

(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
2017 年 12 月 21 日 (21.12.2017)



(10) 国际公布号
WO 2017/214765 A1

(51) 国际专利分类号:
H03M 7/30 (2006.01) **G06F 19/10** (2011.01)

(21) 国际申请号: PCT/CN2016/085426

(22) 国际申请日: 2016 年 6 月 12 日 (12.06.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 深圳大学(SHENZHEN UNIVERSITY) [CN/CN]; 中国广东省深圳市南山区南海大道 3688 号, Guangdong 518060 (CN)。

(72) 发明人: 朱泽轩(ZHU, Zexuan); 中国广东省深圳市南山区南海大道 3688 号, Guangdong 518060 (CN)。 黄志安(HUANG, Zhian); 中国广东省深圳市南山区南海大道 3688 号, Guangdong 518060 (CN)。 孙怡雯(SUN, Yiwenn); 中国广东

省深圳市南山区南海大道 3688 号, Guangdong 518060 (CN)。 文振焜(WEN, Zhenkun); 中国广东省深圳市南山区南海大道 3688 号, Guangdong 518060 (CN)。

(74) 代理人: 深圳市恒申知识产权事务所(普通合伙)(HENSEN INTELLECTUAL PROPERTY FIRM); 中国广东省深圳市福田区南园路 68 号上步大厦 10H, Guangdong 518000 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI,

(54) Title: MULTI-THREAD FAST STORAGE LOSSLESS COMPRESSION METHOD AND SYSTEM FOR FASTQ DATA

(54) 发明名称: 针对FASTQ数据的多线程快速存储无损压缩方法及其系统

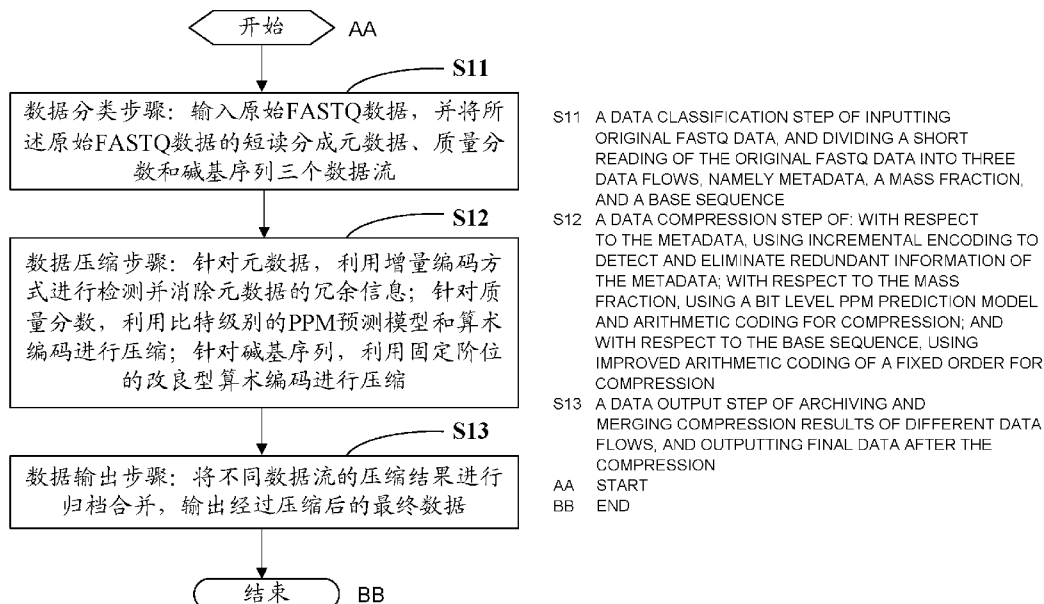


图 1

(57) Abstract: Provided is a multi-thread fast storage lossless compression method for FASTQ data, which is applied to compression of a DNA sequence. The method comprises: a data classification step of inputting original FASTQ data, and dividing a short reading of the original FASTQ data into three data flows, namely metadata, a mass fraction, and a base sequence (S11); a data compression step of: with respect to the metadata, using incremental encoding to detect and eliminate redundant information of the metadata; with respect to the mass fraction, using a bit level PPM prediction model and arithmetic coding for compression; and with respect to the

NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU,
RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH,
TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))。

base sequence, using improved arithmetic coding of a fixed order for compression (S12); and a data output step of archiving and merging compression results of different data flows, and outputting final data after the compression (S13). The solution can improve the compression efficiency and compression speed.

