(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113782102 A
(43)申请公布日 2021.12.10

(21)申请号 202110929436.9

(22)申请日 2021.08.13

(71)申请人 深圳先进技术研究院
地址 518055 广东省深圳市南山区西丽大学城学苑大道1068号

(72)发明人 戴俊彪 黄小罗

(74)专利代理机构 深圳中一联合知识产权代理有限公司 44414
代理人 黄志云

(51)Int.Cl.
G16B 50/30(2019.01)

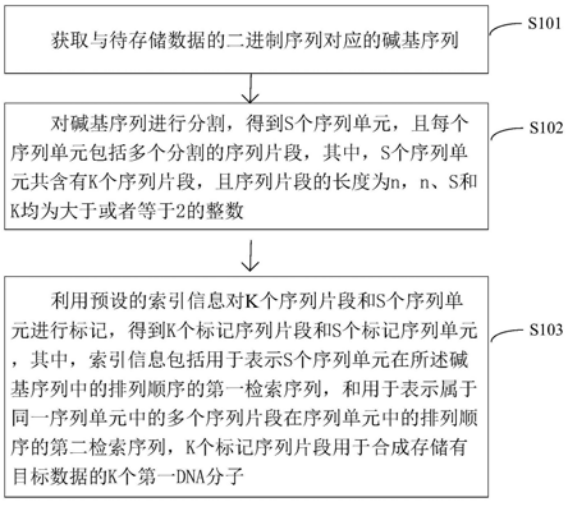
权利要求书2页 说明书14页 附图4页

(54)发明名称

DNA数据的存储方法、装置、设备及可读存储介质

(57)摘要

本申请涉及数据存储技术领域，提供了一种DNA数据的存储方法、装置、设备及可读存储介质领域。该方法，包括：获取与目标数据的二进制序列对应的碱基序列；对碱基序列进行分割，得到S个序列单元，且每个序列单元包括多个分割的序列片段，S个序列单元共含有K个序列片段，且序列片段的长度为n；利用预设的索引信息对K个序列片段和S个序列单元进行标记，得到K个标记序列片段和S个标记序列单元，其中，索引信息包括用于表示S个序列单元在碱基序列中的排列顺序的第一检索序列，和用于表示属于同一序列单元中的多个序列片段在序列单元中的排列顺序的第二检索序列。本申请提供的方法，可以实现大规模的数据信息在DNA中的存储。



CN 113782102 A

1.DNA数据的存储方法,其特征在于,包括:

获取与目标数据的二进制序列对应的碱基序列;

对所述碱基序列进行分割,得到S个序列单元,且每个序列单元包括多个分割的序列片段,其中,S个所述序列单元共含有K个所述序列片段,且所述序列片段的长度为n,n、S和K均为大于或者等于2的整数;

利用预设的索引信息对K个所述序列片段和S个所述序列单元进行标记,得到K个标记序列片段和S个标记序列单元,其中,所述索引信息包括用于表示S个所述序列单元在所述碱基序列中的排列顺序的第一检索序列,和用于表示属于同一所述序列单元中的多个所述序列片段在所述序列单元中的排列顺序的第二检索序列,K个所述标记序列片段用于合成存储有所述目标数据的K个第一DNA分子。

2.根据权利要求1所述的DNA数据的存储方法,其特征在于,利用所述第二检索序列标记属于同一所述序列单元中的多个所述序列片段的方式,包括:

在所述序列片段的任一侧拼接第二检索序列,或

在所述序列片段的两侧同时拼接检索碱基组,两侧的所述检索碱基组形成所述第二检索序列。

3.根据权利要求1所述的DNA数据的存储方法,其特征在于,所述第一检索序列包括i条DNA序列片段,i为大于或等于1的整数,且每条所述DNA序列片段包括用作索引标志的第一碱基序列和用于标示所述序列单元编号的第二碱基序列。

4.根据权利要求1所述的DNA数据的存储方法,其特征在于,所述存储方法还包括:

将K个所述标记序列片段合成存储有所述目标数据的K个第一DNA分子后,将K个所述第一DNA分子存储在S个第一物理空间,其中,同属于一个所述序列单元的所述标记序列片段对应的所述第一DNA分子存储在同一个所述第一物理空间,不属于同一个所述序列单元的所述标记序列片段对应的所述第一DNA分子存储在不同的所述第一物理空间。

5.根据权利要求4所述的DNA数据的存储方法,其特征在于,s个所述第一物理空间集成在一个DNA硬盘中。

6.根据权利要求4所述的DNA数据的存储方法,其特征在于,所述存储方法还包括:

将第二DNA分子存储在与所述第一物理空间对应的第二物理空间,所述第二DNA分子存储有所述索引信息。

7.根据权利要求4至6任一项所述的DNA数据的存储方法,其特征在于,K个所述第一DNA分子的解码方法,包括:

对每个所述第一物理空间中存储的多个所述第一DNA分子进行测序,得到多个所述标记序列片段;

根据所述第二检索序列对属于同一所述标记序列单元的每个所述标记序列片段对应的所述序列片段进行拼接,得到所述序列单元;根据所述第一检索序列将得到的S个所述序列单元进行拼接,得到所述碱基序列;

将所述碱基序列转换为所述目标数据。

8.一种DNA数据存储装置,其特征在于,包括数据处理模块,

所述数据处理模块,用于获取与目标数据的二进制序列对应的碱基序列;对所述碱基序列进行分割,得到S个序列单元,且每个序列单元包括多个分割的序列片段,其中,S个所

述序列单元共含有K个所述序列片段，且所述序列片段的长度为n，n、S和K均为大于或者等于2的整数；利用预设的索引信息对K个所述序列片段和S个所述序列单元进行标记，得到K个标记序列片段和S个标记序列单元，其中，所述索引信息包括用于表示S个所述序列单元在所述碱基序列中的排列顺序的第一检索序列，和用于表示属于同一所述序列单元中的多个所述序列片段在所述序列单元中的排列顺序的第二检索序列，K个所述标记序列片段用于合成存储有所述目标数据的K个第一DNA分子。

9.根据权利要求8所述的DNA数据存储装置，其特征在于，所述装置还包括：DNA分子合成模块，用于将K个所述标记序列片段合成存储有所述目标数据的K个第一DNA分子。

10.根据权利要求8所述的DNA数据存储装置，其特征在于，所述装置还包括：DNA分子存储模块，

用于将K个所述第一DNA分子存储在S个第一物理空间，其中，同属于一个所述序列单元的所述标记序列片段对应的所述第一DNA分子存储在同一个所述第一物理空间，不属于同一个所述序列单元的所述标记序列片段对应的所述第一DNA分子存储在不同的所述第一物理空间。

11.根据权利要求10所述的DNA数据存储装置，其特征在于，所述DNA分子存储模块还用于将第二DNA分子存储在第二物理空间。

12.根据权利要求10所述的DNA数据存储装置，其特征在于，还包括DNA分子测序模块，用于对每个所述第一物理空间中存储的多个所述第一DNA分子进行测序，得到多个所述标记序列片段；

所述数据处理模块，还用于根据所述第二检索序列对属于同一所述标记序列单元的每个所述标记序列片段对应的所述序列片段进行拼接，得到所述序列单元；根据所述第一检索序列将得到的S个所述序列单元进行拼接，得到所述碱基序列；将所述碱基序列转换为所述目标数据。

13.一种DNA数据存储设备，包括一种终端设备，所述终端设备包括存储器、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7任一项所述的DNA数据的存储方法。

14.一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至7任一项所述的DNA数据的存储方法。

15.一种DNA硬盘，其特征在于，包括多个物理空间，所述物理空间之间由物理材料制成，每个所述物理空间用于存储DNA分子。

16.根据权利要求15所述的DNA硬盘，其特征在于，所述物理材料选自SiO_2、金属氧化物、高分子聚合材料中的至少一种。

DNA数据的存储方法、装置、设备及可读存储介质

技术领域

[0001] 本申请属于数据存储技术领域，具体涉及一种DNA数据的存储方法、装置、设备及可读存储介质。

背景技术

[0002] 人工智能及大数据时代的发展对数据存储需求越来越高，迫切需要存储密度高、存储时间长、维护成本低的新型存储介质。脱氧核糖核酸(DeoxyriboNucleic Acid,DNA)作为一种近年来发展起来的信息存储介质，被认为是未来信息存储最有潜力的介质之一。

