



(12) 发明专利申请

(10) 申请公布号 CN 117581304 A

(43) 申请公布日 2024. 02. 20

(21) 申请号 202280045844.0

(22) 申请日 2022.09.21

(30) 优先权数据

63/247296 2021.09.22 US

63/247301 2021.09.22 US

17/944809 2022.09.14 US

17/944948 2022.09.14 US

(72) 发明人 G·D·帕纳比 E·J·奥贾德

D·卡什夫哈吉吉

(74) 专利代理机构 北京泛华伟业知识产权代理
有限公司 11280

专利代理师 王博

(51) Int.Cl.

G16B 40/20 (2006.01)

(85) PCT国际申请进入国家阶段日

2023.12.27

(86) PCT国际申请的申请数据

PCT/US2022/044293 2022.09.21

(87) PCT国际申请的公布数据

WO2023/049215 EN 2023.03.30

(71) 申请人 因美纳有限公司

地址 美国加利福尼亚州

权利要求书3页 说明书90页 附图60页

(54) 发明名称

基于压缩状态的碱基调用

(57) 摘要

所公开的技术包括一种系统。该系统包括：被配置为处理簇的测序图像并且产生空间卷积特征的空间卷积神经网络，被配置为从该空间卷积特征选择包含簇的中心的空间卷积特征子集的滤波逻辑，被配置为将该空间卷积特征子集压缩为压缩特征集合的压缩逻辑，被配置为访问该压缩特征集合中的压缩特征的状态信息的上下文逻辑，被配置为处理有状态压缩特征集合并且产生时间卷积的有状态特征的时间卷积神经网络，以及被配置为基于该时间卷积的有状态特征而产生针对该簇的碱基调用的碱基调用逻辑。



1. 一种系统,包括:

空间卷积神经网络,所述空间卷积神经网络被配置为处理簇的测序图像并且产生空间卷积特征;

滤波逻辑,所述滤波逻辑被配置为从所述空间卷积特征选择包含所述簇的中心的空间卷积特征子集;

压缩逻辑,所述压缩逻辑被配置为将所述空间卷积特征子集压缩为压缩特征集合,其中所述空间卷积特征子集具有M个通道,并且所述压缩特征集合具有N个通道,并且其中 $M > N$;

上下文化逻辑,所述上下文化逻辑被配置为访问所述压缩特征集合中的压缩特征的状态信息,并且将所述状态信息与所述压缩特征相加以生成有状态压缩特征集合;

时间卷积神经网络,所述时间卷积神经网络被配置为处理所述有状态压缩特征集合,并且产生时间卷积的有状态特征;以及

碱基调用逻辑,所述碱基调用逻辑被配置为基于所述时间卷积的有状态特征而生成针对所述簇的碱基调用。

2. 根据权利要求1所述的系统,其中每通道状态为响应于独立地处理所述测序图像而通过线性模型在逐簇的基础上生成的强度变化校正系数。

3. 根据权利要求2所述的系统,其中所述强度变化校正系数包括补偿所述簇的强度分布之间的尺度变化的放大系数。

4. 根据权利要求2或3所述的系统,其中所述强度变化校正系数包括补偿沿着多个通道的所述簇的所述强度分布之间的移位变化的通道特异性偏移系数。

5. 根据权利要求2至4中任一项所述的系统,其中所述强度变化校正系数包括补偿所述移位变化的公共偏移系数。

6. 根据权利要求2至5中任一项所述的系统,其中基于将在测序运行的当前测序循环处确定的当前强度统计值的分析与在所述测序运行的在所述当前测序循环之前的一个或多个先前测序循环处确定的历史强度统计值的分析相结合,在所述逐簇的基础上生成所述强度变化校正系数。

7. 根据权利要求4至6中任一项所述的系统,其中所述簇的所述强度分布的特征在于所述多个通道中的通道的通道特异性强度值。

8. 根据权利要求1至7中任一项所述的系统,其中所述每通道状态的特征在于所述通道的通道特异性状态值。

9. 根据权利要求1至8中任一项所述的系统,其中所述每通道状态是在逐位的基础上利用每簇通道特异性位进行编码的。

10. 根据权利要求9所述的系统,其中所述通道的子集的通道特异性状态值是在所述逐位的基础上利用所述每簇通道特异性位进行编码的。

11. 根据权利要求9或10所述的系统,其中所述通道特异性状态值跨所述通道被平均以生成泛通道状态值,其中所述泛通道状态值是在所述逐位的基础上利用所述每簇通道特异性位进行编码的。

12. 根据权利要求9至11中任一项所述的系统,其中所述每通道状态是在所述逐位的基础上利用所述每簇通道特异性位进行级联的。

13. 根据权利要求9至12中任一项所述的系统, 其中所述每通道状态是在所述逐位的基础上利用所述每簇通道特异性位进行求和的。

14. 根据权利要求1至13中任一项所述的系统, 其中所述每通道状态是基于由空间处理逻辑在所述当前测序循环处和在所述先前测序循环处产生的每簇通道特异性位的压缩值。

15. 根据权利要求14所述的系统, 其中所述每通道状态为所述压缩值的平均值。

16. 根据权利要求14所述的系统, 其中所述每通道状态为所述压缩值中的最大压缩值。

17. 根据权利要求14所述的系统, 其中所述每通道状态为所述压缩值中的最小压缩值。

18. 根据权利要求14所述的系统, 其中所述每通道状态为所述压缩值的指数加权平均值。

19. 根据权利要求18所述的系统, 其中所述指数加权平均值是基于对最近测序循环比对之前测序循环进行更多加权而确定的。

20. 根据权利要求14或15所述的系统, 其中所述每通道状态为所述压缩值的移动平均值。

21. 根据权利要求20所述的系统, 其中所述移动平均值使用来自所述先前测序循环的压缩值的滚动子集。

22. 根据权利要求14至21中任一项所述的系统, 其中所述每通道状态包括启用状态值和非启用状态值。

23. 根据权利要求22所述的系统, 其中所述启用状态值和所述非启用状态值为通道特异性的。

24. 根据权利要求22或23所述的系统, 其中所述启用状态值和所述非启用状态值是从先前碱基调用确定的。

25. 根据权利要求22至24中任一项所述的系统, 其中所述启用状态值和所述非启用状态值是基于所述压缩值的全局最大值和全局最小值而确定的。

26. 根据权利要求22至25中任一项所述的系统, 其中所述启用状态值和所述非启用状态值是基于所述压缩值的指数加权平均值而确定的。

27. 根据权利要求1至26中任一项所述的系统, 其中到所述空间卷积神经网络的输入包括针对当前测序循环、一个或多个先前测序循环以及一个或多个下一测序循环的测序图像的滑动窗口。

28. 根据权利要求27所述的系统, 其中特征集合是从针对所述当前测序循环、所述先前测序循环和下一侧翼测序循环的测序图像的所述滑动窗口中的测序图像的通道特异性像素提取的, 其中每簇中心特征集合是从针对所述当前测序循环、所述先前测序循环和所述下一测序循环的特征集合挑选的。

29. 根据权利要求28所述的系统, 其中通道特异性位集合是从针对所述当前测序循环、所述先前测序循环和所述下一测序循环的所述每簇中心特征集合压缩的。

30. 一种系统, 包括:

空间处理逻辑, 所述空间处理逻辑被配置为处理其中簇的强度分布跨测序图像的通道特异性像素分散的输入, 并且产生其中所述簇的所述强度分布被聚集到从所述通道特异性像素提取的特征中的空间紧凑输出;

簇聚焦逻辑, 所述簇聚焦逻辑被配置为从所述特征挑选表征在所述簇的中心处检测到

的峰强度的那些每簇中心特征;

压缩逻辑,所述压缩逻辑被配置为将所述每簇中心特征提取为每簇通道特异性位,其中所述每簇中心特征具有M个特征通道,其中所述每簇通道特异性位具有N个位通道,并且其中 $M > N$;

状态生成逻辑,所述状态生成逻辑被配置为生成针对所述每簇通道特异性位的每通道状态;

时间处理逻辑,所述时间处理逻辑被配置为处理所述每簇通道特异性位和所述每通道状态,并且产生时间紧凑输出;以及

碱基调用逻辑,所述碱基调用逻辑被配置为基于所述时间紧凑输出而产生碱基调用。

基于压缩状态的碱基调用

[0001] 优先权申请

[0002] 本申请要求2022年9月14日提交的标题为“Compressed State-Based Base Calling”的美国非临时专利申请号17/944,948(代理人案卷号ILLM 1043-4/IP-2208-US)的优先权,该申请继而要求2021年9月22日提交的美国临时专利申请号63/247,301(代理人案卷号ILLM 1043-2/IP-2208-PRV)的权益。

[0003] 本申请要求2022年9月14日提交的标题为“State-Based Base Calling”的美国非临时专利申请号17/944,809(代理人案卷号ILLM 1043-3/IP-2073-US)的优先权,该申请继而要求2021年9月22日提交的美国临时专利申请号63/247,296(代理人案卷号ILLM 1043-1/IP-2073-PRV)的权益。

[0004] 将这些优先权申请以引用方式并入,即如同在本文完整示出一样,以用于所有目的。

技术领域

[0005] 本发明所公开的技术涉及人工智能类型计算机和数字数据处理系统以及对应数据处理方法和用于仿真智能的产品(即,基于知识的系统、推断系统和知识采集系统);并且包括用于不确定性推断的系统(例如,模糊逻辑系统)、自适应系统、机器学习系统和人工神经网络。具体地,所公开的技术涉及将深度神经网络诸如深度卷积神经网络用于分析数据。

[0006] 文献并入

[0007] 以下文献以引用方式并入,即如同在本文完整示出一样,以用于所有目的:

[0008] 2021年5月4日提交的标题为“Equalization-Based Image Processing And Spatial Crosstalk Attenuator”的美国非临时专利申请号17/308,035(代理人案卷号ILLM 1032-2/IP-1991-US);

[0009] 2020年10月27日提交的标题为“Systems And Methods For Per-Cluster Intensity Correction And Base Calling”的美国临时专利申请号63/106,256(代理人案卷号ILLM 1034-1/IP-2026-PRV);

[0010] 2018年3月1日提交的标题为“Optical Distortion Correction For Imaged Samples”的美国非临时专利申请号15/909,437;

[0011] 2020年3月20日提交的标题为“Training Data Generation For Artificial Intelligence-Based Sequencing”的美国非临时专利申请号16/825,987(代理人案卷号ILLM 1008-16/IP-1693-US);

[0012] 2020年3月20日提交的标题为“Artificial Intelligence-Based Generation Of Sequencing Metadata”的美国非临时专利申请号16/825,991(代理人案卷号ILLM 1008-17/IP-1741-US);

[0013] 2020年3月20日提交的标题为“Artificial Intelligence-Based Base Calling”的美国非临时专利申请号16/826,126(代理人案卷号ILLM 1008-18/IP-1744-US);

[0014] 2020年3月20日提交的标题为“Artificial Intelligence-Based Quality

Scoring”的美国非临时专利申请号16/826,134(代理人案卷号ILLM 1008-19/IP-1747-US) ;

[0015] 2020年3月21日提交的标题为“Artificial Intelligence-Based Sequencing”的美国非临时专利申请号16/826,168(代理人案卷号ILLM 1008-20/IP-1752-US) ;

[0016] 2021年2月12日提交的标题为“Artificial Intelligence-Based Base Calling Of Index Sequences”的美国非临时专利申请号17/175,546(代理人案卷号ILLM 1015-2/IP-1857-US) ;

[0017] 2021年2月19日提交的标题为“Artificial Intelligence-Based Many-To-Many Base Calling”的美国非临时专利申请号17/180,542(代理人案卷号ILLM 1016-2/IP-1858-US) ;

[0018] 2021年2月15日提交的标题为“Knowledge Distillation-Based Compression Of Artificial Intelligence-Based Base Caller”的美国非临时专利申请号17/176,151(代理人案卷号ILLM 1017-2/IP-1859-US) ;

[0019] 2020年8月28日提交的标题为“Detecting And Filtering Clusters Based On Artificial Intelligence-Predicted Base Calls”的美国临时专利申请号63/072,032(代理人案卷号ILLM 1018-1/IP-1860-PRV) ;

[0020] 2021年3月16日提交的标题为“Tile Location And/Or Cycle Based Weight Set Selection For Base Calling”的美国临时专利申请号63/161,880(代理人案卷号ILLM 1019-1/IP-1861-PRV) ;

[0021] 2021年3月16日提交的标题为“Neural Network Parameter Quantization For Base Calling”的美国临时专利申请号63/161,896(代理人案卷号ILLM 1019-2/IP-2049-PRV) ;

[0022] 2021年2月15日提交的标题为“Hardware Execution And Acceleration Of Artificial Intelligence-Based Base Caller”的美国非临时专利申请号17/176,147(代理人案卷号ILLM 1020-2/IP-1866-US) ;

[0023] 2021年8月3日提交的标题为“Base Calling Using Multiple Base Caller Models”的美国临时专利申请号63/228,954(代理人案卷号ILLM 1021-1/IP-1856-PRV) ;

[0024] 2021年2月18日提交的标题为“Data Compression For Artificial Intelligence-Based Base Calling”的美国非临时专利申请号17/179,395(代理人案卷号ILLM 1029-2/IP-1964-US) ;

[0025] 2021年2月19日提交的标题为“Split Architecture For Artificial Intelligence-Based Base Caller”的美国非临时专利申请号17/180,480(代理人案卷号ILLM 1030-2/IP-1982-US) ;