(57) 摘要: 一种针对FASTQ数据的多线程快速存储无损压缩方法, 应用于DNA序列的压缩, 其中, 所述方法包括: 数据分类步骤: 输入原始FASTQ数据, 并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流 (S11); 数据压缩步骤: 针对元数据, 利用增量编码方式进行检测并消除元数据的冗余信息; 针对质量分数, 利用比特级别的PPM预测模型和算术编码进行压缩; 针对碱基序列, 利用固定阶位的改良型算术编码进行压缩 (S12); 数据输出步骤: 将不同数据流的压缩结果进行归档合并, 输出经过压缩后的最终数据 (S13)。该方案能提高压缩效率和压缩速度。

发明名称：针对FASTQ数据的多线程快速存储无损压缩方法及其系统

技术领域

- [1] 本发明涉及数据压缩领域，尤其涉及一种针对FASTQ数据的多线程快速存储无损压缩方法及其系统。

背景技术

- [2] 随着DNA测序技术的发展，基因组测序成本越来越低。2014年，测定一个人类基因组的成本控制在1000美元的里程碑已经到来。由于测序效率的提高，DNA序列数据量呈现出爆炸性增长。由于DNA测序数据的增长速度远远超过了计算机微处理器和存储设备的增长速度，存储和分析DNA测序技术和大型基因组项目所产生的DNA数据“海啸”已经成为制约DNA测序产业进一步发展的一个重要瓶颈。而且，由于DNA测序技术正从高通量测序（High-Throughput Sequencing），又称为下一代测序（Next Generation Sequencing）发展到单分子测序技术（又称为第三代测序技术），FASTQ数据的短读从50~200bp不等的固定长度发展到1kbp~300kbp不等的固定长度，数据变化之大进一步制约DNA测序产业的发展，因此迫切需要相关的数据压缩技术投入使用。
- [3] 然而，目前一些主流高效的通用压缩软件如gzip（<http://www.gzip.org/>）、bzip2（<http://gzip.org/>）和LZMA（<http://www.7-zip.org/sdk.html>）。gzip软件对于要压缩的文件首先会采用基于LZ77算法的变种压缩方式，对得到的结果再根据情况使用静态Huffman编码或者动态Huffman编码方法进行压缩。bzip2软件把要压缩的数据进行分块处理（100~900KB每块），对于重复出现的字符序列使用BWT（Burrows-Wheeler transform）算法进行转换处理，然后再采用MTF（Move-To-Front transform）算法和哈弗曼编码（Huffman coding）进行压缩。LZMA软件使用了类似于LZ77算法的字典编码机制，对数据流、重复序列大小以及重复序列位置单独进行了压缩，支持几种散列链变体、二叉树以及基数树作为它的字典查找算法基础。

- [4] 但是，这些压缩方法并未考虑DNA数据的生物学特性，如长重复片段和互补回文结构等，导致对DNA序列数据的压缩效果不甚理想，而且压缩速度较慢。

对发明的公开

技术问题

- [5] 有鉴于此，本发明的目的在于提供一种针对FASTQ数据的多线程快速存储无损压缩方法及其系统，旨在解决现有技术中针对DNA序列数据的压缩效果较差且压缩速度较慢的问题。

问题的解决方案

技术解决方案

- [6] 本发明提出一种针对FASTQ数据的多线程快速存储无损压缩方法，应用于DNA序列的压缩，其特征在于，所述方法包括：

- [7] 数据分类步骤：输入原始FASTQ数据，并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流；

- [8] 数据压缩步骤：针对元数据，利用增量编码方式进行检测并消除元数据的冗余信息；针对质量分数，利用比特级别的PPM预测模型和算术编码进行压缩；针对碱基序列，利用固定阶位的改良型算术编码进行压缩；

- [9] 数据输出步骤：将不同数据流的压缩结果进行归档合并，输出经过压缩后的最终数据。

- [10] 优选的，在所述数据压缩步骤中采用了Pthreads的线程级并行编程方式来同时处理所述三个数据流的压缩。

- [11] 优选的，所述数据压缩步骤具体包括：

- [12] 针对质量分数，采用游程长读编码方式对质量分数的数据流进行初次压缩以实现预处理；

- [13] 利用比特级别的PPM预测模型和算术编码对经过预处理后的压缩数据进行再次压缩。

- [14] 优选的，所述数据压缩步骤具体包括：

- [15] 针对碱基序列，判断DNA序列的压缩模式是基于非参考基因的压缩模式还是基于参考基因的压缩模式；

- [16] 如果是基于非参考基因的压缩模式，则利用固定阶位的改良型算术编码将碱基序列的数据流进行压缩；
- [17] 如果是基于参考基因的压缩模式，则通过DNA数据匹配工具对DNA短读序列进行比对并剔除冗余，记录相应的匹配信息并以SAM格式文件保存，然后利用固定阶位的改良型算术编码将保存的SAM格式文件提取压缩。
- [18] 另一方面，本发明还提供一种针对FASTQ数据的多线程快速存储无损压缩系统，所述系统包括：
- [19] 数据分类模块，用于输入原始FASTQ数据，并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流；
- [20] 数据压缩模块，用于针对元数据，利用增量编码方式进行检测并消除元数据的冗余信息；针对质量分数，利用比特级别的PPM预测模型和算术编码进行压缩；针对碱基序列，利用固定阶位的改良型算术编码进行压缩；
- [21] 数据输出模块，用于将不同数据流的压缩结果进行归档合并，输出经过压缩后的最终数据。
- [22] 优选的，所述数据压缩模块采用Pthreads的线程级并行编程方式来同时处理所述三个数据流的压缩。
- [23] 优选的，所述数据压缩模块具体用于：
- [24] 针对质量分数，采用游程长读编码方式对质量分数的数据流进行初次压缩以实现预处理；
- [25] 利用比特级别的PPM预测模型和算术编码对经过预处理后的压缩数据进行再次压缩。
- [26] 优选的，所述数据压缩模块具体用于：
- [27] 针对碱基序列，判断DNA序列的压缩模式是基于非参考基因的压缩模式还是基于参考基因的压缩模式；
- [28] 如果是基于非参考基因的压缩模式，则利用固定阶位的改良型算术编码将碱基序列的数据流进行压缩；
- [29] 如果是基于参考基因的压缩模式，则通过DNA数据匹配工具对DNA短读序列进行比对并剔除冗余，记录相应的匹配信息并以SAM格式文件保存，然后利用