[0003] DNA分子具有四种碱基，它们分别是：腺嘌呤(Adenine,A)、胞嘧啶(Cytosine,C)、鸟嘌呤(Guanine,G)和胸腺嘧啶(Thymine,T)。基于DNA的数据存储技术是利用上述四种碱基序列来表示二进制“0”和“1”组成的数据系列。相比较于传统存储介质，DNA数据存储具有存储密度高，存储时间久，维护成本低，生物相容性好的特点。如：1g DNA能够存储超过百万部高清电影，其数据存储密度是目前传统硬盘等硅基存储介质7个数量级以上；同时，DNA能够稳定存储数据千年以上，是现有存储介质存储时间的百倍以上。此外，DNA维护成本低，存百年的维护费用仅是目前现有介质的万分之一。

[0004] DNA数据存储流程通常包含以下步骤：(1)从图片、视频、文本等计算机信息中提取二进制信息；(2)根据二进制与碱基A、T、C、G之间的预设对应关系，将二进制序列信息转换为由碱基A、T、C、G编码形成的、存储有数据信息的A/T/C/G序列(即DNA序列)；(3)采用DNA合成技术或其他技术将编码的A/T/C/G序列转换为DNA化学多聚物分子，并存储在合适的环境中。之后，当需要获取存储的数据时，则可以执行以下步骤：(4)利用DNA测序技术，将存储的DNA化学多聚物分子解读成A/T/C/G序列；(5)利用合适的解码方式将A/T/C/G序列转换为二进制信息；(6)将二进制信息转换为图片、视频、文本等计算机信息。

[0005] 其中，数据编码问题是目前的DNA数据存储方法中的核心问题。

发明内容

[0006] 本申请实施例的目的之一在于：提供一种DNA数据的存储方法、装置、设备及可读存储介质，旨在解决DNA数据存储技术中的数据编码问题。

[0007] 本申请实施例采用的技术方案是：

[0008] 第一方面，提供了一种DNA数据的存储方法，包括：

[0009] 获取与目标数据的二进制序列对应的碱基序列；

[0010] 对所述碱基序列进行分割，得到S个序列单元，且每个序列单元包括多个分割的序列片段，其中，S个所述序列单元共含有K个所述序列片段，且所述序列片段的长度为n，n、S和K均为大于或者等于2的整数；

[0011] 利用预设的索引信息对K个所述序列片段和S个所述序列单元进行标记，得到K个标记序列片段和S个标记序列单元，其中，所述索引信息包括用于表示S个所述序列单元在所述碱基序列中的排列顺序的第一检索序列，和用于表示属于同一所述序列单元中的多个

所述序列片段在所述序列单元中的排列顺序的第二检索序列,K个所述标记序列片段用于合成存储有所述目标数据的K个第一DNA分子。

[0012] 在一个实施例中,利用所述第二检索序列标记属于同一所述序列单元中的多个所述序列片段的方式,包括:

[0013] 在所述序列片段的任一侧拼接第二检索序列,或

[0014] 在所述序列片段的两侧同时拼接检索碱基组,两侧的所述检索碱基组形成所述第二检索序列。

[0015] 在一个实施例中,所述第一检索序列包括i条DNA序列片段,i为大于或等于1的整数,且每条所述DNA序列片段包括用作索引标志的第一碱基序列和用于标示所述序列单元编号的第二碱基序列。

[0016] 在一个实施例中,第一检索序列和第二检索序列对应的DNA序列片段利用DNA合成技术获得。示例性的,DNA合成技术包括但不限于酶法合成、亚磷酰胺合成等。

[0017] 在一个实施例中,第一检索序列和第二检索序列对应的DNA序列片段可以从预先合成的DNA通用分子库中扩增获得,比如PCR技术等。

[0018] 在一个实施例中,所述存储方法还包括:

[0019] 将K个所述标记序列片段合成存储有所述目标数据的K个第一DNA分子后,将K个所述第一DNA分子存储在S个第一物理空间,其中,同属于一个所述序列单元的所述标记序列片段对应的所述第一DNA分子存储在同一个所述第一物理空间,不属于同一个所述序列单元的所述标记序列片段对应的所述第一DNA分子存储在不同的所述第一物理空间。

[0020] 在一个实施例中,s个所述第一物理空间集成在一个DNA硬盘中。

[0021] 在一个实施例中,所述存储方法还包括:

[0022] 将第二DNA分子存储在与所述第一物理空间对应的第二物理空间,所述第二DNA分子存储有所述索引信息。

[0023] 在一个实施例中,K个所述第一DNA分子的解码方法,包括:

[0024] 对每个所述第一物理空间中存储的多个所述第一DNA分子进行测序,得到多个所述标记序列片段;根据所述第二检索序列对属于同一所述标记序列单元的每个所述标记序列片段对应的所述序列片段进行拼接,得到所述序列单元;

[0025] 根据所述第一检索序列将得到的S个所述序列单元进行拼接,得到所述碱基序列;

[0026] 将所述碱基序列转换为所述目标数据。

[0027] 第二方面,提供了一种DNA数据存储装置,包括数据处理模块,

[0028] 所述数据处理模块,用于获取与目标数据的二进制序列对应的碱基序列;对所述碱基序列进行分割,得到K个长度为n的序列片段,K个所述序列片段划分为S个序列单元,S和K均为大于或者等于2的整数;利用预设的索引信息对K个所述序列片段和S个所述序列单元进行标记,得到K个标记序列片段和S个标记序列单元,其中,所述索引信息包括用于表示S个所述序列单元在所述碱基序列中的排列顺序的第一检索序列,和用于表示属于同一所述序列单元中的多个所述序列片段在所述序列单元中的排列顺序的第二检索序列,K个所述标记序列片段用于合成存储有所述目标数据的K个第一DNA分子。

[0029] 在一个实施例中,所述装置还包括:DNA合成模块,用于将K个所述标记序列片段合成存储有所述目标数据的K个第一DNA分子。

[0030] 在一个实施例中,所述装置还包括:DNA分子存储模块,用于将K个所述第一DNA分子存储在S个第一物理空间,其中,同属于一个所述序列单元的所述标记序列片段对应的所述第一DNA分子存储在同一个所述第一物理空间,不属于同一个所述序列单元的所述标记序列片段对应的所述第一DNA分子存储在不同的所述第一物理空间。

[0031] 在一个实施例中,所述DNA分子存储模块还用于将第二DNA分子存储在第二物理空间。

[0032] 在一个实施例中,还包括DNA分子测序模块,用于对每个所述第一物理空间中存储的多个所述第一DNA分子进行测序,得到多个所述标记序列片段;所述数据处理模块,还用于对根据所述第二检索序列对属于同一所述标记序列单元的每个所述标记序列片段对应的所述序列片段进行拼接,得到所述序列单元;

[0033] 根据所述第一检索序列将得到的S个所述序列单元进行拼接,得到所述碱基序列;

[0034] 将所述碱基序列转换为所述目标数据。

[0035] 第三方面,提供了一种DNA数据存储设备,包括一种终端设备,所述终端设备包括存储器、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第一方面所述的DNA数据的存储方法。

[0036] 第四方面,提供了一种计算机可读存储介质,计算机可读存储介质存储有计算机程序,计算机程序被处理器执行时实现如第一方面的DNA数据的存储方法。

[0037] 第五方面,提供了一种DNA硬盘,包括多个物理空间,所述物理空间由物理材料制成,每个所述物理空间用于存储DNA分子。

[0038] 在一个实施例中,所述物理材料为SiO_2、金属氧化物、高分子聚合材料中的至少一种。这些物理材料形成将DNA分子包裹的物理空间,同时,将不同的DNA分子隔离。

[0039] 本申请实施例提供的DNA数据的存储方法、装置、设备及可读存储介质的有益效果在于:本申请将与目标数据的二进制序列对应的碱基序列分割为S个序列,每个序列单元包括多个分割的序列片段,S个所述序列单元共含有K个所述序列片段,且所述序列片段的长度为n,n、S和K均为大于或者等于2的整数,利用预设的索引信息标记序列片段和序列单元的位置信息,将标记后的序列片段合成为DNA分子后分别存储。通过该方法,可以提升数据信息的存储量,实现大规模的数据信息在DNA中的存储。

[0040] 在一种实施情形中,第一检索序列的长度和标记序列片段(带有第二检索序列的序列片段)的长度不同,则通过长度区别从序列单元中区分第一检索序列和标记序列片段。具体的,以m表示标记序列片段的碱基数,以q表示第二检索序列的碱基数,以i表示第一检索序列中DNA序列片段的条数,以p表示第一检索序列中第二碱基序列的碱基数,通过本申请提供的方法,可以实现含有D个碱基数的DNA数据的存储,其中,D的计算公式如下:

[0041] $D=4^{q}\times(m-q)\times4^{i\times p}$

[0042] 在特定实施例中,m长度为8碱基的情况下,$q=4$,$i=10$,$p=4$,$D=256\times4\times4^{40}=1.23\times10^{27}$。1个碱基存储2bits信息时候,能够存储的信息量$L=2.46\times10^{27}$bits$=3.075\times10^{26}$bytes$=3.075\times10^{5}$ZB,远大于目前的数据存储规模。

[0043] 在另一种实施情形中,第一检索序列的长度和标记序列片段(带有第二检索序列的序列片段)的长度相同;第一检索序列中用作索引标志的第一碱基序列,可以是第二检索碱基序列的一部分,第一碱基序列的碱基数和第二检索碱基序列的碱基数相同。在这种情

况下，以m表示标记序列片段的碱基数，以q表示第二检索序列的碱基数，以i表示第一检索序列中DNA序列片段的条数，以p表示第一检索序列中第二碱基序列的碱基数，通过本申请提供的方法的方法，可以实现含有D个碱基数的DNA数据的存储，其中，D的计算公式如下：

[0044] $D=(4^{q}-i)\times(m-q)\times4^{i\times p}$

[0045] 在特定实施例中，m长度为8碱基的情况下，q＝4，i＝10，p＝4，$D=(256-10)\times4\times4^{40}=1.18\times10^{27}$。1个碱基存储2bits信息时候，能够存储的信息量$L=2.36\times10^{27}$bits＝2.95×10^{26}bytes＝2.95×10^{5}ZB，远大于目前的数据存储规模。

附图说明

[0046] 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。

[0047] 图1是本申请实施例提供的DNA存储装置的组成示意图；

[0048] 图2是本申请实施例提供的DNA数据的存储写入的工艺流程图；

[0049] 图3是本申请实施例提供的S101中的碱基序列经分割后形成包含多个序列片段的序列单元的示意图；

[0050] 图4是本申请实施例提供的S103中的序列单元经过第一检索序列标记，且序列单元中的各序列片段经过第二检索序列标记后，得到分别包含多个信息序列片段的信息序列单元的示意图；

[0051] 图5是本申请实施例提供的将K个第一DNA分子存储在S个不同的第一物理空间后的DNA存储的示意图；

[0052] 图6是本申请实施例提供的DNA数据的解读工艺流程图；

[0053] 图7本申请一实施例提供的终端设备的结构示意图。

具体实施方式

[0054] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本申请。

[0055] 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。

[0056] 应当理解，在本申请说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合，并且包括这些组合。另外，在本申请说明书和所附权利要求书的描述中，术语“第一”、“第二”、“第三”等仅用于区分描述，而不能理解为指示或暗示相对重要性。

[0057] 还应当理解，在本申请说明书中描述的参考“一个实施例”或“一些实施例”等意味着在本申请的一个或多个实施例中包括结合该实施例描述的特定特征、结构或特点。由此，

在本说明书中的不同之处出现的语句“在一个实施例中”、“在一些实施例中”、“在其他一些实施例中”、“在另外一些实施例中”等不是必然都参考相同的实施例，而是意味着“一个或多个但不是所有的实施例”，除非是以其他方式另外特别强调。术语“包括”、“包含”、“具有”及它们的变形都意味着“包括但不限于”，除非是以其他方式另外特别强调。

[0058] 目前，数据编码问题是DNA数据存储技术中的核心技术问题，尤其针对大规模数据来说，缺乏有效的数据编码方法实现大规模数据存储。有鉴于此，本申请提供了一种DNA数据的存储方法，该方法在进行数据编码的过程中，先将待存储数据的二进制序列转换成对应的碱基序列之后，通过两次分割，将碱基序列分割为S个序列单元，且每个序列单元包括多个分割的序列片段，其中，S个所述序列单元共含有K个所述序列片段。然后利用预设的索引信息对K个所述序列片段和S个所述序列单元进行标记。最后将K个标记序列片段合成DNA分子后存储。基于本申请提供的数据编码方式，能够有效实现DNA数据存储。

[0059] 进一步的，本申请提供的数据编码方式，在数据存储量上具有明显的优势。

[0060] 该存储方法通过图1所示的DNA数据存储装置实现。DNA数据存储装置包括数据处理模块、DNA分子合成模块、DNA分子存储模块、DNA分子测序模块。

[0061] 其中，数据处理模块用于实现数据编解码。例如，将待存储的数据转换为二进制信息，在按照预设的二进制数据与碱基的对应关系，把二进制信息转换成碱基序列。之后再按照预设的索引信息对碱基序列进行编码，得到最终用于生成DNA分子的碱基序列。

[0062] DNA分子合成模块用于根据编码好的碱基序列合成DNA分子。DNA分子存储模块能够存放DNA分子。DNA分子测序模块用于将DNA分子翻译成碱基序列。相应的，数据处理模块也可以根据索引信息对DNA分子中测序得到的碱基序列进行解码，并通过数据转换得到DNA分子中存储的数据。

[0063] 在本申请实施例中，DNA数据存储装置可以是一个完整的DNA数据存储设备，即由多个功能模块集成的一个设备。该设备能够实现数据编码、DNA分子合成、DNA分子存储、DNA分子测序以及数据解码的完整流程。

[0064] 在另一个实施例中，DNA数据存储装置也可以是由各个独立的设备构成系统。

[0065] 例如，其中，数据处理模块可以是电脑、服务器、机器人等计算机设备。用于实现数据编解码。DNA分子合成模块可以DNA分子合成仪，用于根据编码好的碱基序列合成DNA分子。DNA分子存储模块可以是DNA硬盘，能够存放DNA分子。DNA分子测序模块可以是DNA分子测序仪，能够实现DNA分子测序功能。

[0066] 图2为本申请实施例提供的一种DNA数据的存储方法的实现流程示意图，具体包括：

[0067] S101.获取与待存储数据的二进制序列对应的碱基序列。

[0068] 该步骤中，获取与待存储数据的二进制序列对应的碱基序列是指，将待存储数据的二进制序列，转换为由A、T、C、G编码形成的、存储有数据信息的碱基序列。

[0069] 在一些实施例中，获取与待存储数据的二进制序列对应的碱基序列包括：

[0070] S111.提取待存储数据对应的二进制序列。

[0071] 待存储数据为可以在终端设备中存在的任何数据信息，可以包括文字、图片、声音、视频、软件、程序等信息，但不限于此。

[0072] 在提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。

[0073] 示例性的,提取文本“春,已不再是想象之外的那只蝴蝶。”对应的二进制序列为“11100110 10011000 10100101 11101111 10111100 10001100 11100101 10110111 10110010 11100100 10111000 10001101 11100101 10000110 10001101 11100110 10011000 10101111 11100110 10000011 10110011 11101000 10110001 10100001 11100100 10111001 10001011 11100101 10100100 10010110 11100111 10011010 10000100 11101001 10000010 10100011 11100101 10001111 10101010 11101000 10011101 10110100 11101000 10011101 10110110 11100011 10000000 10000010”。

[0074] S121.根据预设的映射规则,将二进制序列转换为碱基序列。

[0075] 本申请实施例中,预设的映射规则是指预设的二进制与碱基之间的映射规则。根据二进制代码与碱基A、T、C、G之间预设的映射规则,将有0/1组成的二进制序列转换为由碱基A、T、C、G编码形成的、存储有数据信息的碱基序列。示例性的,二进制序列与碱基A、T、C、G之间的预设对应关系为:一个碱基A代表一个00,一个碱基T代表一个01,一个碱基C代表一个10,一个碱基G代表一个11。当二级制序列为00110110101100101011011011000011001001时,根据二进制与碱基A、T、C、G之间预设的映射规则,将二进制数据信息转换碱基序列为AGTCCGACCGTCGAAGACT的DNA序列。当然,二进制代码与碱基A、T、C、G之间的预设对应关系不限于上述示例,例如,也可以为:一个碱基T代表一个00,一个碱基A代表一个01,一个碱基G代表一个10,一个碱基C代表一个11,但不限于此。应当理解,二进制序列与碱基A、T、C、G之间预设的映射规则,只需要能够根据预设的映射规则将二进制序列转换碱基序列就行,并不限于上述示例。

