(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 114023392 A
(43)申请公布日 2022.02.08

(21)申请号 202111301348.0

(22)申请日 2021.11.04

(71)申请人 大连大学
地址 116622 辽宁省大连市经济技术开发区学府大街10号

(72)发明人 王宾 郑燕芬 胡轶男 张强

(74)专利代理机构 大连智高专利事务所(特殊普通合伙) 21235
代理人 毕进

(51)Int.Cl.
G16B 50/50(2019.01)
G16B 50/00(2019.01)

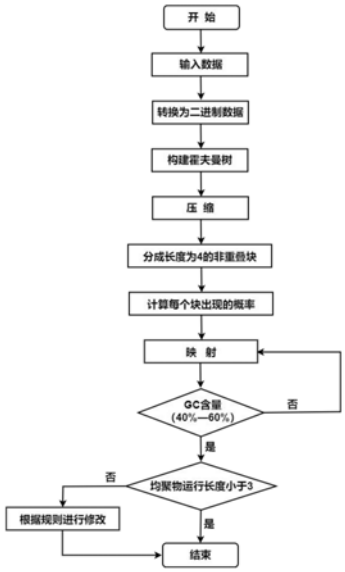
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种DNA存储的码字设计方法

(57)摘要

本发明公开了一种DNA存储的码字设计方法,其具体为:将存储信息转换为DNA序列,首先要将信息转换为二进制数据。其次,构建最小方差霍夫曼树,利用它对二进制数据进行压缩。然后,将压缩后的二进制数据以4位为一组进行不重叠分块,得到至多16种组合,根据组合的概率依次从字典中选择码字进行映射,得到DNA序列。最后,求得DNA序列的GC含量,如果GC含量高于60%或者低于40%,会对映射关系进行调整,使得它在40%到60%之间;再进一步检查DNA序列中是否含有均聚物超过3的情况,如果存在就进行替换修改。本发明不仅具有高的编码率和结构简单的特点,而且编码完成的DNA序列还满足GC含量在40%到60%之间和均聚物运行长度不超过3的约束条件。



CN 114023392 A

1.一种DNA存储的码字设计方法,其特征在于,包括:

步骤1:获取输入数据并转换为二进制数据;

步骤2:将所述二进制数据进行压缩;

步骤3:压缩后的所述二进制数据分成长度为4的不重叠块,所述不重叠块至多有16种组合;

步骤4:获取不重叠块出现的概率;

步骤5:根据字典中的码字对所述不重叠块进行编码:

步骤6:获取编码后DNA序列的GC含量,如果GC含量低于40%或者高于60%,则进行步骤5,改变映射关系继续编码,反之则进行步骤7;

步骤7:寻找均聚物运行长度不小于3的情况,如果有,则进行替换修改,如果没有执行步骤8;

步骤8:获取编码率,并且将DNA序列输出。

2.根据权利要求1所述一种DNA存储的码字设计方法,其特征在于,如果输入数据为文本数据,则将每个字符转换为ACSII码;如果输入数据为图像数据,则将图像转换为像素值;然后再将所述ACSII码、像素值转换为8位的二进制数据。

3.根据权利要求1所述一种DNA存储的码字设计方法,其特征在于,使用最小方差霍夫曼树进行二进制数据压缩。

4.根据权利要求3所述一种DNA存储的码字设计方法,其特征在于,所述霍夫曼树包含n个叶节点,对应n个源符号,其出现的概率是p_i,n个叶节点到根节点的距离l_i的方差σ^2,是通过下面的公式获得:

$$\sigma^2=\sum_{i=1}^{n} p_i(l_i-\bar{R})^2 \qquad (1)$$

其中,$\bar{R}$是每个码字的平均长度,通过下面公式获得:

$$\bar{R}=\sum_{i=1}^{n} p_i l_i \qquad (2)$$

在所有可能的霍夫曼树中,有最小方差的树被称为最小方差霍夫曼树。

5.根据权利要求1所述一种DNA存储的码字设计方法,其特征在于,根据字典中的码字来对不重叠块进行编码,具体为:

字典C_1={C,A}

字典C_2={TA,TG,TC}

字典C_3={GAT,GTA,GAC,GAG,GTC,GTG,GCA,GCT,GGT,GGA,GCG}。

6.根据权利要求1所述一种DNA存储的码字设计方法,其特征在于,GC含量指的是在一条DNA序列中,碱基G和碱基C所占整个DNA序列中碱基的百分比,如下公式所示:

$$GC(s)=\frac{|G|+|C|}{|s|}\times 100\% \qquad (3)$$

其中,GC(s)表示序列s的GC含量,|G|和|C|表示序列s中碱基G和碱基C的个数,|s|表示序列s全部的碱基个数。

7.根据权利要求1所述一种DNA存储的码字设计方法,其特征在于,寻找均聚物运行长度不小于3的情况,具体为:寻找DNA序列中相同的碱基连续出现3次及以上的情况。

8.根据权利要求1所述一种DNA存储的码字设计方法，其特征在于，所述编码率表示的是多少位原始数据能代表一个DNA核苷酸，其利用公式(4)获得：

$$R=\frac{b}{n} \qquad (4)$$

其中，b表示的是原始数据的比特数，n表示的是存储相同数据的DNA核苷酸的总数。

一种DNA存储的码字设计方法

技术领域

[0001] 本发明涉及编码设计技术领域,具体涉及一种DNA存储的码字设计方法。

背景技术

[0002] 目前,全球对数据存储的需求超过了全球存储能力的增长速度。DNA作为自然遗传信息的载体,提供了一种稳定、资源高效、可持续的数据存储解决方案。直到21世纪的头十年,Church和Goldman的开创性工作才使DNA存储成为主流。Church等人成功地在DNA分子中存储了高达659KB的数据,而在这项工作之前,最大的存储数据量小于1KB。Goldman等人存储的数据更多,达到了739KB。值得注意的是,这两项研究中存储的数据不仅包含文本,还包含图像、声音、pdf等,这证实了DNA可以存储多种数据类型。

[0003] 具体来说,DNA数据存储是一项新兴的研究,即将二进制数字信息转化为DNA序列,以合成DNA的形式进行密集而持久的数据存储。但是目前DNA编码方法仅仅简单地将二进制数据映射成DNA序列,存在编码率低、合成成本高的缺点。

发明内容

[0004] 针对现有技术存在上述问题,本申请提出了一种结构简单和高编码率的码字设计方法,其编码得到的序列还满足GC含量在40%-60%之间和均聚物运行长度不超过3的约束条件。

[0005] 为实现上述目的,本申请的技术方案为:一种DNA存储的码字设计方法,包括:

[0006] 步骤1:获取输入数据并转换为二进制数据;

[0007] 步骤2:将所述二进制数据进行压缩;

[0008] 步骤3:压缩后的所述二进制数据分成长度为4的不重叠块,所述不重叠块至多有16种组合;

[0009] 步骤4:获取不重叠块出现的概率;

[0010] 步骤5:根据字典中的码字对所述不重叠块进行编码:

[0011] 步骤6:获取编码后DNA序列的GC含量,如果GC含量低于40%或者高于60%,则进行步骤5,改变映射关系继续编码,反之则进行步骤7;

[0012] 步骤7:寻找均聚物运行长度不小于3的情况,如果有,则进行替换修改,如果没有执行步骤8;

[0013] 步骤8:获取编码率,并且将DNA序列输出。

[0014] 进一步的,如果输入数据为文本数据,则将每个字符转换为ACSII码;如果输入数据为图像数据,则将图像转换为像素值;然后再将所述ACSII码、像素值转换为8位的二进制数据。

[0015] 进一步的,使用最小方差霍夫曼树进行二进制数据压缩。

[0016] 进一步的,所述霍夫曼树包含n个叶节点,对应n个源符号,其出现的概率是p_i,n个叶节点到根节点的距离l_i的方差σ^2,是通过下面的公式获得:

[0017] $$\sigma^2=\sum_{i=1}^{n} p_i(l_i-\bar{R})^2 \quad (1)$$

[0018] 其中，$\bar{R}$是每个码字的平均长度，通过下面公式获得：

[0019] $$\bar{R}=\sum_{i=1}^{n} p_i l_i \quad (2)$$

[0020] 在所有可能的霍夫曼树中，有最小方差的树被称为最小方差霍夫曼树。

[0021] 进一步的，根据字典中的码字来对不重叠块进行编码，具体为：

[0022] 字典C_1＝{C，A}

[0023] 字典C_2＝{TA，TG，TC}

[0024] 字典C_3＝{GAT，GTA，GAC，GAG，GTC，GTG，GCA，GCT，GGT，GGA，GCG}。

[0025] 更进一步的，GC含量指的是在一条DNA序列中，碱基G和碱基C所占整个DNA序列中碱基的百分比，如下公式所示：

[0026] $$GC(s)=\frac{|G|+|C|}{|s|}\times 100\% \quad (3)$$

[0027] 其中，GC(s)表示序列s的GC含量，|G|和|C|表示序列s中碱基G和碱基C的个数，|s|表示序列s全部的碱基个数。

[0028] 更进一步的，寻找均聚物运行长度不小于3的情况，具体为：寻找DNA序列中相同的碱基连续出现3次及以上的情况。

[0029] 更进一步的，所述编码率表示的是多少位原始数据能代表一个DNA核苷酸，其利用公式(4)获得：

[0030] $$R=\frac{b}{n} \quad (4)$$

[0031] 其中，b表示的是原始数据的比特数，n表示的是存储相同数据的DNA核苷酸的总数。

[0032] 本发明由于采用以上技术方案，能够取得如下的技术效果：

[0033] 1、在编码之前本发明首先使用最小方差霍夫曼树进行数据压缩，这样可以有效地降低生物研究的成本，提高了整个DNA编码的编码率；

[0034] 2、在DNA存储过程中，不同的序列出现错误的概率是不一样的，对于存在长均聚物运行、高GC含量的DNA序列在DNA合成和测序的过程中出现错误的概率会明显的提高；因此本发明引入约束条件来限制此类DNA序列的出现，提高了DNA存储过程的可靠性和准确性。

[0035] 3、基于码字设计的DNA存储编码方法不仅能够存储文本还可以存储图像，而且具有结构简单，算法复杂度低的特点。

附图说明

[0036] 图1为一种DNA存储的码字设计方法实现流程图。

具体实施方式

[0037] 下面将结合本发明中的附图，对本发明实施中的技术方案进行清楚、完整的描述，可以理解的是，所描述的实施例仅是本发明的一部分实施例，而不是全部的实施例。基于本

发明的实施例,本领域的技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例,都属于本发明的保护范围。

[0038] 本发明中所涉及的约束条件有GC含量在40％到60％之间和均聚物运行长度不能超过3。其中,GC含量指的是在一条DNA序列中,碱基G和碱基C所占整个DNA序列中碱基的百分比;均聚物运行长度不能超过3表示在一个DNA序列中相同的碱基不能连续出现3次及以上。

[0039] 实施例1

[0040] 本发明的实施例是在以本发明技术方案为前提下进行实施的,给出了详细的实施方式和具体的操作过程,但本发明的保护范围不限于下述实施例。实施例中用该编码算法对一个大小为511B的文本文件进行编码,满足的约束条件如上所述。

[0041] 步骤1:获取输入数据并转换为二进制数据。

[0042] 具体的,首先利用abs函数将文本文件中的字符转换为ACSII码,再将其转换为8位的二进制数,511B的文本文件经过上述操作可以转化为4088bits的二进制数据。

[0043] 步骤2:将所述二进制数据进行压缩。

[0044] 需要说明的是,将每16位的二进制数作为一组,然后利用最小方差霍夫曼树进行压缩,其中霍夫曼树是通过源符号出现的概率构建。所有的源符号根据它们出现的概率进行排序,概率较低的符号被放置在离根节点较远的地方,概率较高的符号被放置在离根节点较近的地方,这使编码之后的字符串的平均长度和期望值降低,从而达到无损压缩数据的目的;

[0045] 具体的,首先将二进制数据分为16位一组,其次利用unique函数去重,然后利用tabulate函数求其概率,再次根据概率构造霍夫曼树,最后利用最小方差霍夫曼树对二进制数据进行压缩;

[0046] 步骤3:压缩后的所述二进制数据分成长度为4的不重叠块,所述不重叠块至多有16种组合;

[0047] 步骤4:获取不重叠块出现的概率;

[0048] 具体的,利用tabulate函数计算不重叠的块出现的概率;

[0049] 步骤5:根据下列字典中的码字对所述不重叠块进行编码;

[0050] 字典C_1＝{C,A}

[0051] 字典C_2＝{TA,TG,TC}

[0052] 字典C_3＝{GAT,GTA,GAC,GAG,GTC,GTG,GCA,GCT,GGT,GGA,GCG};

[0053] 具体的,依照这16种组合的概率高低依次从字典C_1-C_3选择码字进行编码。概率较高的组合就从字典中选择较短的码字进行编码。例如,将概率最高的组合编码为C,将概率第二高的组合编码为A,将概率第三高的组合编码为TA,依次类推。该文本文件的映射关系如表1所示。