固定阶位的改良型算术编码将保存的SAM格式文件提取压缩。

发明的有益效果

有益效果

- [30] 本发明提供的技术方案充分利用了DNA数据的生物学特性，能在两种FASTQ数据（即长读和短读）中获得极高的压缩比，通过设计高效的编码方式把FASTQ数据拆分成三种类型的数据流并对此分别进行单独压缩，大大提高了压缩率，并且，对上游DNA数据匹配工具提供了接口，能应用于基于参考基因的压缩模式，发挥同源物种基因组之间的高度相似性，进一步提高重测序数据的压缩比。另外，通过采用Pthreads（POSIX threads）的线程级并行编程，能够同时压缩处理三种数据流所产生的中间文件，从而大大提高压缩速度，增加了适用性，进而提高了压缩效率。

对附图的简要说明

附图说明

- [31] 图1为本发明一实施方式中针对FASTQ数据的多线程快速存储无损压缩方法流程图；
- [32] 图2为本发明一实施方式中针对FASTQ数据的多线程快速存储无损压缩系统10的内部结构示意图。

发明实施例

本发明的实施方式

- [33] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明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仅用以解释本发明，并不用于限定本发明。
- [34] 本发明具体实施方式提供了一种针对FASTQ数据的多线程快速存储无损压缩方法，应用于DNA序列的压缩，其中，所述方法主要包括如下步骤：
- [35] S11、数据分类步骤：输入原始FASTQ数据，并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流；
- [36] S12、数据压缩步骤：针对元数据，利用增量编码方式进行检测并消除元数据

的冗余信息；针对质量分数，利用比特级别的PPM预测模型和算术编码进行压缩；针对碱基序列，利用固定阶位的改良型算术编码进行压缩；

[37] S13、数据输出步骤：将不同数据流的压缩结果进行归档合并，输出经过压缩后的最终数据。

[38] 本发明提供的一种针对FASTQ数据的多线程快速存储无损压缩方法充分利用了DNA数据的生物学特性，能在两种FASTQ数据（即长读和短读）中获得极高的压缩比，通过设计高效的编码方式把FASTQ数据拆分成三种类型的数据流并对此分别进行单独压缩，大大提高了压缩率。另外，通过采用Pthreads（POSIX threads）的线程级并行编程，能够同时压缩处理三种数据流所产生的中间文件，从而大大提高压缩速度，增加了适用性，进而提高了压缩效率。

[39] 以下将对本发明所提供的一种针对FASTQ数据的多线程快速存储无损压缩方法进行详细说明。

[40] 请参阅图1，为本发明一实施方式中针对FASTQ数据的多线程快速存储无损压缩方法流程图。

[41] 在步骤S11中，数据分类步骤：输入原始FASTQ数据，并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流。

[42] 在本实施方式中，DNA测序技术产生成千上万条短读，这些短读存储于以FASTQ为格式的文件中，包含测序产生的所有信息。在广泛使用的FASTQ格式中，每个短读包含四行，每行由换行符分割。每个短读以字符‘@’开始，后面紧接着元数据作为第一行，用来唯一标识短读。第二行是碱基数据，由仅包含{‘A’，‘T’，‘C’，‘G’，‘N’}五个字符的序列构成，其中字符‘N’表示不明确的碱基，可表示为{‘A’，‘T’，‘C’，‘G’}中任意一个字符。第三行以字符‘+’开始，紧接着与第一行相同的短读标识。最后一行为质量分数行，与碱基一一对应，表示每个碱基字符对应位置测序的可信度。

[43] 在步骤S12中，数据压缩步骤：针对元数据，利用增量编码方式进行检测并消除元数据的冗余信息；针对质量分数，利用比特级别的PPM（Prediction by partial matching）预测模型和算术编码进行压缩；针对碱基序列，利用固定阶位的改良型算术编码进行压缩。