[0076] 示例性的,二进制与碱基之间的映射规则为:一个碱基A代表一个11,一个碱基T代表一个10,一个碱基C代表一个01,一个碱基G代表一个00。根据该映射规则,可以将文本“春,已不再是想象之外的那只蝴蝶。”对应的二进制序列“11100110 10011000 10100101 11101111 10111100 10001100 11100101 10110111 10110010 11100100 10111000 10001101 11100101 10000110 10001101 11100110 10011000 10101111 11100110 10000011 10110011 11101000 10110001 10100001 11100100 10111001 10001011 11100101 10100100 10010110 11100111 10011010 10000100 11101001 10000010 10100011 11100101 10001111 10101010 11101000 10011101 10110100 11101000 10011101 10110110 11100011 10000000 10000010”转换为碱基序列“ATCT TCTG TTCC ATAA TAAG TGAG ATCC TACA TAGT ATCG TATG TGAC ATCC TGCT TGAC ATCT TCTG TTAA ATCT TGGA TAGA ATTG TAGC TTGC ATCG TATC TGTA ATCC TTCG TCCT ATCA TCTT TGCG ATTC TGGT TTGA ATCC TGAA TTTT ATTG TCAC TACG ATTG TCAC TACT ATGA TGGG TGGT”。

[0077] S102.对碱基序列进行分割,得到S个序列单元,且每个序列单元包括多个分割的序列片段,其中,S个序列单元共含有K个序列片段,且序列片段的长度为n,n、S和K均为大于或者等于2的整数。

[0078] 本申请实施例中，对碱基序列进行分割，经过分割后的碱基序列变成s个序列单元，且每个序列单元包括多个分割的序列片段，从而降低了序列长度，便于后续步骤对序列单元进行分别存储。应当理解的是，本申请实施例所指的长度，是指碱基长度，可以理解为碱基个数。

[0079] 碱基序列分割后得到的每个序列单元包括长度为n的多个序列片段，因此，序列单元的长度为序列片段的整数倍。在一些实施例中，经分割后形成的序列单元的长度相同，即碱基序列分割成s个长度相同的序列单元，且s个长度相同的序列单元同时含有相同数量、且长度为n的序列片段。在一些实施例中，经分割后形成的序列单元的长度不同，示例性的，按照预设的序列单元长度对碱基序列依次进行分割，得到S-1个序列单元，最后一次分割后剩余的碱基序列长度不足预设长度，此时，剩余碱基序列作为一个序列单元，其长度小于其他序列单元的长度。在一些实施例中，碱基序列可以按照其他预设的分割规则进行分割，得到序列单元长度不一致的s个序列单元。

[0080] 在一种可能的实现方式中，对碱基序列进行分割，得到s个序列单元，每个序列单元包括多个分割的序列片段，包括：

[0081] 将碱基序列分割为多个序列单元，且序列单元的长度为n的整数倍；

[0082] 将每个序列单元分割成多个长度为n的序列片段。

[0083] 在一些实施例中，将碱基序列分割为多个序列单元时，从碱基序列的一端开始，每隔一个预设的序列单元长度对碱基序列进行一次分割，得到s个序列单元，其中，预设长度为n的整数倍。当最后一次分割后剩余的碱基序列长度不足预设长度时，将剩余碱基序列作为一个序列单元。在其他实施例中，也可以从碱基序列的其他位点开始，对碱基序列进行分割。如，从碱基序列的中间位点开始，同时朝两端依次对碱基序列进行分割，得到长度为n的整数倍的序列单元。

[0084] 在一些实施例中，将每个序列单元分割成多个长度为n的序列片段，包括：

[0085] 从序列单元的一端开始，每隔长度n对序列单元进行一次分割，得到多个序列片段。在其他实施例中，也可以从序列单元的其他位点开始，对序列单元进行分割。如，从序列单元的中间位点开始，同时朝两端依次对序列单元进行分割，得到长度为1的序列片段。

[0086] 示例性的，如图3所示，S101中的碱基序列中，按照序列单元长度为60个碱基的标准，从碱基序列的一端开始对碱基序列依次进行分割，得到3组长度为60个碱基的序列单元，以及一组长度为12个碱基的序列单元；按照序列片段长度为4个碱基的标准，从序列单元的一端开始对序列单元依次进行分割，三组长度为60个碱基的序列单元分别分割成15组序列片段，长度为12个碱基的序列单元分割成3组序列片段。

[0087] 在另一种可能的实现方式中，对碱基序列进行分割，得到s个序列单元，每个序列单元包括长度为n的多个序列片段，包括：

[0088] 将碱基序列分割为多个长度为n的序列片段；

[0089] 按照预设的组合规则将序列片段进行组合，得到s个序列单元。

[0090] 在一些实施例中，将碱基序列分割为多个长度为n的序列片段，从碱基序列的一端开始，每隔长度n对碱基序列进行一次分割，得到多个序列片段。

[0091] 在一些实施例中，按照预设的组合规则将序列片段进行组合时，预设的组合规则是指将多个序列片段归属为一个序列单元的规则，包括归属为一个序列单元的序列片段在

碱基序列中的位置，序列片段的数量以及序列片段组合成序列单元时序列片段的排布顺序。组合成的序列单元的长度可以相同，也可以不同，但均为1的整数倍。示例性的，按序列片段在碱基序列中的顺序，依次将20个碱基数量为6的序列片段进行组合，形成一个序列单元。

[0092] 示例性的，序列片段的长度为4，序列单元包括15个序列片段。此时，步骤S101的碱基序列“ATCT TCTG TTCC ATAA TAAG TGAG ATCC TACA TAGT ATCG TATG TGAC ATCC TGCT TGAC ATCT TCTG TTAA ATCT TGGA TAGA ATTG TAGC TTGC ATCG TATC TGTA ATCC TTCG TCCT ATCA TCTT TGCG ATTC TGGT TTGA ATCC TGAA TTTT ATTG TCAC TACG ATTG TCAC TACT ATGA TGGG TGGT”分割得到48个序列片段，分别为：ATCT、TCTG、TTCC、ATAA、TAAG、TGAG、ATCC、TACA、TAGT、ATCG、TATG、TGAC、ATCC、TGCT、TGAC、ATCT、TCTG、TTAA、ATCT、TGGA、TAGA、ATTG、TAGC、TTGC、ATCG、TATC、TGTA、ATCC、TTCG、TCCT、ATCA、TCTT、TGCG、ATTC、TGGT、TTGA、ATCC、TGAA、TTTT、ATTG、TCAC、TACG、ATTG、TCAC、TACT、ATGA、TGGG、TGGT；根据设定预设的组合规则和序列单元的长度，依次将各序列片段进行组合，且每15个序列片段组合成一个序列单元，得到四个序列单元，分别包含如下序列片段。第一个序列单元包括如下15个序列片段：ATCT、TCTG、TTCC、ATAA、TAAG、TGAG、ATCC、TACA、TAGT、ATCG、TATG、TGAC、ATCC、TGCT、TGAC；第二个序列单元包括如下15个序列片段：ATCT、TCTG、TTAA、ATCT、TGGA、TAGA、ATTG、TAGC、TTGC、ATCG、TATC、TGTA、ATCC、TTCG、TCCT；第三个序列单元包括如下15个序列片段：ATCA、TCTT、TGCG、ATTC、TGGT、TTGA、ATCC、TGAA、TTTT、ATTG、TCAC、TACG、ATTG、TCAC、TACT；第四个序列单元包括如下3个序列片段：ATGA、TGGG、TGGT。

[0093] 本申请实施例中，对碱基序列进行分割后，得到S个序列单元，S个序列单元共包含K个长度为n的序列片段。

[0094] S103.利用预设的索引信息对K个序列片段和S个序列单元进行标记，得到K个标记序列片段和S个标记序列单元，其中，索引信息包括用于表示S个序列单元在所述碱基序列中的排列顺序的第一检索序列，和用于表示属于同一序列单元中的多个序列片段在序列单元中的排列顺序的第二检索序列，K个标记序列片段用于合成存储有目标数据的K个第一DNA分子。

[0095] 本申请实施例中，采用预设的索引信息对序列单元进行标记，以方便后续DNA存储数据的解码。采用预设的索引信息对序列单元进行标记，包括：利用第一检索序列对S个序列单元在所述碱基序列中的排列顺序进行标记，以及利用第二检索序列对属于同一序列单元中的多个序列片段在序列单元中的排列顺序进行标记，得到K个标记序列片段和S个标记序列单元。在这种情况下，碱基序列中的S个序列单元的顺序被记录，同时，属于同一序列单元中的多个序列片段的顺序也被记录下来。

[0096] 本申请实施例中，第二检索序列为碱基形成的序列，可通过预先设定的规则来确定第二检索序列的碱基序列。示例性的，当序列单元中的序列片段的数量小于或等于4时，可以采用单碱基来表示第二检索序列。如：序号1对应双碱基A，序号2对应双碱基C，序号3对应双碱基G，序号4对应双碱基T，当然，序号和碱基之间不限于这种对应方式。例性的，当序列单元中的序列片段的数量小于或等于16时，可以采用双碱基来表示第二检索序列。如：序号1对应双碱基AA，序号2对应双碱基AC，序号3对应双碱基AG，序号4对应双碱基AT，序号5对应双碱基CA，序号6对应双碱基CC，序号7对应双碱基CG，序号8对应双碱基CT，序号9对应双