[0026] 2021年2月19日提交的标题为“Bus Network For Artificial Intelligence-Based Base Caller”的美国非临时专利申请号17/180,513(代理人案卷号ILLM 1031-2/IP-1965-US) ;

[0027] 2021年3月31日提交的标题为“Artificial Intelligence-Based Base Caller With Contextual Awareness”的美国临时专利申请号63/169,163(代理人案卷号ILLM 1033-1/IP-2007-PRV) ;

[0028] 2021年6月29日提交的标题为“Self-Learned Base Caller, Trained Using Oligo Sequences”的美国临时专利申请号63/216,419(代理人案卷号ILLM 1038-1/IP-2050-PRV) ;

[0029] 2021年6月29日提交的标题为“Self-Learned Base Caller, Trained Using Organism Sequences”的美国临时专利申请号63/216,404(代理人案卷号ILLM 1038-2/IP-2094-PRV) ;

[0030] 2021年7月19日提交的标题为“Specialist Signal Profilers For Base Calling”的美国临时专利申请号63/223,408(代理人案卷号ILLM 1041-1/IP-2063-PRV) ;

[0031] 2021年7月28日提交的标题为“Quality Score Calibration Of Basecalling Systems”的美国临时专利申请号63/226,707(代理人案卷号ILLM 1045-1/IP-2093-PRV) ;

[0032] 2021年7月1日提交的标题为“Efficient Artificial Intelligence-Based Base Calling Of Index Sequences”的美国临时专利申请号63/217,644(代理人案卷号ILLM 1046-1/IP-2135-PRV) ;

[0033] 2014年10月31日提交的标题为“Image Analysis Useful For Patterned Objects”的美国非临时专利申请号14/530,299;

[0034] 2014年12月3日提交的标题为“Methods And Systems For Analyzing Image Data”的美国非临时专利申请号15/153,953;

[0035] 2018年1月5日提交的标题为“Phasing Correction”的美国非临时专利申请号15/863,241;

[0036] 2013年9月6日提交的标题为“Centroid Markers For Image Analysis Of High Density Clusters In Complex Polynucleotide Sequencing”的美国非临时专利申请号14/020,570;

[0037] 2009年9月23日提交的标题为“Method And System For Determining The Accuracy Of DNA Base Identifications”的美国非临时专利申请号12/565,341;

[0038] 2007年3月30日提交的标题为“Systems And Devices For Sequence By Synthesis Analysis”的美国非临时专利申请号12/295,337;

[0039] 2008年1月28日提交的标题为“Image Data Efficient Genetic Sequencing Method And System”的美国非临时专利申请号12/020,739;

[0040] 2013年3月15日提交的标题为“Biosensors For Biological Or Chemical Analysis And Systems And Methods For Same”的美国非临时专利申请号13/833,619(代理人案卷号IP-0626-US) ;

[0041] 2016年6月7日提交的标题为“Biosensors For Biological Or Chemical Analysis And Methods Of Manufacturing The Same”的美国非临时专利申请号15/175,489(代理人案卷号IP-0689-US) ;

[0042] 2013年4月26日提交的标题为“Microdevices And Biosensor Cartridges For Biological Or Chemical Analysis And Systems And Methods For The Same”的美国非临时专利申请号13/882,088(代理人案卷号IP-0462-US) ;

[0043] 2012年9月21日提交的标题为“Methods And Compositions For Nucleic Acid Sequencing”的美国非临时专利申请号13/624,200(代理人案卷号IP-0538-US) ;

- [0044] 2011年1月13日提交的标题为“Data Processing System And Methods”的美国非临时专利申请号13/006,206;
- [0045] 2018年3月26日提交的标题为“Detection Apparatus Having A Microfluorometer, A Fluidic System, And A Flow Cell Latch Clamp Module”的美国非临时专利申请号15/936,365;
- [0046] 2019年9月11日提交的标题为“Flow Cells And Methods Related To Same”的美国非临时专利申请号16/567,224;
- [0047] 2019年6月12日提交的标题为“Device For Luminescent Imaging”的美国非临时专利申请号16/439,635;
- [0048] 2017年5月12日提交的标题为“Integrated Optoelectronic Read Head And Fluidic Cartridge Useful For Nucleic Acid Sequencing”的美国非临时专利申请号15/594,413;
- [0049] 2019年3月12日提交的标题为“Illumination For Fluorescence Imaging Using Objective Lens”的美国非临时专利申请号16/351,193;
- [0050] 2009年12月15日提交的标题为“Dynamic Autofocus Method And System For Assay Imager”的美国非临时专利申请号12/638,770;
- [0051] 2013年3月1日提交的标题为“Kinetic Exclusion Amplification Of Nucleic Acid Libraries”的美国非临时专利申请号13/783,043。

背景技术

[0052] 本部分中讨论的主题不应仅因为在本部分中有提及就被认为是现有技术。类似地,在本部分中提及的或与作为背景技术提供的主题相关联的问题不应被认为先前在现有技术中已被认识到。本部分中的主题仅表示不同的方法,这些方法本身也可对应于受权利要求书保护的技術的具体实施。

[0053] 近年来,计算能力的快速提高使得深度卷积神经网络(CNN)在许多准确度显著提高的计算机视觉任务上取得了很大的成功。在推理阶段,许多应用需要以严格的功率消耗要求对一个图像进行低等待时间处理,这降低了图形处理单元(GPU)和其他通用平台的效率,通过定制专用于深度学习算法推理的数字电路,为特定的加速硬件(例如,现场可编程门阵列(FPGA))带来了机会。然而,由于大数据量、密集型计算、变化的算法结构和频繁的存储器访问,在便携式和嵌入式系统上部署CNN仍然具有挑战性。

[0054] 由于卷积在CNN中贡献了大部分运算,因此卷积加速方案显著影响了硬件CNN加速器的效率和性能。卷积涉及具有沿内核和特征图滑动的四个循环级的乘法和累加(MAC)运算。第一循环级计算内核窗口内的像素的MAC。第二循环级跨不同的输入特征图累加MAC的乘积之和。在完成第一循环级和第二循环级之后,通过添加偏置来获得最终的输出像素。第三循环级在输入特征图内滑动内核窗口。第四循环级生成不同的输出特征图。

[0055] FPGA由于其(1)高度可重构性,(2)与专用集成电路(ASIC)相比开发时间更快,以跟上CNN的快速发展,(3)良好的性能,以及(4)与GPU相比优越的能量效率,获得了越来越多的关注和普及,特别是在加速推理任务方面。FPGA的高性能和高效率可以通过合成针对特定计算定制的电路来实现,以利用定制的存储器系统直接处理数十亿次运算。例如,现代

FPGA上的数百至数千个数字信号处理 (DSP) 块以高并行性支持核心卷积运算,例如乘法和加法。外部片上存储器和片上处理引擎 (PE) 之间的专用数据缓冲器可被设计成通过在FPGA芯片上配置数十兆字节的片上块随机存取存储器 (BRAM) 来实现优选的数据流。

[0056] 需要高效的数据流和CNN加速的硬件架构来最小化数据通信,同时最大化资源利用来实现高性能。因此有机会设计出在具有高性能、高效率 and 高度灵活性的加速硬件上加速各种CNN算法的推理过程的方法和框架。

[0057] 下一代测序 (NGS) 技术的关键特征为并行化,并且若干测序平台的主要机制为边合成边测序 (SBS)。简而言之,通过顺序地建立单链DNA模板的互补碱基并且通过在一系列原始荧光图像中捕获合成信息来同时对数千万至数亿个随机DNA片段进行测序。

[0058] 从图像数据提取实际序列信息 (即, {A,C,G,T} 中的字符串) 涉及两个计算任务,即图像分析和碱基调用。图像分析的主要功能是将图像数据转化为每个DNA片段的荧光强度数据,而碱基调用的目标是从获得的强度数据推断序列信息。

[0059] 存在多个随机的和上下文的变化源,这些变化源可能降低碱基调用准确性。例如,碱基调用中的k-mer偏差受测序基因组的GC含量影响。当应用于DNA的富含GC区时,碱基调用器可能表现出偏差,这主要是由于降低的序列复杂性,而且也是由于扩增步骤期间聚合酶链反应 (PCR) 偏差的结果。

[0060] 碱基调用的准确性对于各种下游应用 (包括序列组装、SNP调用和基因型调用) 至关重要。改进碱基调用准确性能够实现具有较小测序覆盖范围的下游应用的期望性能,这转化为测序成本的降低。

[0061] 训练用于碱基调用的神经网络需要大量计算机存储器,这随着图像大小和数量的增加呈指数增长。计算机存储器成为限制因素,因为用于优化深度神经网络的反向传播算法需要存储中间激活。由于这些中间激活的大小和数量与输入大小和数量成比例地增加,因此存储器很快填充有更大且更多的图像。

[0062] 根据一个具体实施,使用神经网络 (例如,在共同拥有的专利申请号16/826,126; 16/826,134; 16/826,168; 17/175,546; 17/180,542; 17/176,151; 63/072,032; 63/161,880; 63/161,896; 17/176,147; 63/228,954; 17/179,395; 17/180,480; 17/180,513; 63/169,163 和 63/217,644 中公开的神经网络) 的碱基调用器使用用于测序循环的滑动窗口的图像数据来进行碱基调用预测。增加滑动窗口的大小以包括来自更多测序循环的图像数据将增加神经网络的复杂性且还增加对可用计算和存储器的附加负担。

[0063] 出现了配置碱基调用操作以并入来自多个过去测序循环的上下文信息的机会。可以产生具有降低的错误率的更准确的碱基调用,特别是对于衰减k-mer偏差。

附图说明

[0064] 在附图中,在所有不同视图中,类似的参考符号通常是指类似的部件。另外,附图未必按比例绘制,而是重点说明所公开的技术的原理。在以下描述中,参考以下附图描述了本发明所公开的技术的各种实施方式,其中:

[0065] 图1为所公开的基于状态的碱基调用的高级图。

[0066] 图2示出使用基于测序图像中的像素的历史强度值而生成的状态信息来执行碱基调用操作的一个具体实施。

[0067] 图3示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每像素状态的一个具体实施。

[0068] 图4示出基于每像素状态而生成碱基调用的一个具体实施。

[0069] 图5示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道状态的一个具体实施。

[0070] 图6示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MIN(最小)状态的一个具体实施。

[0071] 图7示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成泛通道MIN(最小)状态的一个具体实施。

[0072] 图8示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MAX(最大)状态的一个具体实施。

[0073] 图9示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成泛通道MAX(最大)状态的一个具体实施。

[0074] 图10示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道AVG(平均)状态的一个具体实施。

[0075] 图11示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成泛通道AVG(平均)状态的一个具体实施。

[0076] 图12示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MIN MAX(最小和最大)状态1200的一个具体实施。

[0077] 图13示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MIN AVG(最小和平均)状态1300的一个具体实施。

[0078] 图14示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MAX AVG(最大和平均)状态1400的一个具体实施。

[0079] 图15示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MIN MAX AVG(最小、最大和平均)状态的一个具体实施。

[0080] 图16示出根据先前调用的碱基生成状态信息以供用于未来碱基调用的一个具体实施。

[0081] 图17A、图17B和图17C示出针对当前测序循环十一(循环11)的碱基调用引导的开和关状态生成。

[0082] 图18A、图18B和图18C对图17A、图17B和图17C进行了扩展,并且示出关状态如何在下一测序循环十二(循环12)处更新。

[0083] 图19A、图19B和图19C对图18A、图18B和图18C进行了进一步扩展,并且示出开状态如何在又下一测序循环十三(循环13)处更新。

[0084] 图20A、图20B和图20C对图19A、图19B和图19C进行了进一步扩展,并且示出开状态和关状态两者如何在又下一测序循环十四(循环14)处更新。

[0085] 图21示出使用指数加权平均来生成用于碱基调用的状态信息的一个具体实施。

[0086] 图22示出使用指数加权平均来生成用于碱基调用的每通道状态信息的一个具体实施。

[0087] 图23描绘使用指数加权平均来生成用于碱基调用的每通道状态信息的示例。

[0088] 图24示出使用由先前调用的碱基引导的指数加权平均来生成用于碱基调用的每通道状态信息的一个具体实施。

[0089] 图25描绘使用由先前调用的碱基引导的指数加权平均来生成用于碱基调用的每通道状态信息的示例。

[0090] 图26示出使用先前调用的碱基来针对非簇像素生成状态信息的一个具体实施。

[0091] 图27示出生成用于碱基调用的状态数据的不同具体实施。

[0092] 图28示出在碱基调用操作的各种处理方面处馈送状态数据的不同具体实施。

[0093] 图29示出基于状态和基于神经网络的碱基调用的一个具体实施。

[0094] 图30示出将每簇状态数据用于碱基调用簇的一个具体实施。

[0095] 图31示出使用对应簇像素的历史强度值来生成每簇状态的一个具体实施。

[0096] 图32示出使用对应于簇像素的空间卷积特征的历史特征值来生成每簇状态的一个具体实施。

[0097] 图33示出通过内插像素强度来生成簇强度并且使用经内插簇强度来生成用于碱基调用的每簇状态的一个具体实施。

[0098] 图34示出使用与基于神经网络的碱基调用器分开的实时分析 (RTA) 碱基调用器来生成每簇状态的一个具体实施。

[0099] 图35示出生成压缩特征的压缩逻辑的一个具体实施。

[0100] 图36示出使用对应于簇像素的压缩空间卷积特征的历史特征值来生成每簇状态的一个具体实施。

[0101] 图37示出使用与基于神经网络的碱基调用器分开的RTA碱基调用器来生成每簇状态以用于将每簇状态与压缩特征合并的一个具体实施。

[0102] 图38示出使用对应簇像素的历史强度值来生成每簇状态以用于将每簇状态与压缩特征合并的一个具体实施。

[0103] 图39示出从稀疏的每孔状态生成密集的每像素状态的一个具体实施。

[0104] 图40示出提供稀疏的每孔状态、密集的每像素状态和每像素强度值作为到碱基调用器的输入以用于执行碱基调用操作的一个具体实施。

[0105] 图41示出使用稀疏的每孔状态、密集的每像素状态和每像素强度值作为到基于神经网络的碱基调用器的输入以用于执行碱基调用操作的一个具体实施。

[0106] 图42A和图42B描绘包括可配置处理器的测序系统的一个具体实施。

[0107] 图42C是用于分析来自测序系统的传感器数据 (诸如, 碱基调用传感器输出) 的系统的简化框图。

[0108] 图43A是示出碱基调用操作的各方面的简化图, 该碱基调用操作包括由主机处理器执行的运行时程序的功能。

[0109] 图43B是可配置处理器的配置的简化图。

[0110] 图44示出确定CPU上的状态数据并且将状态数据从CPU加载到FPGA以用于碱基调用的一个具体实施。

[0111] 图45比较了非基于神经网络的碱基调用器RTA、无状态信息的基于神经网络的碱基调用器DeepRTA (5ci_k14) 以及所公开的使用每像素MIN状态信息作为附加/补充输入的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC) 的碱基调用性能。