- [44] 在本实施方式中，所述数据压缩步骤采用Pthreads的线程级并行编程方式来同时处理所述三个数据流的压缩，能够同时压缩处理三种数据流所产生的中间文件，从而大大提高压缩速度，增强其适用性。
- [45] 在本实施方式中，针对元数据，利用增量编码方式进行检测并消除元数据的冗余信息，然后利用固定阶位的改良型算术编码进行压缩，采用多线程化并行处理。
- [46] 在本实施方式中，所述数据压缩步骤具体包括：
- [47] 针对质量分数，采用游程长读编码方式对质量分数的数据流进行初次压缩以实现预处理；
- [48] 利用比特级别的PPM预测模型和算术编码对经过预处理后的压缩数据进行再次压缩。
- [49] 在本实施方式中，对于质量分数，由于考虑其保存着大量的连续重复的相同字符，故先采用游程长读编码压缩作为预处理，对质量分数的数据流进行初次压缩，然后再使用比特级别的PPM预测模型和算术编码对经过预处理后的压缩数据进行再次压缩，采用多线程化并行处理。在本实施方式中，这种比特级别的PPM预测模型和算术编码对于由约40种字符并且服从伪随机分布的质量分数进行无损压缩时，具有较高的压缩比和压缩速度。
- [50] 在本实施方式中，所述数据压缩步骤具体包括：
- [51] 针对碱基序列，判断DNA序列的压缩模式是基于非参考基因的压缩模式还是基于参考基因的压缩模式；
- [52] 如果是基于非参考基因的压缩模式，则利用固定阶位的改良型算术编码将碱基序列的数据流进行压缩；
- [53] 如果是基于参考基因的压缩模式，则通过DNA数据匹配工具对DNA短读序列进行比对并剔除冗余，记录相应的匹配信息并以SAM格式文件保存，然后利用固定阶位的改良型算术编码将保存的SAM格式文件提取压缩。
- [54] 在本实施方式中，因为DNA数据匹配工具的性能会直接影响到数据压缩的效果，所以以SAM格式文件为中间文件，对上游DNA数据匹配工具提供接口，为后续压缩性能的提升提供了可能。

- [55] 在本实施方式中，对于碱基序列，本发明可适用于两种DNA数据压缩模式，即基于非参考基因的压缩模式和基于参考基因压缩模式，通过判断DNA序列的压缩模式是属于哪一种模式，然后分别进行处理。本实施方式中的DNA数据匹配工具，如BWA（Burrows-Wheeler Aligner）工具、Bowtie工具和CompMap工具，这些DNA数据匹配工具能对DNA短读序列的比对并剔除冗余，实现对DNA短读序列的并行快速匹配，并应用于DNA序列的压缩存储方法中。
- [56] 对于基于非参考基因的压缩模式：此压缩模式针对只含有{‘A’，‘T’，‘C’，‘G’，‘N’}五个字符组成碱基序列，由于字符种类较少，采取固定阶位的策略，能大大减少不必要空余比特的浪费，从比特级别去降低存储每一个字符所需要的空间大小，故采取固定阶位的改良型算术编码压缩。也就是说，如果是基于非参考基因的压缩模式，则利用固定阶位的改良型算术编码将碱基序列的数据流进行压缩。
- [57] 对于基于参考基因的压缩模式：此压缩模式能基于以上DNA数据匹配工具进行对应的数据压缩，提供相应的接口，使其能应用于碱基序列的预处理，对DNA短读序列进行比对并剔除冗余，记录相应的匹配信息，以SAM格式文件保存，该匹配信息包含：匹配位置、回文串标记、匹配字符长度、匹配类型和非匹配字符，然后，此压缩模式把这五种信息拆分到各自的文件中，分别用固定阶位的改良型算术编码进行压缩。也就是说，如果是基于参考基因的压缩模式，则通过DNA数据匹配工具对DNA短读序列进行比对并剔除冗余，记录相应的匹配信息并以SAM格式文件保存，然后利用固定阶位的改良型算术编码将保存的SAM格式文件提取压缩。
- [58] 在步骤S13中，数据输出步骤：将不同数据流的压缩结果进行归档合并，输出经过压缩后的最终数据。
- [59] 在本实施方式中，本发明的技术方案从两方面去提升对FASTQ数据的压缩性能，即压缩比与压缩速度。其中，在高压压缩模式的比较下，不管以不定长度的长读为主还是以固定长度的短读为主的FASTQ数据中，本发明在两种模式下（即基于参考基因的压缩模式和基于非参考基因的压缩模式）均获得比主流通用压缩软件更高的压缩比和更快的压缩速度。这一优势经过一些数据的测试得以证

明，本发明从美国国家生物技术信息中心（NCBI, <http://www.ncbi.nlm.nih.gov/>）下载的两个FASTQ数据：ERR385912（641MB，短读长度为51，属于短读），ERR654984（1164MB，短读长度为64~502，属于长读），通过测试数据比较，本发明的压缩率平均比bzip2软件和gzip软件分别要高10.7%和15.2%，压缩速度平均比bzip2软件和gzip软件分别要快40.2%和45.5%。

[60] 本发明提供了一种针对FASTQ数据的多线程快速存储无损压缩方法充分利用了DNA数据的生物学特性，能在两种FASTQ数据（即长读和短读）中获得极高的压缩比，通过设计高效的编码方式把FASTQ数据拆分成三种类型的数据流并对此分别进行单独压缩，大大提高了压缩率，并且，对上游DNA数据匹配工具提供了接口，能应用于基于参考基因的压缩模式，发挥同源物种基因组之间的高度相似性，进一步提高重测序数据的压缩比。另外，通过采用Pthreads的线程级并行编程，能够同时压缩处理三种数据流所产生的中间文件，从而大大提高压缩速度，增加了适用性，进而提高了压缩效率。

[61] 本发明具体实施方式还提供一种针对FASTQ数据的多线程快速存储无损压缩系统10，主要包括：

[62] 数据分类模块11，用于输入原始FASTQ数据，并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流；