碱基GA,序号10对应双碱基GC,序号11对应双碱基GG,序号12对应双碱基GT,序号13对应双碱基TA,序号14对应双碱基TC,序号15对应双碱基TG,序号16对应双碱基TT....当然,基准碱基组中的碱基数量并不限于2,当序列单元中的序列片段的数量增加时,采用的第二检索序列中的碱基数量对应增加,如当序列单元中的序列片段的数量小于或等于64时,可以采用三碱基来表示第二检索序列。按照第二检索序列中4的碱基数量次方大于或等于序列单元中的序列片段的数量的规则,以此类推。

[0097] 本申请实施例中,采用第二检索序列标记序列片段时,可以在序列片段的特定位置拼接第二检索序列。在一些实施例中,利用第二检索序列标记属于同一序列单元中的多个序列片段的方式,包括:在序列片段的任一侧拼接第二检索序列。示例性的,在序列片段的起始端即左端拼接第二检索序列;或,在序列片段的终止端即右端拼接第二检索序列。在一个实施例中,利用第二检索序列标记属于同一序列单元中的多个序列片段的方式,包括:在序列片段的两侧同时拼接碱基组,两侧的碱基组形成第二检索序列。

[0098] 本申请实施例中,K个序列片段经过第二检索序列标记,形成K个标记序列片段,又称为信息序列片段。示例性的,序列片段ATGC前标记第二检索序列AA后,形成AAATGC的标记序列片段。

[0099] 本申请实施例中,通过采用第一检索序列用来标示S个序列单元在碱基序列中的位置。在一些实施例中,第一检索序列包括i条DNA序列片段,i为大于或等于1的整数,即第一检索序列可以是一条DNA序列片段,也可以是多条DNA序列片段。其中,每条DNA序列片段包括用作索引标志的第一碱基序列和用于标示序列单元编号的第二碱基序列。其中,第一碱基序列可以根据预先设置置于第一检索序列的特定位置中,示例性的,第一碱基序列位于第一检索序列的起始段(左端);示例性的,第一碱基序列位于第一检索序列的终止段(右端);示例性的,第一碱基序列位于第一检索序列中特定的位置,如第一检索序列的第三位和第四位碱基,不限于此。

[0100] 第一碱基序列可以预先设定。示例性的,TT作为第一碱基序列置于第一检索序列的起始段,用作索引标志,表示以序列单元中以TT开头的DNA序列片段第一检索序列。该示例中,当第一碱基序列位于第一检索序列的起始段时,第一碱基序列与第二检索序列的起始碱基序列不同,以避免识别过程中,误将序列片段识别为第一检索序列。

[0101] 同样的,第二碱基序列也可以预先设定。示例性的,序号1对应四碱基AAAA,序号2对应四碱基AAAC,序号3对应四碱基AAAG,序号4对应四碱基AAAT,当然,序号和碱基之间不限于这种对应方式,基准碱基组中的碱基数量也不限于4。本申请实施例中,序列单元经过第一检索序列标记,且序列单元中的各序列片段经过第二检索序列标记,得到标记后的序列单元,又称为形成信息序列单元。参考图4,将步骤S102中的序列单元(图4箭头左侧所示)经过第一检索序列标记,且序列单元中的各序列片段经过第二检索序列标记后,得到分别包含多个信息序列片段的四组标记序列单元(图4箭头右侧所示)。

[0102] 在一些实施例中,第一检索序列和第二检索序列对应的DNA序列片段利用DNA合成技术获得。示例性的,DNA合成技术包括但不限于酶法合成、亚磷酰胺合成等。

[0103] 在一些实施例中,第一检索序列和第二检索序列对应的DNA序列片段可以从预先合成的DNA通用分子库中扩增获得,比如PCR技术等。

[0104] 上述步骤S101至步骤S103通过图1所示装置中的处理模块实现。

[0105] 在一些实施例中,存储方法还包括:

[0106] S104.将K个标记序列片段合成存储有所述目标数据的K个第一DNA分子后,将K个第一DNA分子存储在S个第一物理空间,其中,同属于一个序列单元的标记序列片段对应的第一DNA分子存储在同一个第一物理空间,不属于同一个序列单元的标记序列片段对应的第一DNA分子存储在不同的第一物理空间。

[0107] 该步骤中,将K个标记序列片段分别合成存储有所述目标数据的K个第一DNA分子,通过图1所示装置中的DNA合成模块实现。本申请实施例可以通过现有的合成技术,将K个标记序列片段分别合成,得到K个第一DNA分子。

[0108] 将K个第一DNA分子存储在S个不同的第一物理空间,并使得同属于一个序列单元的标记序列片段对应的第一DNA分子存储在同一个第一物理空间,不属于同一个序列单元的标记序列片段对应的第一DNA分子存储在不同的第一物理空间,该步骤通过图1所示装置中的DNA存储模块实现。将K个第一DNA分子存储在S个不同的第一物理空间后的DNA存储的示意图如图5所示,其中,每一个小方格表示一个第一物理空间。

[0109] 该步骤中,将合成的K个第一DNA分子分别存储于不同的第一物理空间中,实现各信息序列单元的分别存储。进一步的,将S个所述第一物理空间集成在一个DNA硬盘中,实现存储有目标数据的碱基序列的存储。通过集成,S个第一物理空间中的K个第一DNA分子形成一个完整的整体得以保存,且保存过程中不容易出现遗漏而损失信息,从而有利于提高数据保存的完整性。对应的,在解码释放过程中,通过将集成的第一DNA分子解码释放,可以完整地恢复DNA碱基,保持数据的完整性。

[0110] 在一些实施例中,存储方法还包括:将第二DNA分子存储在与第一物理空间对应的第二物理空间,第二DNA分子存储有索引信息。将存储有索引信息的第二DNA分子存储于与第一物理空间不同的第二物理空间,实现索引信息的保存。

[0111] 本申请实施例提供的DNA数据的存储方法,将待存储数据对应的二进制序列转换为碱基序列后,与目标数据的二进制序列对应的碱基序列分割为S个序列,每个序列单元包括多个分割的序列片段,S个所述序列单元共含有K个所述序列片段,且所述序列片段的长度为n,n、S和K均为大于或者等于2的整数,利用预设的索引信息标记序列片段和序列单元的位置信息,将标记后的序列片段合成为DNA分子后分别存储。通过该方法,可以提升数据信息的存储量,实现大规模的数据信息在DNA中的存储。

[0112] 在某些具体实施例中,第一检索序列的长度和标记序列片段(带有第二检索序列的序列片段)的长度不同,则通过长度区别从序列单元中区分第一检索序列和标记序列片段。具体的,以m表示标记序列片段的碱基数,以q表示第二检索序列的碱基数,以i表示第一检索序列中DNA序列片段的条数,以p表示第一检索序列中第二碱基序列的碱基数,通过本申请提供的方法,可以实现含有D个碱基数的DNA数据的存储,其中,D的计算公式如下:

[0113] $D=4^{q}\times(m-q)\times4^{i\times p}$

[0114] 在特定实施例中,m长度为8碱基的情况下,q=4,i=10,p=4,$D=256\times4\times4^{40}=1.23\times10^{27}$。1个碱基存储2bits信息时候,能够存储的信息量$L=2.46\times10^{27}$bits$=3.075\times10^{26}$bytes$=3.075\times10^{5}$ZB,远大于目前的数据存储规模。在某些具体实施例中,第一检索序列的长度和标记序列片段(带有第二检索序列的序列片段)的长度相同;第一检索序列中用作索引标志的第一碱基序列,可以是第二检索序列的一部分,和第二检索序列的碱基数相

同。在这种情况下，以m表示标记序列片段的碱基数，以q表示第二检索序列的碱基数，以i表示第一检索序列中DNA序列片段的条数，以p表示第一检索序列中第二碱基序列的碱基数，通过本申请的方法，可以实现含有D个碱基数的DNA数据的存储，其中，D的计算公式如下：

[0115] $D=(4^{q}-i)\times(m-q)\times4^{i\times p}$

[0116] 在特定实施例中，m长度为8碱基的情况下，q＝4，i＝10，p＝4，D＝(256-10)×4×4^{40}＝1.18×10^{27}。1个碱基存储2bits信息时候，能够存储的信息量L＝2.36×10^{27}bits＝2.95×10^{26}bytes＝2.95×10^{5}ZB，远大于目前的数据存储规模。