[0054] 表1映射关系

[0055]

符号	概率	码字
1110	0.0819	C
1111	0.0819	A
0110	0.0699	TA

1101	0.0699	TG
0010	0.0675	TC
0011	0.0675	GAT
1010	0.0675	GTA
1001	0.0627	GAC
0101	0.0627	GAG
0100	0.0578	GTC
0111	0.0578	GTG
0000	0.0530	GCA
1011	0.0530	GCT
1000	0.0506	GGT
0001	0.0482	GGA
1100	0.0482	GCG

[0056] 步骤6:获取编码后DNA序列的GC含量,如果GC含量低于40%或者高于60%,则进行步骤5,改变映射关系继续编码,反之则进行步骤7;

[0057] 具体的,该文本编码为DNA序列的GC含量为56.40%。

[0058] 步骤7:寻找均聚物运行长度不小于3的情况,如果有,则进行替换修改,如果没有执行步骤8;

[0059] 具体的,当出现均聚物长度不小于3的情况,将第三个字符进行替换,例如,GGG和AAAA就会分别被修改为GGC和AATA,用C来替换G,T来替换A,这样还不会影响到GC含量。

[0060] 步骤8:获取编码率,并且将DNA序列输出。

[0061] 具体的,文本文件编码为DNA序列如表2所示,编码率约为4.0。

[0062] 表2 DNA序列

[0063]

GCAGGAGACGCACGACTCGCTGCGGTCGAGGCAGACTAGTGGCTCTAGTCTGAGATGATGTAGTGGC
TGTCCGACGCTGTGGCAGGTGTAGGAGCTGTGGCAGTAGCACGAGGTCGTCGCATAGTCGTGGATGT
AGAGTCGATGTATCGATGTAGCGTGTCGTAGTGGCAGCAGATTCGGATGGCATGGACTGGCGGATCGT
AGCAAGCATCGTGGACGTATAAGAGGAGTCGTAAGCTGTGGCTGGTGCTCGATCGTGTGGTGTGGTAT
GATCGCAATGGAGAGACGTAGGTAATATCGCTGTGCGCGGCAGCTGATGGTGATGGTGAGGCTGTGGC
TGACGCTTCTGGACGCGGAGGCTGAGTACCTCGAGTGGCAGTGTGGCTGGTGATTGAAGCGTAGATG
CGTAGTCGAGCGACGACGTCGTCGCATAGTCGCGGTGTCGCGGACGACGGAGTGCATAGTCCGAGTA
GAGACTCCTATACGCAGCGTCGGAGCGCTAGGTCGACGTGGATTAGCTGATGGTTGGACCGCGGAGG
ACGTGGCTCGCAGCATAGTCGATGTATCGGTGTGGACGACTGAGGTGTAGTAGATGGAGACGCAGAGT
ATAGTAGAGGTCGGAGACTGGTGGAGGATGAGGTCGAGGAGGCAGATGCAGGTGATGTCTAGTAGTC
ATAGGACAAGACGTGGACTAGTATGGACGAGGACGAGCAGCGAGCAGTAGCGGTGAATCGGAAGTC
GCTTCTCCGCAGCATAGTCGATGTAGTCGTAAGGAGGAGTCTCGTCGCAGTATGTCTGATCGGAGAGG
CGGCACTGGATGGTTCGTGGCTTCCGTAATCAGCGGCTTATGGTAGGAGTATCGATTGTCTGGATTGGT
CCACGATGCTGCTGGATAGAGCGCGGACCGCGGATAGTCTCTGCGTATGGATGGTGCGGCTAGCTGTA
TGTAGATGAGAGGTTAATGGCATGGATGCTCGCTGTCGGTTCTAATACCAGCGCGCTAGTGTAGACGA
GGTGGTAGGT

[0064] 本发明提出一种DNA存储的码字设计方法,首先将数据信息转换为二进制数据,其次再利用最小方差霍夫曼树对二进制数据进行压缩,以获得更高的编码率,然后将压缩后

的二进制数分成长度为4的不重叠的块，再根据字典中的码字对不重叠的块进行映射，获得DNA序列，最后计算DNA序列的GC含量和检查是否存在均聚物运行长度不小于3的情况，如果GC含量偏高或者偏低会重新生成新的映射关系；如果存在均聚物运行长度不小于3的情况，则根据修改规则进行修改。本发明在Intel(R)CPU 3.6GHz、10.0GB内存、Windows 10运行环境下，借助MATLAB对该算法进行仿真实验，实验结果表明本实施例的方法结果优于其他算法的实验结果。