[63] 数据压缩模块12，用于针对元数据，利用增量编码方式进行检测并消除元数据的冗余信息；针对质量分数，利用比特级别的PPM预测模型和算术编码进行压缩；针对碱基序列，利用固定阶位的改良型算术编码进行压缩；

[64] 数据输出模块13，用于将不同数据流的压缩结果进行归档合并，输出经过压缩后的最终数据。

[65] 本发明提供了一种针对FASTQ数据的多线程快速存储无损压缩系统10，充分利用了DNA数据的生物学特性，能在两种FASTQ数据（即长读和短读）中获得极高的压缩比，通过设计高效的编码方式把FASTQ数据拆分成三种类型的数据流并对此分别进行单独压缩，大大提高了压缩率。另外，通过采用Pthreads的线程级并行编程，能够同时压缩处理三种数据流所产生的中间文件，从而大大提高压缩速度，增加了适用性，进而提高了压缩效率。

- [66] 请参阅图2，所示为本发明一实施方式中针对FASTQ数据的多线程快速存储无损压缩系统10的结构示意图。
- [67] 在本实施方式中，针对FASTQ数据的多线程快速存储无损压缩系统10，应用于DNA序列的压缩，主要包括数据分类模块11、数据压缩模块12以及数据输出模块13。
- [68] 数据分类模块11，用于输入原始FASTQ数据，并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流。
- [69] 在本实施方式中，DNA测序技术产生成千上万条短读，这些短读存储于以FASTQ为格式的文件中，包含测序产生的所有信息。在广泛使用的FASTQ格式中，每个短读包含四行，每行由换行符分割。每个短读以字符‘@’开始，后面紧接元数据作为第一行，用来唯一标识短读。第二行是碱基数据，由仅包含{‘A’，‘T’，‘C’，‘G’，‘N’}五个字符的序列构成，其中字符‘N’表示不明确的碱基，可表示为{‘A’，‘T’，‘C’，‘G’}中任意一个字符。第三行以字符‘+’开始，紧接着与第一行相同的短读标识。最后一行为质量分数行，与碱基一一对应，表示每个碱基字符对应位置测序的可信度。
- [70] 数据压缩模块12，用于针对元数据，利用增量编码方式进行检测并消除元数据的冗余信息；针对质量分数，利用比特级别的PPM预测模型和算术编码进行压缩；针对碱基序列，利用固定阶位的改良型算术编码进行压缩。
- [71] 在本实施方式中，所述数据压缩模块12采用Pthreads的线程级并行编程方式来同时处理所述三个数据流的压缩，能够同时压缩处理三种数据流所产生的中间文件，从而大大提高压缩速度，增强其适用性。
- [72] 针对元数据，数据压缩模块12利用增量编码方式进行检测并消除元数据的冗余信息，然后利用固定阶位的改良型算术编码进行压缩，采用多线程化并行处理。
- [73] 数据压缩模块12具体用于：
- [74] 针对质量分数，采用游程长读编码方式对质量分数的数据流进行初次压缩以实现预处理；
- [75] 利用比特级别的PPM预测模型和算术编码对经过预处理后的压缩数据进行再次

压缩。

- [76] 在本实施方式中，对于质量分数，由于考虑其保存着大量的连续重复的相同字符，故数据压缩模块12先采用游程长读编码压缩作为预处理，对质量分数的数据流进行初次压缩，然后再使用比特级别的PPM预测模型和算术编码对经过预处理后的压缩数据进行再次压缩，采用多线程化并行处理。在本实施方式中，这种比特级别的PPM预测模型和算术编码对于由约40种字符并且服从伪随机分布的质量分数进行无损压缩时，具有较高的压缩比和压缩速度。
- [77] 数据压缩模块12具体用于：
- [78] 针对碱基序列，判断DNA序列的压缩模式是基于非参考基因的压缩模式还是基于参考基因的压缩模式；
- [79] 如果是基于非参考基因的压缩模式，则利用固定阶位的改良型算术编码将碱基序列的数据流进行压缩；
- [80] 如果是基于参考基因的压缩模式，则通过DNA数据匹配工具对DNA短读序列进行比对并剔除冗余，记录相应的匹配信息并以SAM格式文件保存，然后利用固定阶位的改良型算术编码将保存的SAM格式文件提取压缩。
- [81] 在本实施方式中，因为DNA数据匹配工具的性能会直接影响到数据压缩的效果，所以以SAM格式文件为中间文件，对上游DNA数据匹配工具提供接口，为后续压缩性能的提升提供了可能。
- [82] 本实施方式中的DNA数据匹配工具，如BWA（Burrows-Wheeler Aligner）工具、Bowtie工具和CompMap工具，这些DNA数据匹配工具能对DNA短读序列的比对并剔除冗余，实现对DNA短读序列的并行快速匹配，并应用于DNA序列的压缩存储方法中。
- [83] 在本实施方式中，对于碱基序列，本发明可适用于两种DNA数据压缩模式，即基于非参考基因的压缩模式和基于参考基因压缩模式，通过判断DNA序列的压缩模式是属于哪一种模式，然后分别进行处理。其中，两种DNA数据压缩模式的具体处理流程如前述的步骤S12所示，在此就不做重复描述。
- [84] 数据输出模块13，用于将不同数据流的压缩结果进行归档合并，输出经过压缩后的最终数据。