[0117] 即使序列片段中含有8个碱基，也能够实现远大于目前存储量的数据存储。此外，由于每个序列单元及序列片段都采用检索序列进行了标记，因此该方法能够成功实现从序列片段到序列单元，以及从序列单元到碱基序列的数据恢复。

[0118] 此外，本申请实施例通过将第一DNA分子分别存储，并根据实际需要进行扩增保存。在这种情况下，将第一DNA分子纳入DNA分子库中后，可以根据实际需要来提取第一DNA分子的备份信息，从而免去每次重新从头合成的麻烦，可以大幅降低存储成本。

[0119] 在一个实施例中，如图6所示，K个所述第一DNA分子的解码方法，包括：

[0120] S201.对每个第一物理空间中存储的多个第一DNA分子进行测序，得到多个标记序列片段；根据索引信息对属于同一标记序列单元的每个标记序列片段对应的序列片段进行拼接，得到序列单元。

[0121] 对K个第一DNA分子进行测序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，获取待解码的多个信息序列单元。

[0122] 在一个实施例中，对K个第一DNA分子进行测序，分别得到K个标记序列片段。在一些实施例中，当第二DNA分子存储在与第一物理空间对应的第二物理空间，第二DNA分子存储有索引信息时，测序还包括：对第二DNA分子进行测序，获取包括第一检索序列和第二检索序列的索引信息。

[0123] 上述步骤S201，可以通过图1所示存储装置中的DNA测序模块实现。

[0124] S202.根据第二检索序列对属于同一标记序列单元的每个标记序列片段对应的序列片段进行拼接，得到S个序列单元；根据第一检索序列将得到的S个序列单元进行拼接，得到碱基序列。

[0125] 该步骤通过检索信息将分属于S个序列单元中的K个检索片段拼接成碱基序列。本申请实施例根据第一检索序列获取标记序列单元在碱基序列中的位置信息；根据第二检索序列获取属于同一序列单元的多个标记序列片段的位置信息或编号信息。根据得到的多个序列片段的位置信息或编号信息，将多个序列片段拼接成为序列单元；根据得到的序列单元的位置信息或编号信息，S个序列单元拼接为碱基序列。

[0126] 在一些实施例中，根据第二检索序列对属于同一标记序列单元的每个标记序列片段对应的序列片段进行拼接，得到S个序列单元，包括：根据第二检索序列获取属于同一标记序列单元的每个标记序列片段对应的序列片段，以及序列片段在序列单元中的位置；根据序列片段在序列单元中的位置，将序列片段拼接成序列单元。

[0127] 在一个实施例中，根据第一检索序列将得到的S个序列单元进行拼接，得到碱基序列，包括：

[0128] 根据第一检索序列获取序列单元在碱基序列中的位置；

[0129] 根据序列单元在碱基序列中的位置,将s个序列单元拼接成碱基序列。

[0130] 示例性的,根据第一检索序列和第二检索序列,解读出完整的DNA序列“ATCT TCTG TTCC ATAA TAAG TGAG ATCC TACA TAGT ATCG TATG TGAC ATCC TGCT TGAC ATCT TCTG TTAA ATCT TGGA TAGA ATTG TAGC TTGC ATCG TATC TGTA ATCC TTCG TCCT ATCA TCTT TGCG ATTC TGGT TTGA ATCC TGAA TTTT ATTG TCAC TACG ATTG TCAC TACT ATGA TGGG TGGT”。

[0131] S203.将碱基序列转换为目标数据。

[0132] 可以预先设定的与S101的数据写入时匹配的映射关系,将碱基序列转换为二进制序列。比如,按照A对应11,T对应10,C对应01,G对应00,可以将S102得到的碱基序列转换为二进制序列为:“11100110 10011000 10100101 11101111 10111100 10001100 11100101 10110111 10110010 11100100 10111000 10001101 11100101 10000110 10001101 11100110 10011000 10101111 11100110 10000011 10110011 11101000 10110001 10100001 11100100 10111001 10001011 11100101 10100100 10010110 11100111 10011010 10000100 11101001 10000010 10100011 11100101 10001111 10101010 11101000 10011101 10110100 11101000 10011101 10110110 11100011 10000000 10000010”。

[0133] 然后根据二进制序列生成计算机信息序列。

[0134] 根据所生成的二进制序列,结合预先所设定的编码规则,可以将二进制序列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。

[0135] 如利用计算机程序将上述步骤S203得到的二进制序列恢复成“春,已不再是想象之外的那只蝴蝶。”的文本信息。

[0136] 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。

[0137] 第二方面,参考图1,本申请实施例提供了一种DNA数据存储装置,包括数据处理模块,数据处理模块,用于获取与目标数据的二进制序列对应的碱基序列;对碱基序列进行分割,得到S个序列单元,且每个序列单元包括多个分割的序列片段,其中,S个序列单元共含有K个序列片段,且序列片段的长度为n,n、S和K均为大于或者等于2的整数;利用预设的索引信息对K个序列片段和S个序列单元进行标记,得到K个标记序列片段和S个标记序列单元,其中,索引信息包括用于表示S个所述序列单元在碱基序列中的排列顺序的第一检索序列,和用于表示属于同一序列单元中的多个序列片段在序列单元中的排列顺序的第二检索序列,K个标记序列片段用于合成存储有目标数据的K个第一DNA分子。

[0138] 在一些实施例中,数据处理模块,还用于根据第二检索序列对属于同一标记序列单元的每个标记序列片段对应的序列片段进行拼接,得到序列单元;根据第一检索序列将得到的S个序列单元进行拼接,得到碱基序列;将碱基序列转换为目标数据。

[0139] 在一些实施例中,存储装置还包括DNA合成模块,用于将K个标记序列片段合成存储有目标数据的K个第一DNA分子。在一些实施例中,存储装置还包括DNA分子存储模块,用于将K个第一DNA分子存储在S个第一物理空间,其中,同属于一个序列单元的标记序列片段对应的第一DNA分子存储在同一个第一物理空间,不属于同一个序列单元的标记序列片段

对应的第一DNA分子存储在不同的第一物理空间。

[0140] 在一些实施例中，DNA分子存储模块还用于将第二DNA分子存储在第二物理空间。

[0141] 在一些实施例中，存储装置还包括DNA分子测序模块，对每个第一物理空间中存储的多个第一DNA分子进行测序，得到多个标记序列片段；

[0142] 图1含有上述模块的装置能利用DNA进行数据存储写入，与图2所示的DNA数据存储写入的方法对应。

[0143] 本申请实施例还提供了一种DNA数据存储设备。如图7所示，本实施例提供的终端设备70包括：处理器710、存储器720以及存储在存储器720中并可在处理器710上运行的计算机程序721。处理器710执行计算机程序721时实现上述DNA数据的存储方法各个实施例中的步骤，例如图2所示的步骤S101至S103。

[0144] 示例性的，计算机程序721可以被分割成一个或多个模块/单元，一个或者多个模块/单元被存储在存储器720中，并由处理器710执行，以完成本申请。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，该指令段可以用于描述计算机程序721在终端设备中的执行过程。例如，计算机程序721可以被分割成数据处理模块。在一些实施例中，计算机程序721还可以被分割成数据处理模块、DNA分子合成模块、DNA分子存储模块和DNA分子存储模块，各模块具体功能如文所述。为了节约篇幅，此处不再赘述。

[0145] 本领域技术人员可以理解，图7仅仅是终端设备70的一种示例，并不构成对终端设备70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。

[0146] 处理器710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。

[0147] 存储器720可以是终端设备70的内部存储单元，例如终端设备70的硬盘或内存。存储器720也可以是终端设备70的外部存储设备，例如终端设备70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)等等。进一步地，存储器720还可以既包括终端设备70的内部存储单元也包括外部存储设备。存储器720用于存储计算机程序721以及终端设备70所需的其他程序和数据。存储器720还可以用于暂时地存储已经输出或者将要输出的数据。

[0148] 本申请实施例还提供了一种计算机可读存储介质，计算机可读存储介质存储有计算机程序，计算机程序被处理器执行时实现前述各实施例的处理方法。

[0149] 本申请实施例还提供了一种计算机程序产品，当计算机程序产品在终端设备上运行时，使得终端设备执行前述各实施例的DNA数据的存储方法。

[0150] 本申请实施例提供一种DNA硬盘，参考图5，包括多个物理空间，物理空间由物理材料制成，每个物理空间用于存储DNA分子。

[0151] 本申请实施例中，物理材料制成的物理空间包裹DNA分子，DNA分子通过物理材料隔离。其中，物理空间的形状没有严格限定，可以是圆形的、方形的，还可以是其他任意形状。