[0112] 图46比较了所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC) 跨不同状态通道的碱基调用性能。

[0113] 图47示出所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (DC) 对k-mer (特别是5mer) 的碱基调用性能。5mer是指五个碱基位置的重复碱基模式 (例如, ACGCG、GGGGG、TCGCG)。

[0114] 图48比较了非基于神经网络的碱基调用器RTA、无状态信息的基于神经网络的碱基调用器DeepRTA (5ci_k14)、所公开的使用每像素MIN状态信息作为附加/补充输入的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN) 以及所公开的使用每像素MIN和MAX状态信息作为附加/补充输入的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 的碱基调用性能。

[0115] 图49比较了非基于神经网络的碱基调用器RTA、无状态信息和5循环侧翼输入窗口的基于神经网络的碱基调用器DeepRTA (5ci_k14)、所公开的使用每像素MIN状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN)、所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同3循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (3ci_k14_DC_MIN_MAX) 以及所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 的碱基调用性能。

[0116] 图50将所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 对k-mer (特别是5mer) 的碱基调用性能与无状态信息和5循环侧翼输入窗口的基于神经网络的碱基调用器DeepRTA (5ci_k14) 的碱基调用性能进行比较。5mer是指五个碱基位置的重复碱基模式 (例如, ACGCG、GGGGG、TCGCG)。

[0117] 图51比较了具有均衡器具体实施的非基于神经网络的碱基调用器RTA、所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 以及所公开的使用每像素AVG状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_AVG) 的碱基调用性能。

[0118] 图52比较了具有均衡器具体实施的非基于神经网络的碱基调用器RTA以及所公开的使用每像素MIN、MAX和AVG状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX_AVG) 的碱基调用性能。

[0119] 图53比较了具有均衡器具体实施的非基于神经网络的碱基调用器RTA、所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口和大小为14 (K=14) 的滤波器组的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 以及所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口和大小为32 (K=32) 的滤波器组的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 的碱基调用性能。

[0120] 图54比较了所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA针

对指数加权平均具体实施的不同 α 参数配置(例如,0.05、0.07、0.10、0.12)的碱基调用性能,如上文所论述。

[0121] 图55和图56是根据所公开的技术的一个具体实施追踪由指数加权平均值确定的状态值的图表。

[0122] 图57从次级分析任务和度量的角度比较了具有均衡器具体实施的RTA (RTA+Eq)、无状态信息的DeepRTA (DeepRTA (k14_1m)) 以及所公开的带状态的DeepRTA (DeepRTA_extrachannels (k14_1, _onoff_expavg)), 这些次级分析任务和度量如调用读数的数量、读段1和读段1以及其平均值的读取错配错误率、单核苷酸多态性 (SNP) 重新调用、SNP精度、SNP调用准确性 (F1分数)、插入/缺失 (Indel) 重新调用、Indel精度和Indel调用准确性 (F1分数)。

具体实施方式

[0123] 呈现以下讨论以使得本领域的任何技术人员能够实现和使用所公开的技术,并且在特定应用及其要求的上下文中提供以下讨论。对所公开的具体实施的各种修改对于本领域的技术人员而言将是显而易见的,并且在不脱离所公开的技术的实质和范围的情况下,本文所定义的一般原理可以应用于其它具体实施和应用。因此,所公开的技术并非旨在限于所示的具体实施,而是要符合与本文所公开的原理和特征一致的最广范围。

[0124] 当结合附图阅读时,将更好地理解各种具体实施的详细描述。就附图例示各种具体实施的功能块的图而言,功能块不一定指示硬件电路之间的划分。因此,例如,功能块(例如,模块、处理器或存储器)中的一个或多个功能块可在单件硬件(例如,通用信号处理器或随机存取存储器块、硬盘等)或多件硬件中实现。类似地,程序可以是独立程序,可作为子例程并入操作系统中,可以是已安装软件包中的功能等。应当理解,各种具体实施不限于附图中所示的布置和工具。

[0125] 附图中被指定为模块的处理引擎和数据库可在硬件或软件中实现,并且不需要按如附图所示那样精确地划分成相同的块。这些模块中的一些模块还可在不同的处理器、计算机或服务器上实现,或者分布在多个不同的处理器、计算机或服务器之间。此外,应当理解,在不影响所实现的功能的情况下,可组合、同步操作或以与图中所示不同的序列操作模块中的一些。附图中的模块也可被认为是方法中的流程图步骤。模块也不一定需要将其所有代码连续地放置在存储器中;代码的一些部分可与代码的其他部分分离,来自其他模块或其他功能的代码设置在两者之间。

[0126] 引言

[0127] 所公开的技术涉及基于状态的碱基调用。具体地,所公开的技术涉及当生成针对当前测序循环的碱基调用时将关于来自先前测序循环的数据的状态信息并入到来自当前测序循环的数据的分析中。例如,当生成针对第N测序循环的碱基调用时,所公开的技术可以将关于来自测序循环1至N-1的数据的状态信息并入到碱基调用逻辑中。

[0128] 以下论述描述了所公开的基于状态的碱基调用的各种具体实施。具体实施根据不同的数据和处理方面而变化。例如,生成状态信息的不同“方法”或“逻辑”产生不同类型的状态。而且,其历史值被追踪以用于状态生成的“事物”可以随具体实施而变化。此外,一旦生成,“何时”、“在何处”以及“如何”处理状态信息以用于碱基调用就会产生所公开的技术

的各种具体实施。

[0129] 基于状态的碱基调用

[0130] 图1为所公开的基于状态的碱基调用100的高级图。碱基调用是确定序列的核苷酸组成的过程。在一个具体实施中,碱基调用涉及分析图像数据,即测序图像,该图像数据在由测序系统,诸如Illumina的iSeq、HiSeqX、HiSeq 3000、HiSeq 4000、HiSeq 2500、NovaSeq 6000、NextSeq550、NextSeq 1000、NextSeq 2000、NextSeqDx、MiSeq和MiSeqDx执行的测序运行(或测序反应)期间产生。在其它具体实施中,碱基调用可以涉及从非图像测序数据推断序列读数。

[0131] 测序系统104可以用于核酸的测序。可应用技术包括其中核酸附着在阵列中的固定位置(例如,流通池的孔)并且对阵列反复成像的技术。在此类具体实施中,测序系统104可以获得两个不同颜色通道中的图像,这些图像可以用于将特定核苷酸碱基类型与另一种碱基类型区分开来。更具体地,测序系统104可以实现被称为“碱基调用”的过程,该过程通常指在成像循环中针对图像的给定斑点位置确定碱基调用(例如,腺嘌呤(A)、胞嘧啶(C)、鸟嘌呤(G)或胸腺嘧啶(T))的过程。在双通道碱基调用期间,例如,从两个图像提取的图像数据可以用于通过将碱基身份编码为两个图像强度的组合来确定四种碱基类型中的一种碱基类型的存在。对于两个图像中的每个图像中的给定斑点或位置,可以基于信号标识的组合是[开,开]、[开,关]、[关,开]还是[关,关]来确定碱基身份。

[0132] 来自测序系统104的输出数据可以被传送到实时分析模块(未示出)。在各种具体实施中,实时分析模块执行用于分析图像数据(例如,图像质量评分、碱基调用等)、向图形用户界面(GUI)报告或显示光束的特性(例如,焦点、形状、强度、功率、亮度、位置)等的计算机可读指令。这些操作可以在成像循环期间实时执行,以最小化下游分析时间,并且在成像运行期间提供实时反馈和故障排除。在具体实施中,实时分析模块可以为通信地耦合到并且控制测序系统104的成像子系统的计算设备。

[0133] 以下讨论概述根据一个具体实施的生成测序图像的方法及其描绘的内容。

[0134] 在一些具体实施中,碱基调用将在测序图像中编码的强度数据解码到核苷酸序列中。在一个具体实施中,Illumina测序平台采用循环可逆终止(CRT)化学来进行碱基调用。该过程依赖于生长与具有经荧光标记的核苷酸的模板链互补的新生链,同时跟踪每个新添加的核苷酸的发射信号。经荧光标记的核苷酸具有锚定核苷酸类型的荧光团信号的3'可移除块。

[0135] 测序在重复的循环中进行,每个循环包括三个步骤:(a)通过添加经荧光标记的核苷酸延长新生链;(b)使用测序系统104的光学子系统的一个或多个激光器激发荧光团,并且通过光学子系统的不同滤波器成像,从而产生测序图像;以及(c)裂解荧光团并移除3'块,以为下一个测序循环做准备。重复掺入和成像循环直至达到指定数量的测序循环,从而限定读段长度。使用该方法,每个循环询问沿着模板链的新位置。

[0136] Illumina测序仪的巨大能力源于其同时执行和感测数百万或甚至数十亿个经历CRT反应的簇(也称为“簇”)的能力。簇包括模板链的约一千个相同副本,但簇的尺寸和形状不同。在测序运行之前,通过对输入文库进行桥式扩增或排除扩增,使来自模板链的簇生长。扩增和簇生长的目的是增加发射信号的强度,因为成像设备不能可靠地感测单个链的荧光团信号。然而,簇内的链的物理距离较小,因此成像设备将链的簇感知为单个点。

[0137] 测序发生在流通池(或生物传感器),即容纳输入链的小载玻片中。流通池连接到光学系统,该光学系统包括显微镜成像、激发激光器和荧光滤波器。流通池包括被称为槽道的多个腔室。槽道在物理上彼此分离,并且可以包含不同的带标记的测序文库,这些带标记的测序文库在无样品交叉污染的情况下可以进行区分。在一些具体实施中,该流通池包括图案化表面。“图案化表面”是指在固体载体的暴露层中或该暴露层上的不同区域的布置。

[0138] 测序系统104的成像设备(例如,固态成像器件,诸如电荷耦合器件(CCD)或互补金属氧化物半导体(CMOS)传感器)在沿槽道的多个位置、在一系列非重叠区域(被称为区块)中拍摄快照。例如,每个槽道可以存在六十四个区块或九十六个区块。区块容纳数十万至数百万的簇。

[0139] 测序运行的输出是测序图像。测序图像使用经像素化的单元(例如,像素、超像素、子像素)的网格(或阵列)描绘簇及其周围背景。强度发射被存储为经像素化的单元的强度值。测序图像具有经像素化的单元的网格的尺寸 $w \times x \times h$,其中 w (宽度)和 h (高度)为在1至100,000范围内的任何数字(例如,115×115、200×200、1800×2000、2200×25000、2800×3600、4000×400)。在一些具体实施中, w 和 h 相同。在其他具体实施中, w 和 h 不同。测序图像描绘由于在测序运行期间核苷酸掺入核苷酸序列中而生成的强度发射。强度发射来自相关联的簇及其周围背景。

[0140] 根据一个具体实施,数据流逻辑(未示出)将测序图像提供到碱基调用器144以用于碱基调用。根据一个具体实施,碱基调用器144在逐块的基础(或逐区块的基础)上访问测序图像。块中的每个块是经像素化的单元的网格中的经像素化的单元的子网格(或子阵列),该子网格形成测序图像。块具有经像素化的单元的子网格的尺寸 $q \times r$,其中 q (宽度)和 r (高度)为在1至10000范围内的任何数字(例如,3×3、5×5、7×7、10×10、15×15、25×25、64×64、78×78、115×115)。在一些具体实施中, q 和 r 相同。在其他具体实施中, q 和 r 不同。在一些具体实施中,从测序图像中提取的块具有相同的尺寸。在其他具体实施中,块具有不同的尺寸。在一些具体实施中,块可以具有重叠的经像素化的单元(例如,在边缘上)。