[0065] 以上所述仅是本发明的优选实施方式，并不用于限制本发明，应当指出，对于本技术领域的普通技术人员，在不脱离本发明技术原理的前提下，还可以做出若干改进和变型，这些改进和变型也应视为本发明的保护范围。

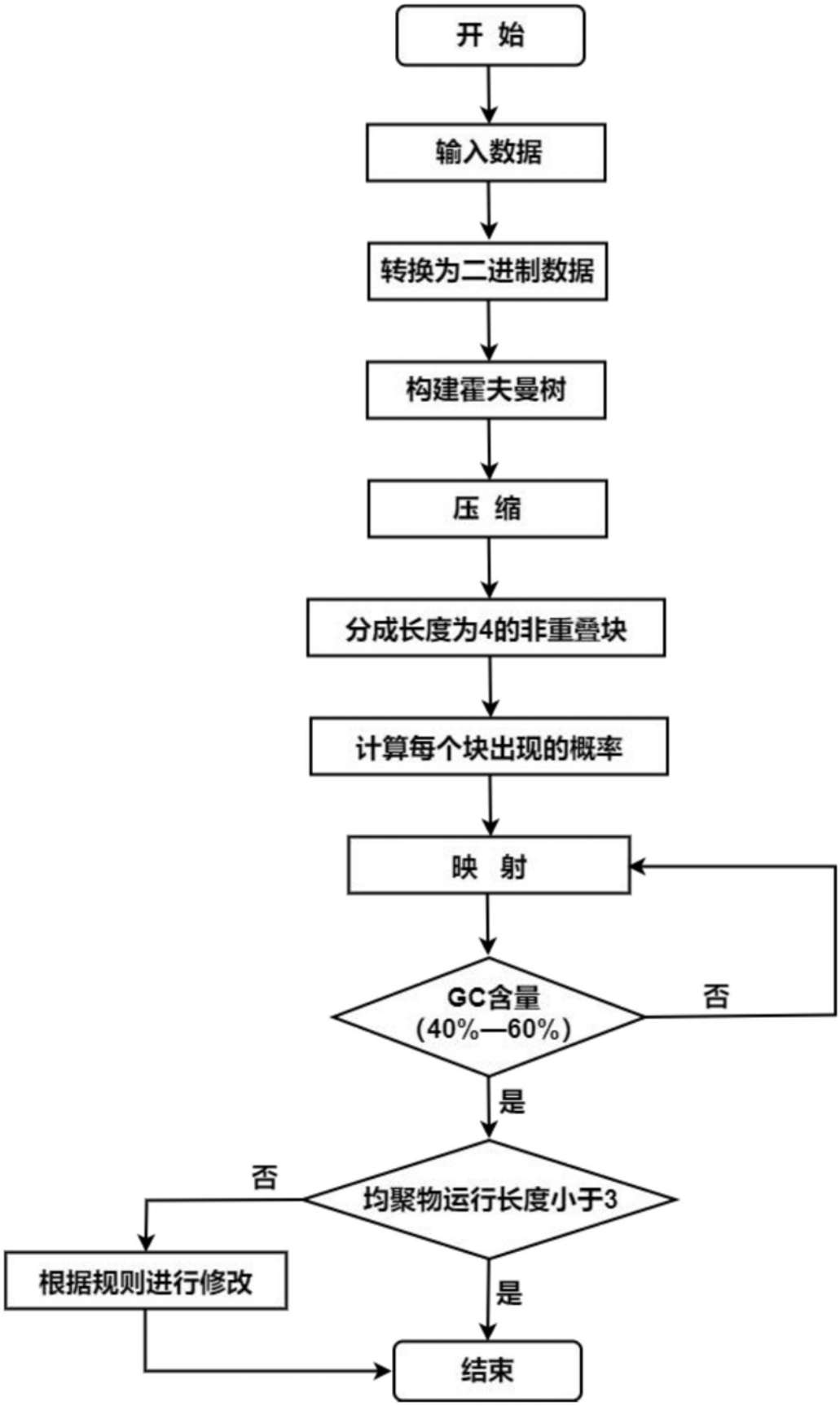

开 始
输入数据
转换为二进制数据
构建霍夫曼树
压 缩
分成长度为4的非重叠块
计算每个块出现的概率
映 射
GC含量
(40%—60%)
否
是
否
均聚物运行长度小于3
根据规则进行修改
是
结束

图1