[0152] 在一些实施例中，物理空间中存储的DNA分子包括上文中的第一DNA分子，此时，物理空间为第一物理空间。对应的，第一DNA分子包括第二检索序列和序列片段。每一个第一物理空间还存储有上文所述的第二检索序列。

[0153] 在一些实施例中，物理空间中存储的DNA分子包括上文所述的第二DNA分子，此时，物理空间为第二物理空间。

[0154] 在一些实施例中，物理材料选自SiO_2、金属氧化物、高分子聚合材料中的至少一种。其中，高分子聚合材料包括树脂，但不限于树脂。

[0155] 以上仅为本申请的可选实施例而已，并不用于限制本申请。对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

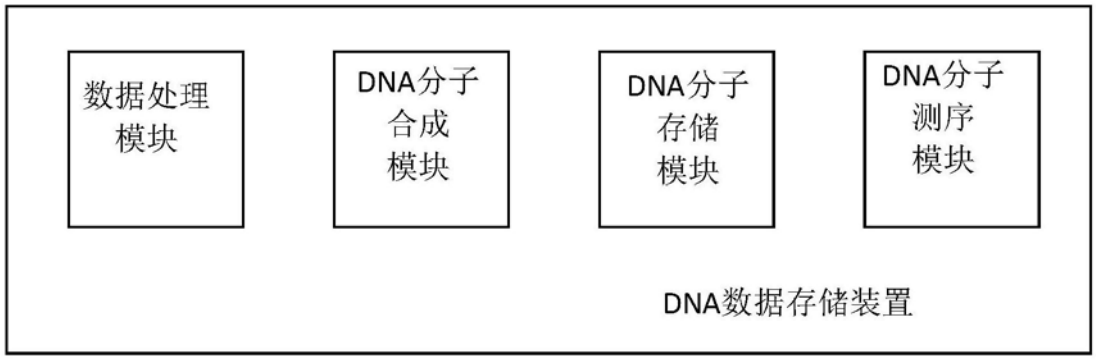

数据处理模块
DNA分子合成模块
DNA分子存储模块
DNA分子测序模块
DNA数据存储装置

图1

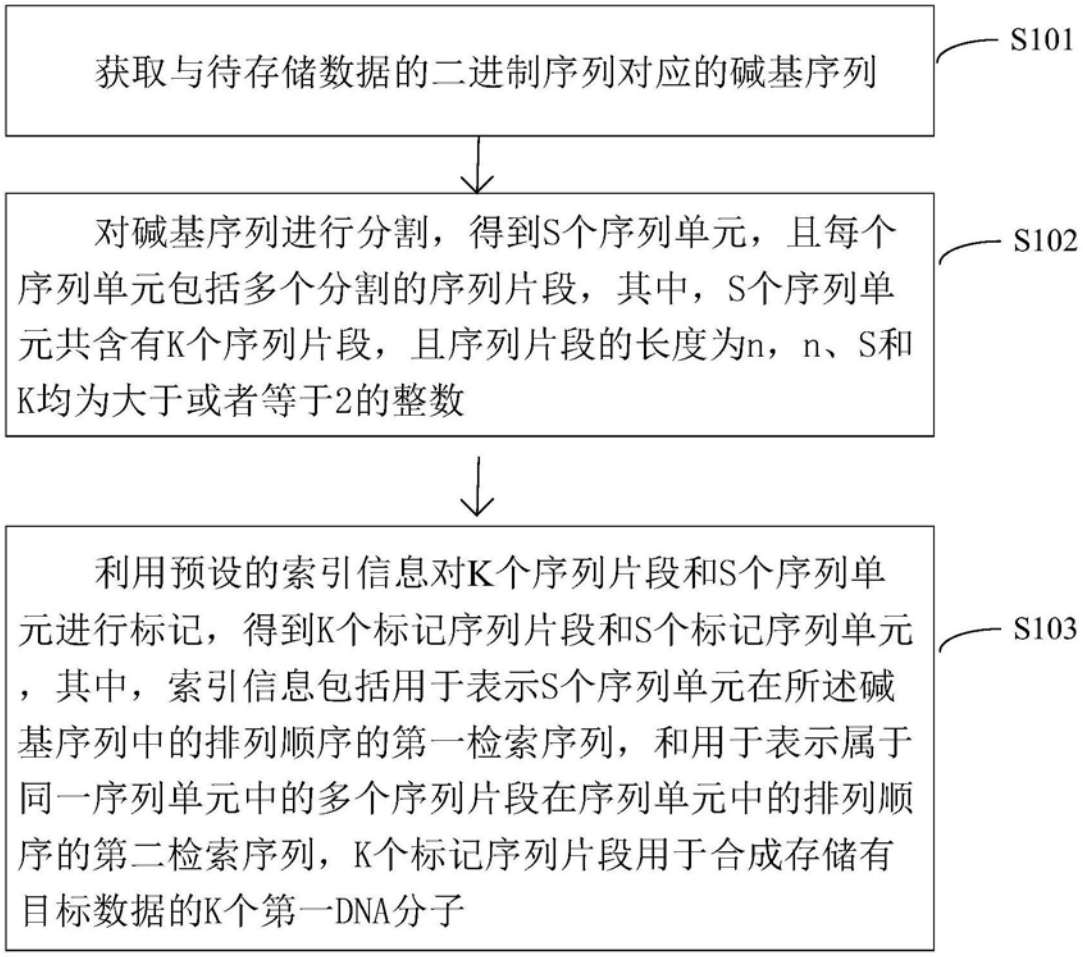

获取与待存储数据的二进制序列对应的碱基序列
S101
对碱基序列进行分割，得到S个序列单元，且每个序列单元包括多个分割的序列片段，其中，S个序列单元共含有K个序列片段，且序列片段的长度为n，n、S和K均为大于或者等于2的整数
S102
利用预设的索引信息对K个序列片段和S个序列单元进行标记，得到K个标记序列片段和S个标记序列单元，其中，索引信息包括用于表示S个序列单元在所述碱基序列中的排列顺序的第一检索序列，和用于表示属于同一序列单元中的多个序列片段在序列单元中的排列顺序的第二检索序列，K个标记序列片段用于合成存储有目标数据的K个第一DNA分子
S103

图2

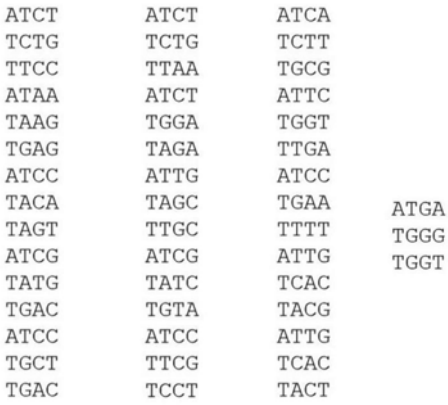

ATCT TCTG TTCC ATAA TAAG TGAG ATCC TACA TAGT ATCG TATG TGAC ATCC TGCT TGAC
ATCT TCTG TTAA ATCT TGGA TAGA ATTG TAGC TTGC ATCG TATC TGTA ATCC TTCG TCCT
ATCA TCTT TGCG ATTC TGGT TTGA ATCC TGAA TTTT ATTG TCAC TACG ATTG TCAC TACT
ATGA TGGG TGGT