[0141] 根据一些具体实施,测序产生针对对应的 m 个图像通道的每测序循环 m 个测序图像。即,测序图像中的每个测序图像具有一个或多个图像(或强度)通道(类似于彩色图像的红色、绿色、蓝色(RGB)通道)。在一个具体实施中,每个图像通道对应于多个滤波器波长带中的一个滤波器波长带。在另一具体实施中,每个图像通道对应于处于测序循环的多个成像事件中的一个成像事件。在又一具体实施中,每个图像通道对应于利用特定激光器照明与通过特定光学滤波器成像的组合。从针对特定测序循环的 m 个图像通道中的每个图像通道平铺(或访问)图像块。在诸如四通道化学、双通道化学和单通道化学的不同具体实施中, m 为4或2。在其他具体实施中, m 为1、3或大于4。在其它具体实施中,图像可以在蓝色通道和紫色通道中,而不是在红色通道和绿色通道中,或者除了红色和绿色颜色通道之外。

[0142] 例如,考虑测序运行是使用两个不同的图像通道(蓝色通道和绿色通道)来实施的。然后,在每个测序循环,测序运行产生蓝色图像和绿色图像。这样,针对测序运行的一系列的 k 个测序循环,产生 k 对蓝色和绿色图像的序列作为输出,并将其存储为测序图像。因此,由碱基调用器144生成用于块级处理的 k 对蓝色和绿色图像块的序列。

[0143] 根据其它具体实施,测序系统104还可以生成非图像测序数据。在一个具体实施中,测序数据可以基于在分子延伸期间由氢离子的释放所诱导的pH变化。可以检测pH变化

并将其转换为与掺入的碱基的数量成比例的电压变化。在又一个具体实施中,测序数据可以例如根据纳米孔感测构造而成,该纳米孔感测使用生物传感器来测量当簇穿过纳米孔或靠近其孔口时电流的中断,同时确定碱基的种类。在一个具体实施中,基于纳米孔的测序可以基于以下概念:使单链DNA(或RNA)经由纳米孔穿过膜,并且跨膜施加电压差。孔中存在的核苷酸可以影响孔的电阻,因此随时间推移的电流测量结果可以指示DNA碱基穿过孔的序列。该电流信号(由于其在绘制时的外观而被称为“波形曲线(squiggle)”)是由测序仪收集的原始数据。这些测量结果可以被存储为在(例如)4kHz频率下获得的16位整数数据采集(DAC)值。在DNA链速度为约450碱基对/秒的情况下,这可以给出平均每种碱基大约九个原始观察结果。然后可以处理该信号以识别对应于各个读数的开孔信号的中断。对原始信号的这些最大限度的利用可以为进行碱基调用,即将DAC值转换为DNA碱基序列的过程。在一些具体实施中,测序数据可以包括归一化或缩放的DAC值。

[0144] 当前测序数据112包括针对测序运行的当前测序循环由测序数据104生成的测序数据。在图1中,当前测序循环被识别为“N”测序循环。

[0145] 先前测序数据116包括针对测序运行的一个或多个先前测序循环由测序数据104生成的测序数据。先前测序循环先于当前测序循环。在图1中,先前测序循环被识别为“1至N-1”测序循环。在其它具体实施中,先前测序数据116包括针对1至N-1测序循环的子集的测序数据。

[0146] 状态发生器126使用当前测序数据112和先前测序数据116来生成针对当前测序循环的当前状态数据136。状态发生器126可以为应用于测序数据以生成期望结果的值或函数。状态发生器126可以通过多种数学操纵中的任一种数学操作应用于测序数据,这些数学操作包括但不限于加法、减法、除法、乘法或它们的组合。状态发生器126可以为数学公式、逻辑函数、计算机实现的算法等。测序数据可以为图像数据、电数据或它们的组合。

[0147] 在一个具体实施中,状态发生器126通过累加当前测序数据112和先前测序数据116的概括统计值来生成当前状态数据136。概括统计值的示例包括最大值、最小值、平均值(均值)、指数加权平均值、移动(运行)平均值、指数移动平均值、众数、标准偏差、方差、偏度、尖峰值、百分位和熵。在其它具体实施中,状态发生器126基于概括统计值而确定中等统计值。中等统计的示例包括增量、总和、一系列最大值、一系列最小值、系列中最大值中的最小值以及系列中最小值中的最大值。

[0148] 碱基调用器144响应于处理当前测序数据112和当前状态数据136而生成针对当前测序循环的当前碱基调用数据154。当前碱基调用数据154可以包括针对一个或多个簇的碱基调用。在一些具体实施中,当前测序数据112和当前状态数据136在由碱基调用器144处理之前被组合。该组合可以通过例如求和运算、逐元素乘法运算、逐元素乘法和求和(卷积)运算以及级联运算来实现。

[0149] 碱基调用器144的示例包括可用于Illumina平台的不同的碱基调用程序,诸如实时分析(RTA)、BlindCall、freeIbis、Softy、AYB、OnlineCall、BM-BC、ParticleCall、TotalReCaller、naiveBayesCall、Srfim、BayesCall、Ibis、Rolexa、Alta-Cyclic和Bustard。碱基调用器144的示例还包括Illumina的基于神经网络的产品,诸如在共同拥有的专利申请号16/825,987;16/825,991;16/826,126;16/826,134;16/826,134;16/826,168;17/175,546;17/180,542;17/176,151;63/072,032;63/161,880;63/161,896;17/176,

147;63/228,954;17/179,395;17/180,480;17/180,513;63/169,163;和63/217,644中公开的产品,在本文中统称为“DeepRTA”或“深度学习初级分析”。碱基调用器144的又其它示例包括可用于牛津纳米孔技术(ONT)的不同碱基调用程序,诸如Metrichor、Nanocall、DeepNano、Nanonet、Scrappie、Albacore、Guppy、Basecaller、Chiron、Halcyon、MinCall、SACall、Causalcall和WaveNano。

[0150] 像素化状态数据

[0151] 图2示出使用基于测序图像中的像素的历史强度值而生成的状态信息200来执行碱基调用操作的一个具体实施。

[0152] 在动作202处,存储器针对测序运行的当前测序循环存储(例如,在数据存储库112中)多个像素中的相应像素的相应当前强度值。在动作212处,存储器针对测序运行的在当前测序循环之前的一个或多个先前测序循环存储(例如,在数据存储库116中)相应像素的相应先前强度值的序列。在一个具体实施中,相应像素的相应强度值各自由多个通道的通道特异性强度值表征。在一个具体实施中,其中通道对应于利用特定激光器照明和通过特定光学滤波器成像的组合。在另一个具体实施中,通道对应于滤波器波长带。在又一个具体实施中,通道对应于在测序循环处的成像事件。

[0153] 在动作222处,对存储器具有访问权限的状态发生器126根据(i)相应当前强度值以及(ii)相应先前强度值的序列中的相应先前强度值来生成相应像素的相应当前状态值。而且,状态发生器126将相应当前状态值存储在存储器中。在一些具体实施中,相应像素的相应状态值各自由多个通道的通道特异性状态值表征。在一个具体实施中,相应当前状态值是利用相应当前强度值进行逐像素编码的。在一些具体实施中,多个通道中的通道子集的通道特异性状态值是利用相应当前强度值进行逐像素编码的。在一些具体实施中,通道特异性状态值跨多个通道中的通道被平均以生成相应泛通道当前状态值。在一个具体实施中,相应泛通道当前状态值是利用相应当前强度值进行逐像素编码的。

[0154] 在一些具体实施中,逐像素编码包括逐像素级联。在其它具体实施中,逐像素编码包括逐像素求和。在又其它具体实施中,逐像素编码包括逐元素乘法。在又其它具体实施中,逐像素编码包括逐元素乘法和求和(卷积)。

[0155] 在一个具体实施中,相应状态值被配置为表征相应像素的历史强度模式。在另一个具体实施中,历史强度模式被配置为补偿碱基调用器144的碱基调用准确性的损失。在又一个具体实施中,历史强度模式被配置为当针对k-mer的碱基进行碱基调用时补偿碱基调用器144的碱基调用准确性的损失。在又一个具体实施中,相应状态值被配置为辨别簇的相应信号分布。

[0156] 在一个具体实施中,相应当前状态值为从(i)相应先前强度值以及(ii)相应当前强度值在当前测序循环处针对相应像素确定的相应当前平均强度。在一些具体实施中,相应当前平均强度各自由通道特异性当前平均强度表征。

[0157] 在另一个具体实施中,相应当前状态值为从(i)相应先前强度值以及(ii)相应当前强度值在当前测序循环处针对相应像素确定的相应当前最大强度。在一些具体实施中,相应当前最大强度各自由通道特异性当前最大强度表征。

[0158] 在又一个具体实施中,相应当前状态值为从(i)相应先前强度值以及(ii)相应当前强度值在当前测序循环处针对相应像素确定的相应当前最大强度。在一些具体实施中,

相应当前最大强度各自由通道特异性当前最大强度表征。

[0159] 在又一个具体实施中,相应当前状态值为从(i)相应先前强度值以及(ii)相应当前强度值在当前测序循环处针对相应像素确定的相应当前最小强度。在一些具体实施中,相应当前最小强度各自由通道特异性当前最小强度表征。

[0160] 在又一个具体实施中,相应当前状态值为从(i)相应先前强度值以及(ii)相应当前强度值在当前测序循环处针对相应像素确定的相应当前指数加权平均强度。在一些具体实施中,相应当前指数加权平均强度是基于对最近测序循环比对之前测序循环进行更多加权而确定的。在其它具体实施中,相应当前指数加权平均强度各自由通道特异性当前指数加权平均强度表征。

[0161] 在又一个具体实施中,相应当前状态值为从(i)相应当前强度值以及(ii)相应先前强度值的滚动子集在当前测序循环处针对相应像素确定的相应当前移动平均强度。在一些具体实施中,相应当前移动平均强度各自由通道特异性当前移动平均强度表征。

[0162] 在一些具体实施中,相应当前强度值中的每个相应当前强度值基于相应当前强度值与相应当前启用状态强度和相应当前非启用状态强度的比较而归属于启用状态桶或非启用状态桶。在一个具体实施中,相应当前启用状态强度为从(i)相应先前强度值在当前测序循环处针对相应像素确定的相应当前全局最大强度。在一些具体实施中,相应当前全局最大强度各自由通道特异性当前全局最大强度表征。

[0163] 在一个具体实施中,相应当前非启用状态强度为从(i)相应先前强度值在当前测序循环处针对相应像素确定的相应当前全局最小强度。在一些具体实施中,相应当前全局最小强度各自由通道特异性当前全局最小强度表征。

[0164] 在一个具体实施中,相应当前状态值还包括在当前测序循环处针对相应像素生成的相应当前启用状态值和相应当前非启用状态值。在一些具体实施中,相应当前启用状态值各自由通道特异性当前启用状态值表征。在其它具体实施中,相应当前非启用状态值各自由通道特异性当前非启用状态值表征。

[0165] 在一个具体实施中,目标通道中的像素的当前启用状态值是在当前测序循环处从(i)目标通道中的像素的在当前测序循环处被检测到并且归属于启用状态桶的当前强度值以及(ii)目标通道中的像素的在先前测序循环处被检测到的先前强度值确定的。

[0166] 在一些具体实施中,当前强度值基于当前强度值与从目标通道中的先前强度值确定的全局最大值和全局最小值的比较而归属于启用状态桶。在其它具体实施中,当前强度值基于当前强度值与在目标通道中的先前测序循环处确定的先前启用状态值和先前非启用状态值的比较而归属于启用状态桶。在一个具体实施中,当前启用状态值为从当前强度值和先前强度值确定的指数加权平均值。在另一个具体实施中,当前启用状态值为从当前强度值和先前强度值确定的平均值。在又一个具体实施中,当前启用状态值为从当前强度值和先前强度值的滚动子集确定的移动平均值。在又一个具体实施中,当前启用状态值为从当前强度值和先前强度值确定的最小值。在又一个具体实施中,当前启用状态值为从当前强度值和先前强度值确定的最大值。

[0167] 在一个具体实施中,当前启用状态值是从先前测序循环携带的,并且当当前强度值在当前测序循环处归属于非启用状态桶时不在当前测序循环处重新确定。

[0168] 在一个具体实施中,目标通道中的像素的当前非启用状态值是在当前测序循环处

从 (i) 目标通道中的像素的在当前测序循环处被检测到并且归属于非启用状态桶的当前强度值以及 (ii) 目标通道中的像素的在先前测序循环处被检测到的先前强度值确定的。

[0169] 在一些具体实施中,当前强度值基于当前强度值与从目标通道中的先前强度值确定的全局最大值和全局最小值的比较而归属于非启用状态桶。在其它具体实施中,当前强度值基于当前强度值与在目标通道中的先前测序循环处确定的先前非启用状态值和先前非启用状态值的比较而归属于非启用状态桶。

[0170] 在一些具体实施中,当当前强度值与全局最小值的接近度大于与全局最大值的接近度时,当前强度值归属于非启用状态桶。在一些具体实施中,当当前强度值与全局最大值的接近度大于与全局最小值的接近度时,当前强度值归属于启用状态桶。

[0171] 在一个具体实施中,当前非启用状态值为从当前强度值和先前强度值确定的指数加权平均值。在另一个具体实施中,当前非启用状态值为从当前强度值和先前强度值确定的平均值。在又一个具体实施中,当前非启用状态值为从当前强度值和先前强度值的滚动子集确定的移动平均值。在又一个具体实施中,当前非启用状态值为从当前强度值和先前强度值确定的最小值。在又一个具体实施中,当前非启用状态值为从当前强度值和先前强度值确定的最大值。