- [85] 本发明提供的一种针对FASTQ数据的多线程快速存储无损压缩系统10，充分利用了DNA数据的生物学特性，能在两种FASTQ数据（即长读和短读）中获得极高的压缩比，通过设计高效的编码方式把FASTQ数据拆分成三种类型的数据流并对此分别进行单独压缩，大大提高了压缩率，并且，对上游DNA数据匹配工具提供了接口，能应用于基于参考基因的压缩模式，发挥同源物种基因组之间的高度相似性，进一步提高重测序数据的压缩比。另外，通过采用Pthreads的线程级并行编程，能够同时压缩处理三种数据流所产生的中间文件，从而大大提高压缩速度，增加了适用性，进而提高了压缩效率。
- [86] 值得注意的是，上述实施例中，所包括的各个单元只是按照功能逻辑进行划分的，但并不局限于上述的划分，只要能够实现相应的功能即可；另外，各功能单元的具体名称也只是为了便于相互区分，并不用于限制本发明的保护范围。
- [87] 另外，本领域普通技术人员可以理解实现上述各实施例方法中的全部或部分步骤是可以通程序来指令相关的硬件来完成，相应的程序可以存储于一计算机可读取存储介质中，所述的存储介质，如ROM/RAM、磁盘或光盘等。
- [88] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

权利要求书

- [权利要求 1] 一种针对FASTQ数据的多线程快速存储无损压缩方法，应用于DNA序列的压缩，其特征在于，所述方法包括：
- 数据分类步骤：输入原始FASTQ数据，并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流；
- 数据压缩步骤：针对元数据，利用增量编码方式进行检测并消除元数据的冗余信息；针对质量分数，利用比特级别的PPM预测模型和算术编码进行压缩；针对碱基序列，利用固定阶位的改良型算术编码进行压缩；
- 数据输出步骤：将不同数据流的压缩结果进行归档合并，输出经过压缩后的最终数据。
- [权利要求 2] 如权利要求1所述的针对FASTQ数据的多线程快速存储无损压缩方法，其特征在于，在所述数据压缩步骤中采用了Pthreads的线程级并行编程方式来同时处理所述三个数据流的压缩。
- [权利要求 3] 如权利要求2所述的针对FASTQ数据的多线程快速存储无损压缩方法，其特征在于，所述数据压缩步骤具体包括：
- 针对质量分数，采用游程长读编码方式对质量分数的数据流进行初次压缩以实现预处理；
- 利用比特级别的PPM预测模型和算术编码对经过预处理后的压缩数据进行再次压缩。
- [权利要求 4] 如权利要求2所述的针对FASTQ数据的多线程快速存储无损压缩方法，其特征在于，所述数据压缩步骤具体包括：
- 针对碱基序列，判断DNA序列的压缩模式是基于非参考基因的压缩模式还是基于参考基因的压缩模式；
- 如果是基于非参考基因的压缩模式，则利用固定阶位的改良型算术编码将碱基序列的数据流进行压缩；
- 如果是基于参考基因的压缩模式，则通过DNA数据匹配工具对DNA短读序列进行比对并剔除冗余，记录相应的匹配信息并以SAM

格式文件保存，然后利用固定阶位的改良型算术编码将保存的SAM格式文件提取压缩。

[权利要求 5]

一种针对FASTQ数据的多线程快速存储无损压缩系统，其特征在于，所述系统包括：

数据分类模块，用于输入原始FASTQ数据，并将所述原始FASTQ数据的短读分成元数据、质量分数和碱基序列三个数据流；

数据压缩模块，用于针对元数据，利用增量编码方式进行检测并消除元数据的冗余信息；针对质量分数，利用比特级别的PPM预测模型和算术编码进行压缩；针对碱基序列，利用固定阶位的改良型算术编码进行压缩；

数据输出模块，用于将不同数据流的压缩结果进行归档合并，输出经过压缩后的最终数据。

[权利要求 6]

如权利要求5所述的针对FASTQ数据的多线程快速存储无损压缩系统，其特征在于，所述数据压缩模块采用Pthreads的线程级并行编程方式来同时处理所述三个数据流的压缩。

[权利要求 7]

如权利要求6所述的针对FASTQ数据的多线程快速存储无损压缩系统，其特征在于，所述数据压缩模块具体用于：

针对质量分数，采用游程长读编码方式对质量分数的数据流进行初次压缩以实现预处理；

利用比特级别的PPM预测模型和算术编码对经过预处理后的压缩数据进行再次压缩。

[权利要求 8]