图3

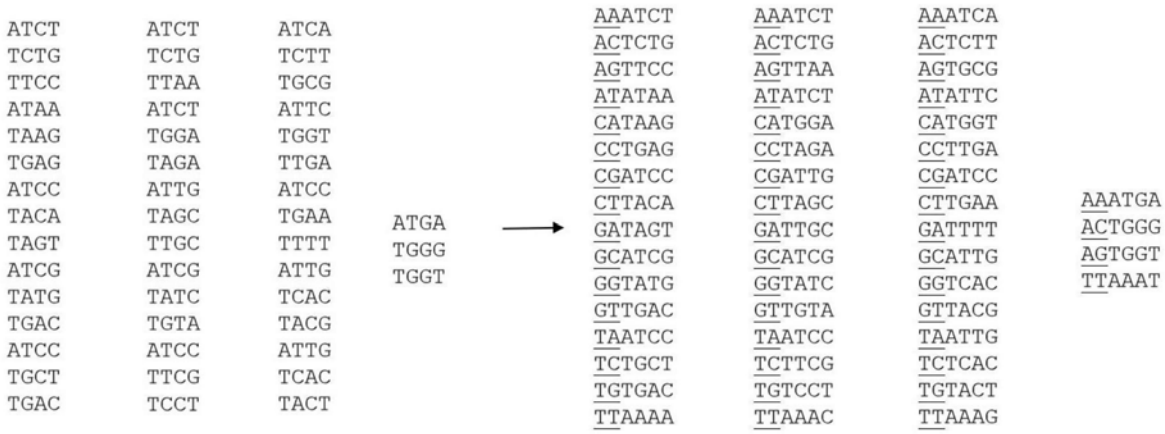

ATCT TCTG TTCC ATAA TAAG TGAG ATCC TACA TAGT ATCG TATG TGAC ATCC TGCT TGAC
ATCT TCTG TTAA ATCT TGGA TAGA ATTG TAGC TTGC ATCG TATC TGTA ATCC TTCG TCCT
ATCA TCTT TGCG ATTC TGGT TTGA ATCC TGAA TTTT ATTG TCAC TACG ATTG TCAC TACT
ATGA TGGG TGGT
AAATCT ACTCTG AGTTCC ATATAA CATAAG CCTGAG CGATCC CTTACA GATAGT GCATCG GGTATG GTTGAC TAATCC TCTGCT TGTGAC TTAAAA
AAATCT ACTCTG AGTTAA ATATCT CATGGA CCTAGA CGATTG CTTAGC GATTGC GCATCG GGTATC GTTGTA TAATCC TCTTCG TGTCCT TTAAAC
AAATCA ACTCTT AGTGCG ATATTC CATGGT CCTTGA CGATCC CTTGAA GATTTT GCATTG GGTCAC GTTACG TAATTG TCTCAC TGTACT TTAAAG
AAATGA ACTGGG AGTGGT TTAAAT

图4

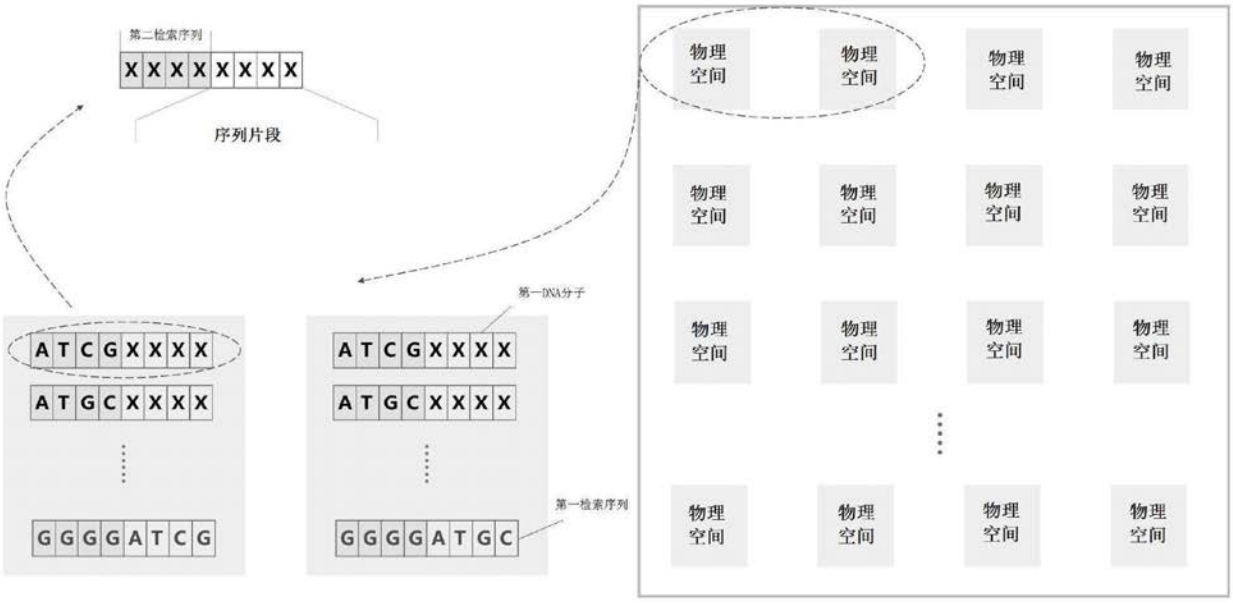

第二检索序列
X X X X X X X X
序列片段
第一DNA分子
A T C G X X X X
A T G C X X X X
G G G G A T C G
A T C G X X X X
A T G C X X X X
G G G G A T G C
第一检索序列
物理空间

图5

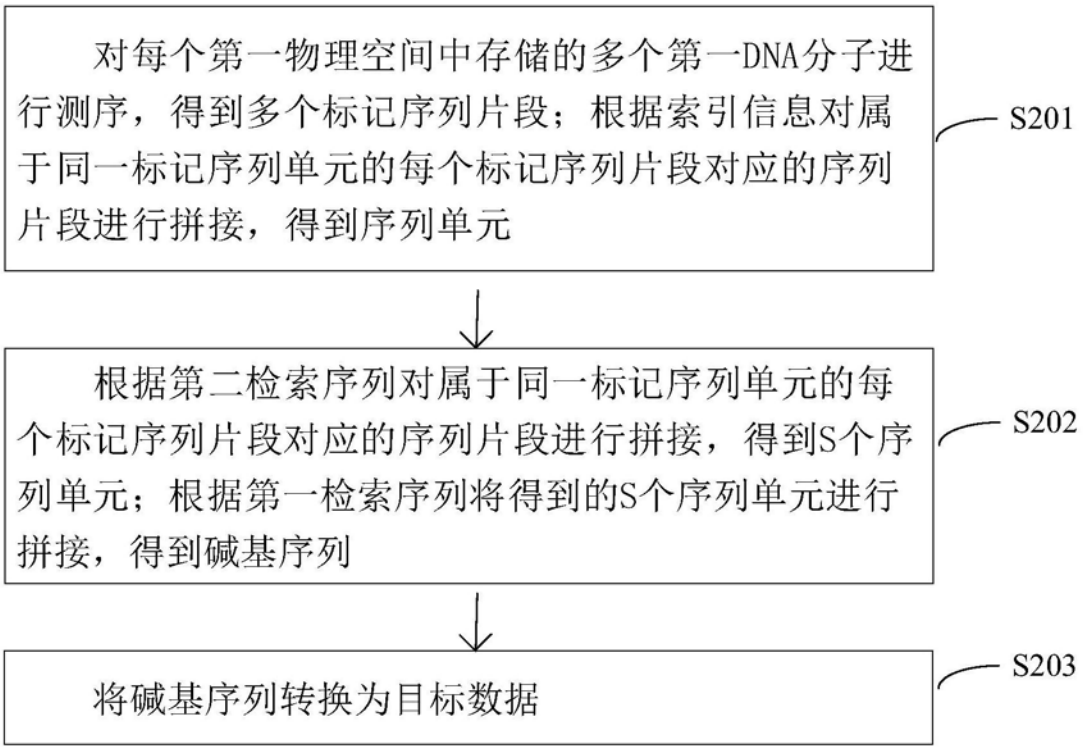

对每个第一物理空间中存储的多个第一DNA分子进行测序，得到多个标记序列片段；根据索引信息对属于同一标记序列单元的每个标记序列片段对应的序列片段进行拼接，得到序列单元
S201
根据第二检索序列对属于同一标记序列单元的每个标记序列片段对应的序列片段进行拼接，得到S个序列单元；根据第一检索序列将得到的S个序列单元进行拼接，得到碱基序列
S202
将碱基序列转换为目标数据
S203

图6

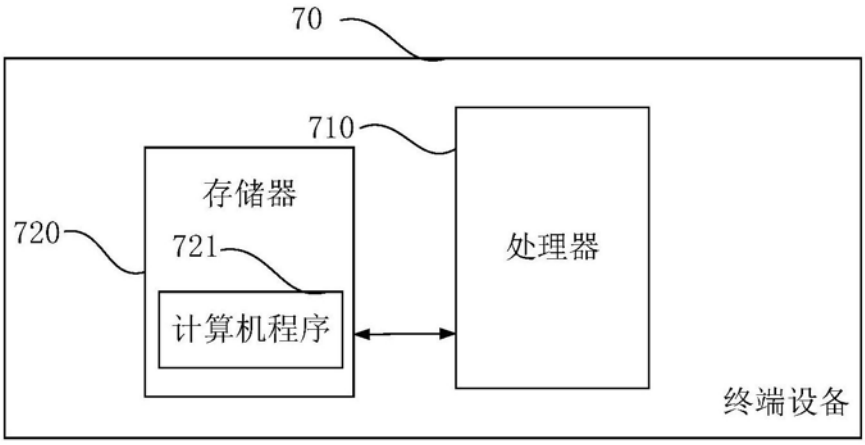


图7