[0172] 在一些具体实施中,当前非启用状态值是从先前测序循环携带的,并且当当前强度值在当前测序循环处归属于启用状态桶时不在当前测序循环处重新确定。

[0173] 在动作232处,对存储器具有访问权限的碱基调用器144响应于处理 (i) 相应当前强度值以及 (ii) 相应当前状态值而生成针对当前测序循环的碱基调用。在一些具体实施中,针对当前测序循环的碱基调用包括针对检测到相应当前信号值和相应先前信号值的一个或多个簇的碱基调用。

[0174] 在一个具体实施中,碱基调用器144为神经网络。在一些具体实施中,神经网络为卷积神经网络。在一个具体实施中,卷积神经网络包括多个空间卷积层和多个时间卷积层。

[0175] 在一些具体实施中,使用针对已经执行的测序运行离线生成的测序图像来训练碱基调用器144。在一个具体实施中,在由碱基调用器144进行碱基调用之前,针对已经执行的测序运行的每个测序循环离线计算相应像素的相应状态值。在一个具体实施中,碱基调用器144被训练为使用相应状态值来补偿碱基调用准确性的损失。

[0176] 在一些具体实施中,在测序运行的每个测序循环处迭代地生成相应当前状态值。在一些具体实施中,存储器被进一步配置为针对测序运行的下一测序循环存储相应像素的相应下一强度值的序列。在一些具体实施中,碱基调用器144被进一步配置为响应于处理 (i) 相应当前强度值、(ii) 针对先前测序循环中的一个或多个先前测序循环的相应先前强度值的序列中的相应先前强度值、(iii) 针对下一测序循环中的一个或多个下一测序循环的相应下一强度值的序列中的相应下一强度值以及 (iv) 相应状态值而生成针对当前测序循环的碱基调用。在一些具体实施中,相应状态值是利用相应先前强度值和相应下一强度值进行逐像素编码的。

[0177] 每像素状态

[0178] 图3示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每像素状态300的一个具体实施。图3描绘大小为 3×3 的示例性像素网格302N、302N-1、……、302。图3还示出每像素状态网格304。像素网格302N、302N-1、……、302表示被提供作为碱基调用器

144的输入以用于碱基调用的测序图像。像素网格302N、302N-1、……、302中的像素可以具有多个图像通道(例如,红色通道和蓝色通道)的强度值。类似地,像素网格302N是针对测序运行的当前测序循环N生成的,并且描绘由于在当前测序循环N处的簇集合中的核苷酸掺入而生成的强度发射。类似地,像素网格302N-1是针对测序运行的先前测序循环N-1生成的,并且描绘由于在先前测序循环N-1处的簇集合中的核苷酸掺入而生成的强度发射。类似地,像素网格302是针对测序运行的第一测序循环1生成的,并且描绘由于在第一测序循环1处的簇集合中的核苷酸掺入而生成的强度发射。

[0179] 从空间角度来看,像素网格302N、302N-1、……、302可以被认为共享九个像素1-9。在图3中,九个像素1-9由跨像素网格302N、302N-1、……、302的相同的每像素上标索引。从时间角度来看,九个像素1-9可以被认为描绘了当前时间步长N处的第一组强度值、先前时间步长N-1处的第二组强度值以及第一时间步长1处的第三组强度值。第一组强度值、第二组强度值和第三组强度值中的强度值可以被认为是按时间序列布置的。在图3中,时变强度值由跨像素网格302N、302N-1、……、302的不同的每像素上标索引。

[0180] 可以基于给定像素在当前测序N处的当前强度值312和给定像素在相应先前测序循环N-1至1处的先前强度值314、……、316而计算给定像素在当前测序N处的当前像素状态318。给定像素的先前强度值314、……、316可以被认为布置在开始于先前强度位置N-1并且终止于先前强度位置1的先前强度序列中。

[0181] 在一个具体实施中,当前像素状态318为将当前强度值312和先前强度值314、……、316作为输入的状态生成逻辑(或函数)的输出。在一个具体实施中,状态生成逻辑可以从当前强度值312和先前强度值314、……、316中选择最大值,并且使用该最大值作为当前像素状态318。在另一个具体实施中,状态生成逻辑可以从当前强度值312和先前强度值314、……、316中选择最小值,并且使用该最小值作为当前像素状态318。在又一个具体实施中,状态生成逻辑可以确定当前强度值312和先前强度值314、……、316的平均值(均值),并且使用该平均值作为当前像素状态318。在又一个具体实施中,状态生成逻辑可以确定当前强度值312和先前强度值314、……、316的指数加权平均值,并且使用该指数加权平均值作为当前像素状态318。在又一个具体实施中,状态生成逻辑可以确定当前强度值312和先前强度值314、……、316的移动(运行)平均值,并且使用该移动平均值作为当前像素状态318。在又一个具体实施中,状态生成逻辑可以确定当前强度值312和先前强度值314、……、316的指数移动平均值,并且使用该指数移动平均值作为当前像素状态318。在又一个具体实施中,状态生成逻辑可以确定当前强度值312和先前强度值314、……、316的标准偏差,并且使用该标准偏差作为当前像素状态318。对九个像素1-9中的每个像素实现该过程,使得针对九个像素1-9中的每个像素确定相应当前像素状态。根据图3所示的示例,由状态生成逻辑针对九个像素1-9生成的当前像素状态形成每像素状态网络304。在一个具体实施中,状态生成逻辑由状态发生器126实现。

[0182] 逐像素状态编码

[0183] 图4示出基于每像素状态304而生成碱基调用的一个具体实施。在一个具体实施中,编码逻辑402例如通过求和、级联、逐元素乘法、或逐元素乘法和求和(卷积)来逐像素地组合400像素网格302N和每像素状态网络304。像素网格302N和每像素状态网络304的组合由碱基调用器144处理以生成当前碱基调用数据154。

[0184] 每通道状态

[0185] 图5示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道状态500的一个具体实施。图5涉及示例性像素P。在图5中,像素P具有跨两个示例性通道(或强度通道)的强度值,该两个示例性通道为:通道1 502和通道2 506。在其它具体实施中,像素P可以具有更少或更多数量的通道。在图5所示的示例中,第一通道502具有针对五个测序循环1-5的第一组强度值512、513、……、516。同样,第二通道506具有针对五个测序循环1-5的第二组强度值522、523、……、526。

[0186] 然后,基于第一组强度值而针对第一通道502确定第一通道状态532。类似地,基于第二组强度值而针对第二通道506确定第二通道状态536。第一通道状态532和第二通道状态536可以由状态生成逻辑例如通过实现最小值选择函数、最大值选择函数、平均函数、指数加权平均函数、移动平均函数、指数移动平均函数或标准偏差函数来确定。

[0187] 然后,基于第一通道状态532和第二通道状态536而针对第五测序循环(循环5)生成至少一个碱基调用。在一个具体实施中,这包括将针对循环5的通道1强度值516与第一通道状态532组合以生成通道1 502的第一组合,以及将针对循环5的通道2强度值526与第二通道状态536组合以生成通道2 506的第二组合。然后,碱基调用器144处理针对通道1 502的第一组合和针对通道2 506的第二组合,以生成针对循环5(例如,针对一个或多个簇)的一个或多个碱基调用。

[0188] 在另一个具体实施中,碱基调用器144将针对循环5的通道1强度值516、针对循环5的通道2强度值526、针对循环5的第一通道状态532和针对循环5的第二通道状态536处理为四个独立输入通道(例如,RGB样式图像通道),并且生成针对循环5的碱基调用。所属领域的技术人员应当了解,在本公开的范围内可以类似地应用且等同地使用组合数据/通道且将其处理为组合的其它当前或未来方法。

[0189] 在又一个具体实施中,例如通过对第一通道状态532和第二通道状态536应用平均函数来组合第一通道状态532和第二通道状态536以生成所谓的“泛通道状态”。然后,针对循环5的该泛通道状态与针对循环5的通道1强度值516和针对循环5的通道2强度值526一起被处理,以生成针对循环5的碱基调用,即,与每通道具体实施中的四个独立输入通道相比,总共三个独立输入通道。

[0190] 尽管图5出于示出和简化的目的仅涉及一个像素P,但应当理解,针对输入测序图像中的多个像素在逐像素和逐通道的基础上实现状态生成和状态编码响应碱基调用的步骤,这部分地是因为簇强度分布和其状态由连续像素的群组限定且因此作为群组进行分析以生成针对簇的碱基调用。

[0191] 每通道MIN状态

[0192] 图6示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MIN(最小)状态600的一个具体实施。在图6中,列601索引由碱基调用器144处理以用于碱基调用的输入测序图像的多个像素。输入测序图像表示在测序运行的当前测序循环N处捕获的一组簇的强度分布。在图6中,列602表示第一通道(图像通道)中针对列601中所索引的像素且针对当前测序循环N的逐像素强度值。在图6中,列603表示第二通道(图像通道)中针对列601中所索引的像素且针对当前测序循环N的逐像素强度值。

[0193] 在图6中,列604表示第一通道中的逐像素最小强度值,这些逐像素最小强度值选

自在第一测序循环1与当前测序循环N之间在第一通道中针对列601中所索引的像素观察到的强度值。在图6中,列605表示第二通道中的逐像素最小强度值,这些逐像素最小强度值选自在第一测序循环1与当前测序循环N之间在第二通道中针对列601中所索引的像素观察到的强度值。

[0194] 在一个具体实施中,碱基调用器144将列602、603、604和605处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0195] 泛通道MIN状态

[0196] 图7示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成泛通道MIN(最小)状态700的一个具体实施。在图7中,列708逐像素地对图6的列604和605求平均。在一个具体实施中,碱基调用器144将列602、603和708处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0197] 每通道MAX状态

[0198] 图8示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MAX(最大)状态800的一个具体实施。在图8中,列804表示第一通道中的逐像素最大强度值,这些逐像素最大强度值选自在第一测序循环1与当前测序循环N之间在第一通道中针对列601中所索引的像素观察到的强度值。在图8中,列805表示第二通道中的逐像素最大强度值,这些逐像素最大强度值选自在第一测序循环1与当前测序循环N之间在第二通道中针对列601中所索引的像素观察到的强度值。

[0199] 在一个具体实施中,碱基调用器144将列602、603、804和805处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0200] 泛通道MAX状态

[0201] 图9示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成泛通道MAX(最大)状态900的一个具体实施。在图9中,列908逐像素地对图8的列804和805求平均。在一个具体实施中,碱基调用器144将列602、603和908处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0202] 每通道AVG状态

[0203] 图10示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道AVG(平均)状态1000的一个具体实施。在图10中,列1004表示第一通道中的逐像素平均强度值,这些逐像素平均强度值被计算为在第一测序循环1与当前测序循环N之间在第一通道中针对列601中所索引的像素观察到的强度值的相应每像素均值。在图10中,列1005表示第二通道中的逐像素平均强度值,这些逐像素平均强度值被计算为在第一测序循环1与当前测序循环N之间在第二通道中针对列601中所索引的像素观察到的强度值的相应每像素均值。

[0204] 在一个具体实施中,碱基调用器144将列602、603、1004和1005处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0205] 泛通道AVG状态

[0206] 图11示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成泛通道AVG(平均)状态1100的一个具体实施。在图11中,列1108逐像素地对图10的列1004和1005求平均。在一个具体实施中,碱基调用器144将列602、603和1108处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0207] 每通道MIN MAX状态

[0208] 图12示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MIN MAX (最小和最大) 状态1200的一个具体实施。在一个具体实施中,碱基调用器144将列602、603、604、605、804和805处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0209] 每通道MIN AVG状态

[0210] 图13示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MIN AVG (最小和平均) 状态1300的一个具体实施。在一个具体实施中,碱基调用器144将列602、603、604、605、1004和1005处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0211] 每通道MAX AVG状态

[0212] 图14示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MAX AVG (最大和平均) 状态1400的一个具体实施。在一个具体实施中,碱基调用器144将列602、603、804、805、1004和1005处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0213] 每通道MIN MAX AVG状态

[0214] 图15示出针对被处理作为用于碱基调用的输入的测序图像中的像素生成每通道MIN MAX AVG (最小、最大和平均) 状态1500的一个具体实施。在一个具体实施中,碱基调用器144将列602、603、604、605、804、805、1004和1005处理为独立输入通道,以生成针对当前测序循环N的一个或多个碱基调用。

[0215] 所属领域的技术人员应当了解,可以提供上文所论述的通道和状态的任何组合或排序或布置作为到碱基调用器144的输入以用于碱基调用。组合、排序和布置上文所论述的通道和状态的其它当前和未来方式在本公开的范围之内。例如,可以通过使用其它聚合/累加函数(如指数加权平均、移动平均、标准偏差、方差等)组合每通道状态来生成泛通道状态。在另一个示例中,每通道MIN状态可以与泛通道AVG MAX状态级联,等等。

[0216] 碱基调用引导的状态生成

[0217] 图16示出根据先前调用的碱基生成1600状态信息以供用于未来碱基调用的一个具体实施。图17A、图17B和图17C示出针对当前测序循环十一(循环11)的碱基调用引导的开和关状态生成1700。图18A、图18B和图18C对图17A、图17B和图17C进行了扩展,并且示出关状态如何在下一测序循环十二(循环12)处更新1800。图19A、图19B和图19C对图18A、图18B和图18C进行了进一步扩展,并且示出开状态如何在又下一测序循环十三(循环13)处更新1900。图20A、图20B和图20C对图19A、图19B和图19C进行了进一步扩展,并且示出开状态和关状态两者如何在又下一测序循环十四(循环14)处更新2000。