如权利要求6所述的针对FASTQ数据的多线程快速存储无损压缩系统，其特征在于，所述数据压缩模块具体用于：

针对碱基序列，判断DNA序列的压缩模式是基于非参考基因的压缩模式还是基于参考基因的压缩模式；

如果是基于非参考基因的压缩模式，则利用固定阶位的改良型算术编码将碱基序列的数据流进行压缩；

如果是基于参考基因的压缩模式，则通过DNA数据匹配工具对DN

A短读序列进行比对并剔除冗余，记录相应的匹配信息并以SAM格式文件保存，然后利用固定阶位的改良型算术编码将保存的SAM格式文件提取压缩。

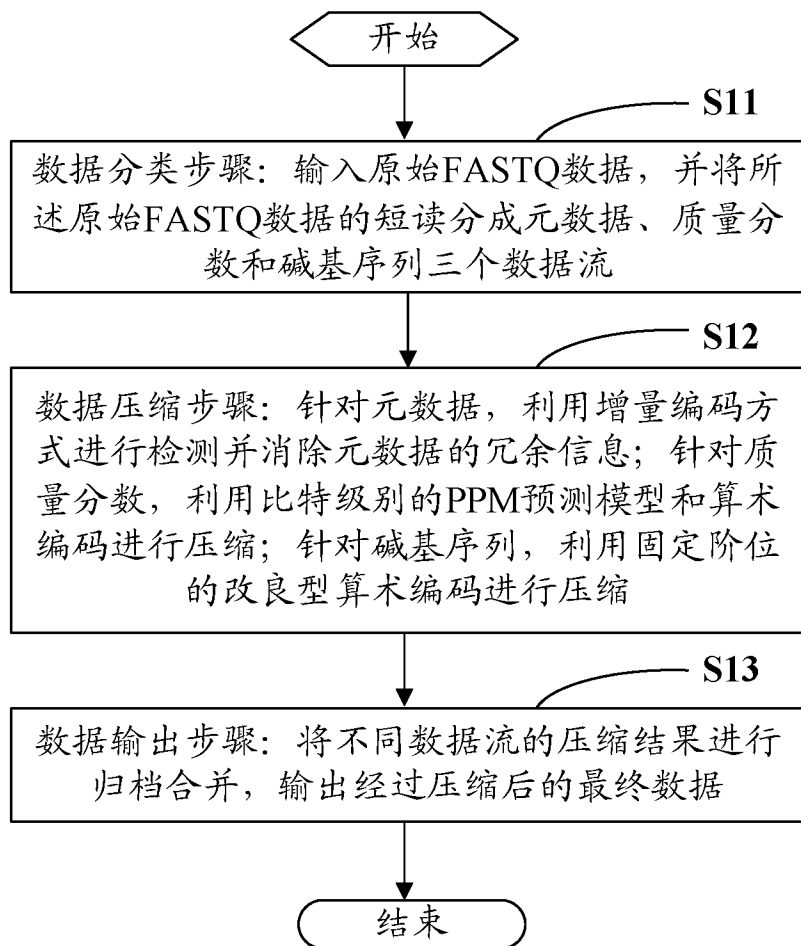


图 1

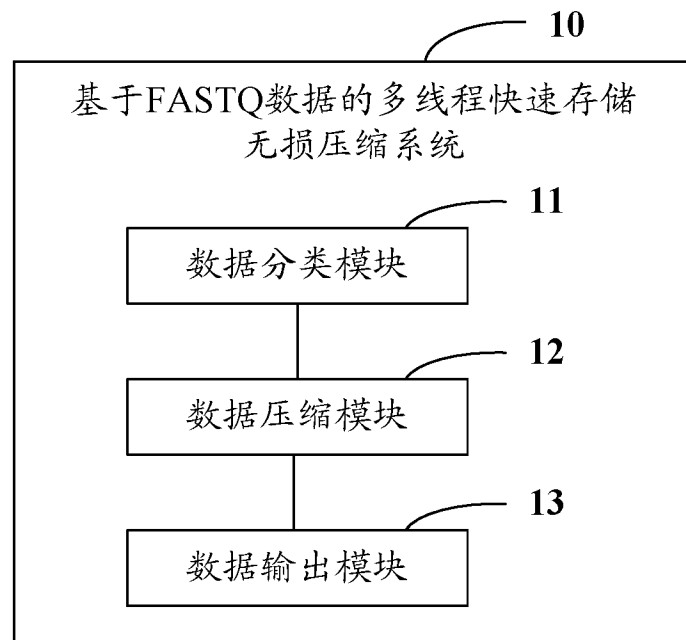


图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/085426

A. CLASSIFICATION OF SUBJECT MATTER

H03M 7/30 (2006.01) i; G06F 19/10 (2011.01) n

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)

H03M; 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)

CNPAT; EPODOC; WPI; google; CNKI: DNA, FASTQ, base, metadata, quality, score, compress, code, arithmetic, incremental, fixed, order, run, length, compare, difference, parallel, combination, mass fraction, encoding

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ZHANG, Yongpeng, "Lossless Comprssion of High-through DNA Sequence Data," China Master's Theses Full-text Database Basic Science, no. 12, 15 December 2015 (15.12.2015), ISSN: 1674-0246, the abstract, pp. 3-5, 12-15 and 20-25	1-8
Y	AKGUN, Mete et al., "A new PPM model for quality score compression. Signal Processing and Communications Applications Conference (SIU), 26 April 2013 (26.04.2013), ISSN: 2165-0608, section V	1-8
Y	BEHZADI, Behshad et al. "DNA Compression Challenge Revisited: A Dynamic Programming Approach," Combinatorial Pattern Matching, 22 June 2005 (22.06.2005), ISSN: 0302-9743, section 2	1-8
A	CN 103559020 A (INSTITUTE OF SOFTWARE, CHINESE ACADEMY OF SCIENCES et al.) 05 February 2014 (05.02.2014) the whole document	1-8

