DWPI, CNKI: 脱氧核糖核酸; 数据; 存储; 碱基; 重复; 重叠; 活字; 索引; 映射; 序列; 密钥; 位置; DNA; data; storage; base sequence; index; mapping; sequence; key; location

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 112382340 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 19 February 2021 (2021-02-19) description, paragraphs [0005]-[0172]	10-16
Y	CN 112382340 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 19 February 2021 (2021-02-19) description, paragraphs [0005]-[0172]	1-9, 16
Y	CN 111858510 A (BEIJING INSTITUTE OF GENOMICS, CHINESE ACADEMY OF SCIENCES (CHINA NATIONAL CENTER FOR BIOINFORMATION)) 30 October 2020 (2020-10-30) description, paragraphs [0007]-[0112]	1-9, 16
A	WO 2020094810 A1 (BIOSISTEMIKA D.O.O.) 14 May 2020 (2020-05-14) entire document	1-16
A	CN 105022935 A (QINGDAO INSTITUTE OF BIOENERGY AND BIOPROCESS TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 04 November 2015 (2015-11-04) entire document	1-16



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

25 February 2022

Date of mailing of the international search report

15 March 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/101247

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112382340	A	19 February 2021	None			
CN	111858510	A	30 October 2020	CN	111858510	B	20 August 2021
WO	2020094810	A1	14 May 2020	EP	3651156	A1	13 May 2020
				EP	3877981	A1	15 September 2021
CN	105022935	A	04 November 2015	None			

国际检索报告

国际申请号

PCT/CN2021/101247

A. 主题的分类

G16B 30/00(2019.01)i; G16B 50/30(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G16B G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNABS, SIPOABS, DWPI, CNKI: 脱氧核糖核酸; 数据; 存储; 碱基; 重复; 重叠; 活字; 索引; 映射; 序列; 密钥; 位置; DNA; data; storage; base sequence; index; mapping; sequence; key; location

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 112382340 A (中国科学院深圳先进技术研究院) 2021年2月19日 (2021 - 02 - 19) 说明书第[0005]-[0172]段	10-16
Y	CN 112382340 A (中国科学院深圳先进技术研究院) 2021年2月19日 (2021 - 02 - 19) 说明书第[0005]-[0172]段	1-9, 16
Y	CN 111858510 A (中国科学院北京基因组研究所国家生物信息中心) 2020年10月30日 (2020 - 10 - 30) 说明书第[0007]-[0112]段	1-9, 16
A	WO 2020094810 A1 (BIOSISTEMIKA D00) 2020年5月14日 (2020 - 05 - 14) 全文	1-16
A	CN 105022935 A (中国科学院青岛生物能源与过程研究所) 2015年11月4日 (2015 - 11 - 04) 全文	1-16

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年2月25日

国际检索报告邮寄日期

2022年3月15日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

张妍

电话号码 86-(10)-62412178

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/101247

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	112382340	A	2021年2月19日	无	
CN	111858510	A	2020年10月30日	CN 111858510	B 2021年8月20日
WO	2020094810	A1	2020年5月14日	EP 3651156	A1 2020年5月13日
				EP 3877981	A1 2021年9月15日
CN	105022935	A	2015年11月4日	无	