[0218] 区分在测序反应中使用一种荧光染料(或相同或类似的激发/发射光谱的两种或更多种染料)检测核苷酸掺入的不同策略的一种途径是通过在测序循环期间发生的荧光转变的存在或相对不存在或在中间的水平方面表征该掺入。因此,测序策略可以通过其在测序循环中的荧光分布来说明。对于本文所公开的策略,“1”和“0”表示荧光状态,其中,核苷酸处于信号状态(例如,可通过荧光检测的)(1)/(开)或核苷酸是否处于暗态(例如,在成像步骤中未检测到,或最低限度地检测到)(0)/(关)。“0”、“关”或“暗”状态不一定是指信号完

全缺失或不存在。尽管在一些具体实施中可能存在信号(例如,荧光)的完全缺失或不存在,但在其它具体实施中,即使在关状态下也可能存在一些可检测信号。只要可以可靠地区分从第一图像到第二图像(或反之亦然)的荧光变化,最低限度的或减弱的荧光信号(例如,背景信号)也被预期包括在“0”、“关”或“暗”状态的范围内。

[0219] 如本文所用,术语“暗”或“关”旨在指由检测器检测到的期望信号的量与由检测器检测到的背景信号相比不显著。例如,当对象的特征的信噪比基本上低(例如,小于1)时,该特征可以被认为是暗的或关的。在一些具体实施中,暗或关特征可能不产生任何量的期望信号(即,不产生或检测到信号)。在其它具体实施中,相对于背景而言非常低量的信号可以被认为是暗的或关的。

[0220] 在一个具体实施中,使用一种荧光染料(或具有相同或类似的激发/发射光谱的两种染料)和两个成像事件在测序反应中检测和确定核苷酸掺入的示例性策略由以下检测表例示:

		图像 1	图像 2
[0221]	A	开	关
	C	关	开
	G	关	关
	T	开	开

[0222] 用于检测核苷酸掺入的其它策略,诸如四通道化学和单通道化学,在本公开的范围之内并且不单独论述。

[0223] 在动作1602处,所公开的技术针对给定像素访问在先前测序循环中所调用的碱基(即,先前调用的碱基)下的逐通道启用(开)分类和逐通道非启用(关)分类。这由图17A中的示例示出。在图17A中,像素P描绘对应簇的强度发射,对于该簇,针对循环1至10调用碱基。在一些具体实施中,像素P包含对应簇的中心,如由对应簇的中心的位置坐标所确定。对于第一通道(通道1)中的强度值以及对于第二通道(通道2)中的强度值,从逐通道启用(开)分类1754和逐通道非启用(关)分类1756发出所调用的碱基。

[0224] 在动作1612处,所公开的技术针对给定像素和当前测序循环基于逐通道启用(开)分类1754和逐通道非启用(关)分类1756而累加针对每通道启用(开)状态和每通道非启用(关)状态的逐通道概括统计值。这由图17B中的示例示出。在图17B中,通过将累加函数应用于列1754中被分类为非启用(关)状态的那些强度值,针对测序循环十一(循环11)确定第一通道的非启用(关)状态1702。累加函数的示例包括最大值选择、最小值选择、平均值(均值)、指数加权平均值、移动(运行)平均值、指数移动平均值、众数、标准偏差、方差、偏度、尖峰值、百分位和熵。在图17B中,通过将累加函数应用于列1754中被分类为启用(开)状态的那些强度值,针对循环11确定第一通道的启用(开)状态1712。在图17B中,通过将累加函数应用于列1756中被分类为非启用(关)状态的那些强度值,针对循环11确定第二通道的非启用(关)状态1722。在图17B中,通过将累加函数应用于列1756中被分类为启用(开)状态的那些强度值,针对循环11确定第二通道的启用(开)状态1732。

[0225] 在动作1622处,所公开的技术针对给定像素和当前测序循环将针对每通道启用(开)状态和每通道非启用(关)状态的所累加的逐通道概括统计值与给定像素的当前强度通道组合。这由图17C中的示例示出。在图17C中,1764和1766分别为在循环11处针对像素P登记的第一通道和第二通道的强度值。在图17C中,碱基调用器144将1764、1766、1702、

1712、1722和1732处理为独立输入通道,以在测序循环11处针对像素P的对应簇生成碱基调用。

[0226] 在动作1632处,所公开的技术基于针对每通道启用(开)状态和每通道非启用(关)状态的所累加的逐通道概括统计值与给定像素的强度通道的组合而针对当前测序循环生成至少一个碱基调用。这由图18A中的示例示出。在图18A中,在循环11处基于“关”逐通道分类1854和“关”逐通道分类1856针对像素P的对应簇生成碱基调用“G”,这又由碱基调用器144响应于在图18C中处理1764、1766、1702、1712、1722和1732而进行。

[0227] 在图18B中,响应于循环11处的“关”逐通道分类1854和“关”逐通道分类1856(并且没有任何“开”分类),在循环12处仅更新第一通道和第二通道的关通道状态1802和1822,使得累加函数被重新应用以现在也考虑具有“关”分类1854和1856两者的循环11的逐通道强度值。在一些具体实施中,逐通道开和关状态作为单个变量被维持在存储器中,并且在每个更新实例处被反向计算,例如,通过与当累加函数为平均函数时用于计算平均值的除数相乘。第一通道和第二通道的开通道状态1812和1832在循环12处保持不变,并且分别从循环11继承相同的值,即,1712和1732。

[0228] 在图18C中,1864和1866分别为在循环12处针对像素P登记的第一通道和第二通道的强度值。在图18C中,碱基调用器144将1864、1866、1802、1812、1822和1832处理为独立输入通道,以在测序循环12处针对像素P的对应簇生成碱基调用。

[0229] 在图19A中,在循环12处基于“开”逐通道分类1954和“开”逐通道分类1956针对像素P的对应簇生成碱基调用“T”,这又由碱基调用器144响应于在图18C中处理1864、1866、1802、1812、1822和1832而进行。

[0230] 在图19B中,响应于循环12处的“开”逐通道分类1954和“开”逐通道分类1956(并且没有任何“关”分类),在循环13处仅更新第一通道和第二通道的开通道状态1912和1932,使得累加函数被重新应用以现在也考虑具有“开”分类1954和1956两者的循环12的逐通道强度值。第一通道和第二通道的关通道状态1902和1922在循环13处保持不变,并且分别从循环12继承相同的值,即,1802和1822。

[0231] 在图19C中,1964和1966分别为在循环13处针对像素P登记的第一通道和第二通道的强度值。在图19C中,碱基调用器144将1964、1966、1902、1912、1922和1932处理为独立输入通道,以在测序循环13处针对像素P的对应簇生成碱基调用。

[0232] 在图20A中,在循环13处基于“开”逐通道分类2054和“关”逐通道分类2056针对像素P的对应簇生成碱基调用“A”,这又由碱基调用器144响应于在图19C中处理1964、1966、1902、1912、1922和1932而进行。

[0233] 在图20B中,响应于循环13处的“开”逐通道分类2054和“关”逐通道分类2056,在循环14处分别更新第一通道和第二通道的开通道状态2012和2032以及关通道状态2002和2022,使得累加函数被重新应用以现在也考虑具有“开”分类2054和“关”分类2056的循环13的逐通道强度值。

[0234] 在图20C中,2064和2066分别为在循环14处针对像素P登记的第一通道和第二通道的强度值。在图20C中,碱基调用器144将2064、2066、2002、2012、2022和2032处理为独立输入通道,以在测序循环14处针对像素P的对应簇生成碱基调用。

[0235] 需注意,在碱基调用引导的状态生成的情况下,状态信息滞后一个测序循环。即,

用于碱基调用循环12的状态信息1802、1812、1822和1832是基于循环1-11,并且不包括循环12。类似地,用于碱基调用循环13的状态信息1902、1912、1922和1932是基于循环1-12,并且不包括循环13。

[0236] 指数加权平均

[0237] 图21示出使用指数加权平均来生成2100用于碱基调用的状态信息的一个具体实施。在一个具体实施中,当计算状态信息时,使用指数加权平均来对来自更近测序循环的输入信号(例如,强度值)给予更多权重。因此,通过指数加权平均而生成的状态信息估计跟随信号中的变化并且对于异常值更稳健。

[0238] 在动作2102处,所公开的技术针对给定像素累加来自先前测序循环(例如,前十或二十个测序循环)的历史强度值的概括统计值。累加概括统计值的累加函数的示例包括最大值选择、最小值选择、平均值(均值)、指数加权平均值、移动(运行)平均值、指数移动平均值、众数、标准偏差、方差、偏度、尖峰值、百分位和熵。

[0239] 在动作2112处,所公开的技术针对给定像素基于所累加的概括统计值而初始化起始指数加权平均值。在一个具体实施中,起始指数加权平均值可以为历史强度值的平均值(例如,所有其它像素或所有其它簇像素的历史强度值的平均值)、选自历史强度值的全局最大值(全局MAX)(例如,所有其它像素或所有其它簇像素的历史强度值的全局MAX)或选自历史强度值的全局最小值(全局MIN)(例如,所有其它像素或所有其它簇像素的历史强度值的全局MIN)。

[0240] 在动作2122处,所公开的技术针对给定像素和当前测序循环基于起始指数加权平均值与当前像素强度值的加权组合而确定当前指数加权平均值。

[0241] 在动作2132处,所公开的技术使用当前指数加权平均值作为用于在当前测序循环处进行碱基调用(例如,碱基调用对应于给定像素的簇)的当前状态值。在一个具体实施中,这包括碱基调用器144处理当前状态值和当前像素强度值并且针对当前测序循环生成至少一个碱基调用。

[0242] 在动作2142处,所公开的技术针对给定像素和下一测序循环基于当前指数加权平均值与下一像素强度值的加权组合而确定下一指数加权平均值。

[0243] 在动作2152处,所公开的技术使用下一指数加权平均值作为用于在下一测序循环处进行碱基调用(例如,碱基调用对应于给定像素的簇)的下一状态值。在一个具体实施中,这包括碱基调用器144处理下一状态值和下一像素强度值并且针对下一测序循环生成至少一个碱基调用。

[0244] 在一个具体实施中,指数加权平均逻辑表示如下:

[0245] $y[k] = (1-\alpha) * y[k-1] + \alpha * x[k]$

[0246] 其中:

[0247] $y[k]$ 为针对当前测序循环的指数加权平均值

[0248] $y[k-1]$ 为针对先前测序循环的指数加权平均值

[0249] $x[k]$ 为针对当前测序循环的输入信号(例如,强度值)

[0250] α 为加权参数

[0251] 加权参数“ α ”可以为介于零与一之间的值 $\{0 \leq \alpha \leq 1\}$ 。当 α 等于零 $\{\alpha=0\}$ 时,输出为 $y[k-1]$ (没有平均)。当 α 等于一 $\{\alpha=1\}$ 时,输出为 $x[k]$ 。在其它 α 值处,输出为强度的指数加

权平均值。 α 表示滤波器对更新的反应速度;较小的值反应较慢并且进行更多平均,而较大的值对更近输入值施加更大的权重。

[0252] 需注意, $y[.]$ 可以被存储作为单个变量并在适当位置被更新,因为除了更新上述表达式时之外不使用历史值。 $y[k-1]$ 可以被初始化为例如预期强度(例如,所有簇/孔上的平均初始强度)。在一个具体实施中,前二十个测序循环(循环1-20)可以用于估计初始平均强度,而没有在线估计。初始平均强度可以被插入作为 $y[k-1]$ 的初始估计值,并且然后在测序循环二十(循环20)之后可以使用在线指数加权平均。

[0253] 为了在CPU上实现,可以通过将上述表达式重写为下式来使用一乘有效地实现更新:

[0254] $y[k] = y[k-1] * \alpha (x[k] - y[k-1])$

[0255] 在一些具体实施中, α 可以为接近零的小值,因此执行更多平均。在其它具体实施中,可以改变 α 以找到最适合于特定应用的对v/s响应时间进行平均的折衷的值。

[0256] 每通道指数加权平均

[0257] 图22示出使用指数加权平均来生成2200用于碱基调用的每通道状态信息的一个具体实施。图23描绘使用指数加权平均来生成2300用于碱基调用的每通道状态信息的示例。

[0258] 在动作2202处,所公开的技术针对给定像素累加来自先前测序循环(例如,前十或二十个测序循环)的每通道历史强度值的逐通道概括统计值。累加逐通道概括统计值的每通道累加函数的示例包括最大/全局最大值选择、最小/全局最小值选择、平均值(均值)、指数加权平均值、移动(运行)平均值、指数移动平均值、众数、标准偏差、方差、偏度、尖峰值、百分位和熵。这由图23中的示例示出。

[0259] 在图23中,第一通道(通道1)具有针对前五个测序循环(循环1-5)的历史强度值2374。在图23中,第二通道(通道2)具有针对循环1-5的历史强度值2376。此外,针对通道1从历史强度值2374选择全局最大(全局MAX)值2312,并且针对通道1从历史强度值2374选择全局最小(全局MIN)值2314。类似地,针对通道2从历史强度值2376选择全局MAX值2316,并且针对通道2从历史强度值2376选择全局MIN值2318。