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 24 February 2017	Date of mailing of the international search report 22 March 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer OU, Xiaodan Telephone No. (86-10) 82246933

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/085426

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2015120170 A1 (BIGDATABIO, LLC.) 13 August 2015 (13.08.2015) the whole document	1-8
A	US 2015227686 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 13 August 2015 (13.08.2015) the whole document	1-8

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/085426

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103559020 A	05 February 2014	CN 103559020 B	06 July 2016
WO 2015120170 A1	13 August 2015	US 2016352699 A1	01 December 2016
US 2015227686 A1	13 August 2015	None	

A. 主题的分类 H03M 7/30(2006.01)i; G06F 19/10(2011.01)n 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) H03M; 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)) CNPAT; EPODOC; WPI; google; CNKI; DNA, FASTQ, 碱基, 元数据, 质量分数, 压缩, 编码, 算术, PPM, 增量, 固定, 阶, 游程, 比对, 差, 并行, SAM, 合并, base, metadata, quality, score, compress, code, arithmetic, incremental, fixed, order, run, length, compare, difference, parallel, combination																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>Y</td> <td>张永朋, . "高通量DNA测序数据无损压缩研究," 中国优秀硕士学位论文全文数据库 基础科学辑, 第12期, 2015年 12月 15日 (2015 - 12 - 15), ISSN: 1674-0246, , 摘要, 第3-5, 12-15, 20-25页</td> <td>1-8</td> </tr> <tr> <td>Y</td> <td>AKGÜN, Mete 等, . "A new PPM model for quality score compression," Signal Processing and Communications Applications Conference (SIU), 2013年 4月 26日 (2013 - 04 - 26), ISSN: 2165-0608, , 第V节</td> <td>1-8</td> </tr> <tr> <td>Y</td> <td>BEHZADI, Behshad 等, . "DNA Compression Challenge Revisited: A Dynamic Programming Approach," Combinatorial Pattern Matching, 2005年 6月 22日 (2005 - 06 - 22), ISSN: 0302-9743, , 第2节</td> <td>1-8</td> </tr> <tr> <td>A</td> <td>CN 103559020 A (中国科学院软件研究所 等) 2014年 2月 5日 (2014 - 02 - 05) 全文</td> <td>1-8</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	Y	张永朋, . "高通量DNA测序数据无损压缩研究," 中国优秀硕士学位论文全文数据库 基础科学辑, 第12期, 2015年 12月 15日 (2015 - 12 - 15), ISSN: 1674-0246, , 摘要, 第3-5, 12-15, 20-25页	1-8	Y	AKGÜN, Mete 等, . "A new PPM model for quality score compression," Signal Processing and Communications Applications Conference (SIU), 2013年 4月 26日 (2013 - 04 - 26), ISSN: 2165-0608, , 第V节	1-8	Y	BEHZADI, Behshad 等, . "DNA Compression Challenge Revisited: A Dynamic Programming Approach," Combinatorial Pattern Matching, 2005年 6月 22日 (2005 - 06 - 22), ISSN: 0302-9743, , 第2节	1-8	A	CN 103559020 A (中国科学院软件研究所 等) 2014年 2月 5日 (2014 - 02 - 05) 全文	1-8
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
Y	张永朋, . "高通量DNA测序数据无损压缩研究," 中国优秀硕士学位论文全文数据库 基础科学辑, 第12期, 2015年 12月 15日 (2015 - 12 - 15), ISSN: 1674-0246, , 摘要, 第3-5, 12-15, 20-25页	1-8															
Y	AKGÜN, Mete 等, . "A new PPM model for quality score compression," Signal Processing and Communications Applications Conference (SIU), 2013年 4月 26日 (2013 - 04 - 26), ISSN: 2165-0608, , 第V节	1-8															
Y	BEHZADI, Behshad 等, . "DNA Compression Challenge Revisited: A Dynamic Programming Approach," Combinatorial Pattern Matching, 2005年 6月 22日 (2005 - 06 - 22), ISSN: 0302-9743, , 第2节	1-8															
A	CN 103559020 A (中国科学院软件研究所 等) 2014年 2月 5日 (2014 - 02 - 05) 全文	1-8															
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
<table border="0"> <tr> <td style="vertical-align: top;"> * 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 </td> <td style="vertical-align: top;"> "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 </td> </tr> </table>			* 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													
* 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		国际检索报告邮寄日期															
2017年 2月 24日		2017年 3月 22日															
ISA/CN的名称和邮寄地址		受权官员															
中华人民共和国国家知识产权局(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		欧晓丹 电话号码 (86-10)82246933															

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
A	WO 2015120170 A1 (BIGDATABIO, LLC) 2015年 8月 13日 (2015 - 08 - 13) 全文	1-8
A	US 2015227686 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2015年 8月 13日 (2015 - 08 - 13) 全文	1-8

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/085426

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	103559020	A	2014年 2月 5日	CN 103559020 B	2016年 7月 6日
WO	2015120170	A1	2015年 8月 13日	US 2016352699 A1	2016年 12月 1日
US	2015227686	A1	2015年 8月 13日	无	