[0260] 在动作2212处,所公开的技术基于所累加的逐通道概括统计值而针对每个通道初始化一对起始启用(开)状态和起始非启用(关)状态。这由图23中的示例示出。在图23中,针对通道1并且在循环5之后,全局MAX 2312被初始化为通道1的启用(开)状态,并且全局MIN 2314被初始化为通道1的非启用(关)状态。在图23中,针对通道2并且在循环5之后,全局MAX 2316被初始化为通道2的启用(开)状态,并且全局MIN 2318被初始化为通道2的非启用(关)状态。

[0261] 在动作2222处,所公开的技术基于当前通道强度值与该对起始启用(开)和非启用(关)状态的比较而针对给定像素和当前测序循环在逐通道的基础上将当前通道强度值归属于起始启用(开)状态或起始非启用(关)状态。这由图23中的示例示出。在图23中,针对通道1且针对循环6,登记了强度值2322。由于强度值2322在数量上更接近全局MAX 2312,所以强度值2322归属于启用(开)状态,因为全局MAX 2312充当通道1的启用(开)状态的代理,并且用作处于启用(开)状态的通道1的起始指数加权平均值。在图23中,针对通道2且针对循环6,登记了强度值2326。

[0262] 由于强度值2326在数量上更接近全局MIN 2318,所以强度值2326归属于非启用(关)状态,因为全局MIN 2318充当通道2的非启用(关)状态的代理,并且用作非启用(关)状态中的通道2的起始指数加权平均值。

[0263] 在动作2232处,所公开的技术针对给定像素且在给定通道内使用指数加权平均值以基于当前通道强度值与归属状态的先前状态值的加权组合而更新归属状态。

[0264] 在动作2242处,所公开的技术针对给定像素且在给定通道内维持来自无归属状态的先前状态值的无归属状态。

[0265] 动作2232和2242由图23中的示例示出。在图23中,针对循环6,如下确定通道1和2的状态信息。使用指数加权平均值2332来确定针对循环6和通道1的启用(开)状态2342,因为针对循环6和通道1的强度值2322归属于通道1的启用(开)状态。指数加权平均值2332使用“0.12”作为 α 值,如关于上述表达式所论述。通道1的非启用(关)状态2344保持不变,并且从循环5继承为通道1的全局MIN 2314。

[0266] 使用指数加权平均值2338来确定针对循环6和通道2的非启用(关)状态2348,因为针对循环6和通道2的强度值2326归属于通道2的非启用(关)状态。指数加权平均值2338也使用“0.12”作为 α 值,如关于上述表达式所论述。通道2的启用(开)状态2346保持不变,并且从循环5继承为通道2的全局MAX 2316。

[0267] 然后,针对循环6的逐通道启用(开)和非启用(关)状态2342、2344、2346和2348与针对循环6登记的强度值2322和2326一起使用,以生成针对循环6的至少一个碱基调用。

[0268] 在图23中,针对循环7,如下确定通道1和2的状态信息。使用指数加权平均值2364来确定针对循环7和通道1的非启用(关)状态2374,因为针对循环7和通道1的强度值2352归属于通道1的非启用(关)状态。指数加权平均值2364也使用“0.12”作为 α 值,如关于上述表达式所论述。通道1的启用(开)状态2372保持不变,并且从循环6继承为通道1的指数加权平均值2342。

[0269] 使用指数加权平均值2366来确定针对循环7和通道2的启用(开)状态2376,因为针对循环7和通道2的强度值2356归属于通道2的启用(开)状态。指数加权平均值2366也使用“0.12”作为 α 值,如上述表达式中所论述。通道2的非启用(关)状态2378保持不变,并且从循环6继承为通道2的指数加权平均值2348。

[0270] 然后,针对循环7的逐通道启用(开)和非启用(关)状态2372、2374、2376和2378与针对循环7登记的强度值2352和2356一起使用,以生成针对循环7的至少一个碱基调用。

[0271] 需注意,在图23中,在给定测序循环处对启用(开)状态和非启用(关)状态的逐通道归属是基于针对给定测序循环(例如,图23中的循环6和7)登记的每通道强度值。论述现在转向碱基调用引导的每通道指数加权平均,其中在给定测序循环处对启用(开)状态和非启用(关)状态的逐通道归属是基于先前测序循环进行的碱基调用。

[0272] 碱基调用引导的每通道指数加权平均

[0273] 图24示出使用由先前调用的碱基引导的指数加权平均来生成2400用于碱基调用的每通道状态信息的一个具体实施。图25示出使用由先前调用的碱基引导的指数加权平均来生成2500用于碱基调用的每通道状态信息的一个具体实施。

[0274] 在图24中,动作2402和2412类似于图22的动作2202和2212。在动作2422处,对启用(开)状态和非启用(关)状态的逐通道归属滞后一个测序循环,并且基于在先前测序循环中

调用的碱基而进行。这由图25中的示例示出。在图25中,在循环6处,分别基于通道1和2的“开”分类2504和“关”分类2506而进行碱基调用“A”2502。

[0275] 在动作2432处,循环6处的“开”分类2504和“关”分类2506以及碱基调用2502用于通过应用指数加权平均来确定针对循环7的状态信息2342、2344、2346和2348。在动作2442处,针对循环7的无归属状态也基于循环6处的碱基调用2502而被标识,并且保持不变。

[0276] 需注意,在循环7处,基于在循环6处确定的状态信息2342、2344、2346和2348以及在循环7处登记的每通道强度值2372和2376而生成“关”分类2514和“开”分类2516以及进而生成碱基调用“C”2512。

[0277] 继续向前,循环7处的“关”分类2514和“开”分类2516以及碱基调用2512用于通过应用指数加权平均来确定针对循环8的状态信息2372、2374、2376和2378。针对循环8的无归属状态也基于循环7处的碱基调用2512而被标识,并且保持不变。

[0278] 尽管未示出,但基于针对循环8登记的逐通道强度值和循环7处针对循环8确定的状态信息2372、2374、2376和2378而生成针对循环8的碱基调用。

[0279] 在不同的具体实施中,使用指数加权平均值(EWA)确定的状态信息可以用于替代、补充或结合使用诸如最小值选择逻辑、最大值选择逻辑、平均逻辑等一些其它逻辑确定的状态信息。

[0280] 非簇像素

[0281] 论述现在转向簇像素与非簇像素之间的区别。簇像素为包含簇的中心的像素,如由簇中心的位置坐标所确定。非簇像素不包含簇中心。需注意,非簇像素的确描绘了簇强度(或背景强度);它们正好与簇中心不一致。

[0282] 此处,对簇-像素-碱基调用关系的简要论述也是有帮助的。簇由多个像素描绘,例如,由 3×3 像素的像素网格描绘。在一些具体实施中,碱基调用器144处理针对给定像素网格中的像素具有不同强度值的测序图像,以表征在不同测序循环处的给定簇的相应强度分布。响应于该处理,在一个具体实施中,碱基调用器144生成输出,该输出通过仅参考包含给定簇的中心的给定像素网格中的中心像素而在不同测序循环处指定针对给定簇的相应碱基调用。即,即使给定像素网格整体表征了给定簇的强度分布,也仅针对中心像素进行相应碱基调用。给定像素网格的非中心像素由碱基调用器144分析以生成碱基调用;然而,仅使用中心像素来表示碱基调用。

[0283] 在仅使用中心像素来表示碱基调用的具体实施中,碱基调用引导的状态生成可以为挑战性的,因为没有碱基调用可用于确定非簇像素的状态。这种限制可以通过下面论述的技术来补偿。

[0284] 图26示出使用先前调用的碱基来针对非簇像素生成2600状态信息的一个具体实施。对于包含簇中心的像素(即,簇像素2602),使用2612簇的先前碱基调用来生成状态数据,如上文例如关于图16至图20C以及图24和图25所论述。

[0285] 根据一个具体实施,对于不包含簇中心的像素(即,非簇像素2608),使用2618最近簇像素的状态数据来生成状态数据。在另一个具体实施中,使用2628所有其它簇像素的状态数据的平均值来生成非簇像素2608的状态数据。

[0286] 状态输入

[0287] 图27示出生成用于碱基调用的状态数据2700的不同具体实施。所公开的技术基于

“实体”2702的当前配置 i 和过去配置 $i-1$ 、 $i-2$ 、……、 1 而生成该实体的状态信息2722。实体2702的示例包括任何类型的信号测量结果、像素测量结果、电压测量结果、电流测量结果、pH标度测量结果、中间处理输出、卷积特征(例如,特征图)、压缩特征(例如,压缩特征图)、不同类型的输出(例如,softmax分数、sigmoid分数、回归分数)、碱基调用和读数。然后根据状态信息2722和实体2702的当前配置 i 来进行当前碱基调用。

[0288] 需注意,为了将状态信息2722与实体2702组合,在一些具体实施中,状态信息2722的维数可能需要与实体2702的维数协调(例如,通过匹配使其兼容),或反之亦然。在不同的具体实施中,这可以通过如克隆、填充(例如,零填充)、级联、卷积、求和、转置卷积等维度改变操作来实现。例如,当状态信息2722具有 1×1 的维数并且实体2702具有 3×3 的维数时,状态信息2722的九个克隆可以与实体2702级联。

[0289] 馈送状态输入

[0290] 图28示出在碱基调用操作的各种处理方面处馈送2800状态数据的不同具体实施。在一个具体实施中,碱基调用器144可以具有各种处理模块1-n(例如,预处理层、神经网络层、后处理层、输出层)。状态数据2802可以被提供给碱基调用器144的处理模块1-n中的任一个处理模块以生成碱基调用2808。例如,状态数据2802可以与输入数据2802(例如,输入图像)组合以供碱基调用器144的第一处理模块(例如,第一卷积层)处理。在另一个示例中,状态数据2802可以与碱基调用器144的中间输出2812(例如,由先前卷积层产生的中间特征图)组合。在又一个示例(未描绘)中,状态数据2802可以与碱基调用器144的最终输出(例如,由最终卷积层产生的最终特征图)组合。在此类具体实施中,状态数据2802与最终输出的组合可以由输出层/碱基调用层(例如,softmax层、回归层、sigmoid层)处理以生成碱基调用2808。

[0291] 基于状态和基于神经网络的碱基调用器

[0292] 以下讨论聚焦于本文所述的基于神经网络的碱基调用器2900。基于神经网络的碱基调用器2900为碱基调用器144的一个具体实施,并且在本文中统称为“DeepRTA”。首先,根据一个具体实施,描述对基于神经网络的碱基调用器2900的输入。然后,提供基于神经网络的碱基调用器2900的结构和形式的示例。最后,根据一个具体实施,描述基于神经网络的碱基调用器2900的输出。

[0293] 数据流逻辑将测序图像提供到基于神经网络的碱基调用器2900以用于碱基调用。基于神经网络的碱基调用器2900在逐块的基础(或逐区块的基础)上访问测序图像。块中的每个块是经像素化的单元的网格中的经像素化的单元的子网格(或子阵列),该子网格形成测序图像。块具有经像素化的单元的子网格的尺寸 $q \times r$,其中 q (宽度)和 r (高度)为在1至10000范围内的任何数字(例如, 3×3 、 5×5 、 7×7 、 10×10 、 15×15 、 25×25 、 64×64 、 78×78 、 115×115)。在一些具体实施中, q 和 r 相同。在其他具体实施中, q 和 r 不同。在一些具体实施中,从测序图像中提取的块具有相同的尺寸。在其他具体实施中,块具有不同的尺寸。在一些具体实施中,块可以具有重叠的经像素化的单元(例如,在边缘上)。

[0294] 测序产生针对对应的 m 个图像通道的每测序循环 m 个测序图像。即,测序图像中的每个测序图像具有一个或多个图像(或强度)通道(类似于彩色图像红色、绿色、蓝色(RGB)通道)。在一个具体实施中,每个图像通道对应于多个滤波器波长带中的一个滤波器波长带。在另一具体实施中,每个图像通道对应于处于测序循环的多个成像事件中的一个

成像事件。在又一具体实施中,每个图像通道对应于利用特定激光器照明与通过特定光学滤波器成像的组合。从针对特定测序循环的 m 个图像通道中的每个图像通道平铺(或访问)图像块。在诸如四通道化学、双通道化学和单通道化学的不同具体实施中, m 为4或2。在其他具体实施中, m 为1、3或大于4。在其它具体实施中,图像可以在蓝色通道和紫色通道中,而不是在红色通道和绿色通道中,或者除了红色和绿色颜色通道之外。

[0295] 例如,考虑测序运行是使用两个不同的图像通道(蓝色通道和绿色通道)来实施的。然后,在每个测序循环,测序运行产生蓝色图像和绿色图像。这样,针对测序运行的一系列的 k 个测序循环,产生 k 对蓝色和绿色图像的序列作为输出,并将其存储为测序图像。因此,由基于神经网络的碱基调用器2900生成用于块级处理的 k 对蓝色图像块和绿色图像块的序列。

[0296] 用于碱基调用的单次迭代的对基于神经网络的碱基调用器2900的输入图像数据(或前向传递的单个示例或单个前向遍历)包括用于多个测序循环的滑动窗口的数据。滑动窗口可以包括例如当前测序循环、一个或多个先前测序循环和一个或多个后续测序循环。

[0297] 在一个具体实施中,输入图像数据包括用于三个测序循环的数据,使得用于当前(时间 t)测序循环的待进行碱基调用的数据伴随有:(i)用于左侧翼/上下文/先前/前一个/之前(时间 $t-1$)测序循环的数据和(ii)用于右侧翼/上下文/下一个/后续/之后(时间 $t+1$)测序循环的数据。

[0298] 在另一具体实施中,输入图像数据包括用于五个测序循环的数据,使得用于待碱基调用的当前(时间 t)测序循环的数据伴随有:(i)用于第一左侧翼/上下文/先前/前一个/之前(时间 $t-1$)测序循环的数据;(ii)用于第二左侧翼/上下文/先前/前一个/之前(时间 $t-2$)测序循环的数据;(iii)用于第一右侧翼/上下文/下一个/后续/之后(时间 $t+1$)测序循环的数据;以及(iv)用于第二右侧翼/上下文/下一个/后续/之后(时间 $t+2$)测序循环的数据。

[0299] 在又一具体实施中,输入图像数据包括用于七个测序循环的数据,使得用于待碱基调用的当前(时间 t)测序循环的数据伴随有:(i)用于第一左侧翼/上下文/先前/前一个/之前(时间 $t-1$)测序循环的数据;(ii)用于第二左侧翼/上下文/先前/前一个/之前(时间 $t-2$)测序循环的数据;(iii)用于第三左侧翼/上下文/先前/前一个/之前(时间 $t-3$)测序循环的数据;(iv)用于第一右侧翼/上下文/下一个/后续/之后(时间 $t+1$)测序循环的数据;(v)用于第二右侧翼/上下文/下一个/后续/之后(时间 $t+2$)测序循环的数据;以及(vi)用于第三右侧翼/上下文/下一个/后续/之后(时间 $t+3$)测序循环的数据。在其他具体实施中,输入图像数据包括用于单个测序循环的数据。在其他具体实施中,输入图像数据包括用于10、15、20、30、58、75、92、130、168、175、209、225、230、275、318、325、330、525或625个测序循环的数据。

[0300] 根据一个具体实施,基于神经网络的碱基调用器2900通过其卷积层处理图像块,并产生替代表示。然后,由输出层(例如,softmax层)使用替代表示以生成仅针对当前(时间 t)测序循环或测序循环中的每个测序循环,即,当前(时间 t)测序循环、第一和第二先前(时间 $t-1$ 、时间 $t-2$)测序循环,以及第一和第二后续(时间 $t+1$ 、时间 $t+2$)测序循环的碱基调用。所得的碱基调用形成测序读段。

[0301] 在一个具体实施中,基于神经网络的碱基调用器2900输出针对特定测序循环的单个目标簇的碱基调用。在另一个具体实施中,基于神经网络的碱基调用器2900输出针对用

于特定测序循环的多个目标簇中的每个目标簇的碱基调用。在又一个具体实施中,基于神经网络的碱基调用器2900输出针对用于多个测序循环中的每个测序循环的多个目标簇中的每个目标簇的碱基调用,从而产生每个目标簇的碱基调用序列。

[0302] 在一个具体实施中,基于神经网络的碱基调用器2900为多层感知器(MLP)。在另一个具体实施中,基于神经网络的碱基调用器2900为前馈神经网络。在又一个具体实施中,基于神经网络的碱基调用器2900为全连接神经网络。在另外的具体实施中,基于神经网络的碱基调用器2900为全卷积神经网络。在又另外的具体实施中,基于神经网络的碱基调用器2900为语义分割神经网络。在又一个另外的具体实施中,基于神经网络的碱基调用器2900为生成对抗网络(GAN)。在又一个具体实施中,基于神经网络的碱基调用器2900包括如变压器、BERT和DETR的多头注意机制。

[0303] 在一个具体实施中,基于神经网络的碱基调用器2900为具有多个卷积层的卷积神经网络(CNN)。在另一个具体实施中,基于神经网络的碱基调用器2900为递归神经网络(RNN),诸如长短期记忆网络(LSTM)、双向LSTM(Bi-LSTM)或门控递归单元(GRU)。在又一个具体实施中,基于神经网络的碱基调用器2900包括CNN和RNN两者。

[0304] 在其它具体实施中,基于神经网络的碱基调用器2900可以使用1D卷积、2D卷积、3D卷积、4D卷积、5D卷积、扩张或空洞卷积、转置卷积、深度可分离卷积、逐点卷积、 1×1 卷积、分组卷积、扁平卷积、空间和跨通道卷积、混洗分组卷积、变压器、BERT、空间可分离卷积和去卷积。基于神经网络的碱基调用器2900可以使用一个或多个损失函数,诸如逻辑回归/对数损失函数、多类交叉熵/softmax损失函数、二元交叉熵损失函数、均方误差损失函数、L1损失函数、L2损失函数、平滑L1损失函数和Huber损失函数。基于神经网络的碱基调用器2900可以使用任何并行性、效率性和压缩方案,诸如TFRecords、压缩编码(例如,PNG)、分片、图变换的并行调用、批处理、预取、模型并行性、数据并行性和同步/异步随机梯度下降(SGD)。基于神经网络的碱基调用器2900可以包括上采样层、下采样层、递归连接、栅极和栅极存储器单元(如LSTM或GRU)、残差块、残差连接、高速连接、跳跃连接、窥视孔连接、激活函数(例如,非线性变换函数如修正线性单元(ReLU)、泄露ReLU、指数衬里单元(ELU)、sigmoid和双曲线正切(tanh))、批量归一化层、正则化层、丢弃层、池化层(例如,最大或平均池化)、全局平均池化层和注意机制。

[0305] 基于神经网络的碱基调用器2900使用基于反向传播的梯度更新技术来训练。可以用于训练基于神经网络的碱基调用器2900的示例性梯度下降技术包括随机梯度下降、批量梯度下降和微型批量梯度下降。可以用于训练基于神经网络的碱基调用器2900的梯度下降优化算法的一些示例为Momentum、Nesterov加速梯度、Adagrad、Adadelat、RMSprop、Adam、AdaMax、Nadam和AMSGrad。

[0306] 在一个具体实施中,基于神经网络的碱基调用器2900使用特化的架构来分离用于不同测序循环的数据的处理。首先描述使用特化的架构的动机。如上所述,基于神经网络的碱基调用器2900处理用于当前测序循环、一个或多个先前测序循环和一个或多个后续测序循环的图像块。附加测序循环的数据提供序列特异性上下文。基于神经网络的碱基调用器2900在训练期间学习序列特异性上下文,并对该序列特异性上下文进行碱基调用。此外,前测序循环和后测序循环的数据为当前测序循环提供了预定相和定相信号的二阶贡献。

[0307] 然而,在不同测序循环处和不同图像通道中捕获的图像相对于彼此未对准并且具

有残差配准误差。考虑到这种未对准,特化的架构包括空间卷积层,该空间卷积层不混合测序循环之间的信息并且仅混合测序循环内的信息。

[0308] 空间卷积层(或空间逻辑)使用所谓的“隔离卷积”,该隔离卷积通过经由“专用非共享”卷积序列独立地处理多个测序循环中的每个测序循环的数据来实现隔离。该隔离卷积对仅给定测序循环(即,循环内)的数据和所得特征映射图进行卷积,而不对任何其他测序循环的数据和所得特征映射图进行卷积。

[0309] 例如,考虑输入图像数据包括:(i)用于待进行碱基调用的当前(时间 t)测序循环的当前图像块;(ii)用于先前(时间 $t-1$)测序循环的先前图像块;以及(iii)用于下一个(时间 $t+1$)测序循环的下一个图像块。然后,特化的架构发起三个单独的卷积流水线,即当前卷积流水线、先前卷积流水线和下一个卷积流水线。当前数据处理流水线接收用于当前(时间 t)测序循环的当前图像块作为输入,并且通过多个空间卷积层独立地处理该当前图像块,以产生所谓的“当前空间卷积表示”作为最终空间卷积层的输出。先前卷积流水线接收用于先前(时间 $t-1$)测序循环的先前图像块作为输入,并且通过多个空间卷积层独立地处理该先前图像块,以产生所谓的“先前空间卷积表示”作为最终空间卷积层的输出。下一个卷积流水线接收用于下一个(时间 $t+1$)测序循环的下一个图像块作为输入,并且通过多个空间卷积层独立地处理该下一个图像块,以产生所谓的“下一个空间卷积表示”作为最终空间卷积层的输出。

[0310] 在一些具体实施中,并行执行当前卷积流水线、先前卷积流水线和下一个卷积流水线。在一些具体实施中,空间卷积层是特化的架构内的空间卷积网络(或子网络)的一部分。

[0311] 基于神经网络的碱基调用器2900还包括混合测序循环之间(即,循环间)的信息的时间卷积层(或时间逻辑)。时间卷积层从空间卷积网络接收其输入,并且对由相应数据处理流水线的最终空间卷积层产生的空间卷积表示进行操作。

[0312] 时间卷积层的循环间可操作性自由源于以下事实:未对准属性通过由空间卷积层序列执行的隔离卷积的堆叠或级联而从空间卷积表示清除,该未对准属性存在于作为输入馈送到空间卷积网络的图像数据中。

[0313] 时间卷积层使用所谓的“组合卷积”,该组合卷积在滑动窗口的基础上逐组地对后续输入中的输入通道进行卷积。在一个具体实施中,这些后续输入是由先前的空间卷积层或先前时间卷积层产生的后续输出。

[0314] 在一些具体实施中,时间卷积层是特化的架构内的时间卷积网络(或子网络)的一部分。时间卷积网络从空间卷积网络接收其输入。在一个具体实施中,时间卷积网络的第一时间卷积层逐组地组合测序循环之间的空间卷积表示。在另一具体实施中,时间卷积网络的之后的时间卷积层组合先前时间卷积层的后续输出。最终时间卷积层的输出被馈送到产生输出的输出层。输出用于在一个或多个测序循环处对一个或多个簇进行碱基调用。

[0315] 关于基于神经网络的碱基调用器2900的更多细节可以在2019年3月21日提交的标题为“Artificial Intelligence-Based Sequencing”的美国临时专利申请号62/821,766(代理人案卷号ILLM 1008-9/IP-1752-PRV)中找到,该专利申请以引用方式并入本文。

[0316] 每像素状态输入

[0317] 图29示出基于状态和基于神经网络的碱基调用的一个具体实施。在图29所示的示

例中,测序循环的第一窗口包括测序循环N-2、N-1、N、N+1和N+2。在一个具体实施中,用于相应测序循环N-2、N-1、N、N+1和N+2的相应图像块(测序图像/簇图像)2914、2924、2934、2944和2954与用于当前测序循环N的状态数据2912逐像素地组合(例如,级联)。状态数据2912可以使用上述技术中的任何技术来确定。

[0318] 图像块2914、2924、2934、2944和2954统称为簇图像(测序图像)2950。每像素每通道状态数据2940包括状态数据2912的五个实例,因为存在五个测序循环N-2、N-1、N、N+1和N+2。

[0319] 状态数据2912具有簇图像2950中的对应像素的每像素状态值,类似于上文关于图3和图4所论述的具体实施。而且,在一个具体实施中,因为簇图像2950具有两个通道(强度通道),所以状态数据2912也具有两个状态通道。在其它具体实施中,状态数据2912可以仅具有一个通道或具有多于两个通道,如上文在不同的具体实施中所论述。

[0320] 在高级别下,每像素每通道状态数据2940与簇图像2950组合以供空间逻辑106(或空间网络或空间子网络或空间卷积神经网络)处理。通过空间逻辑106分别处理每像素每通道状态数据2940与簇图像2950的组合,以生成用于相应测序循环N-2、N-1、N、N+1和N+2的相应空间图2916、2926、2936、2946和2956(或中间结果或空间输出集或空间特征图集)。空间卷积网络106可以使用1D、2D或3D卷积。

[0321] 空间逻辑106包括空间卷积层的序列(或级联)。每个空间卷积层具有滤波器组,该滤波器组具有实施隔离卷积的多个空间卷积滤波器。因此,每个空间卷积层产生多个空间特征图作为输出。由目标空间卷积层产生的空间特征图的数量是在目标空间卷积层中配置的空间卷积滤波器的数量的函数。例如,如果目标空间卷积层具有14个空间卷积滤波器,则目标空间卷积层产生14个空间特征图。从整体的角度来看,14个空间特征图可以被认为是具有14个通道(或深度维度=14)的空间特征图体积(或张量)。

[0322] 此外,在目标空间卷积层之后的下一个空间卷积层也可以配置有14个空间卷积滤波器。在此类情况下,下一空间卷积层将生成目标空间卷积层的十四空间特征图作为输入进行处理,并且其自身生成十四新的空间特征图作为输出。图29示出由空间网络106的最终空间卷积层针对相应测序循环N-2、N-1、N、N+1和N+2生成的五个空间特征图集2916、2926、2936、2946和2956。例如,五个空间特征图集2916、2926、2936、2946和2956中的每个空间特征图集具有十四个特征图。

[0323] 在另一个示例中,可以通过空间网络106的七个空间卷积层的级联来生成七个空间特征图集的序列。目标测序循环i的每循环输入块数据和状态数据的组合可以具有例如 115×115 的空间维数和深度维数2(由于原始测序图像中的两个图像通道)。在一个具体实施中,七个空间卷积层的每一个空间卷积层使用 3×3 卷积,将后续空间特征图体积的空间维数减少2,例如从 10×10 减少到 8×8 。

[0324] 第一空间特征图体积可以具有空间维度 113×113 (即,从 115×115 减小第一空间卷积层的 3×3 卷积)和深度维度14(即,由于第一空间卷积层中的十四空间卷积滤波器而导致的十四特征图或十四通道)。第二空间特征图体积可以具有空间维度 111×111 (即,从 113×113 减小第二空间卷积层的 3×3 卷积)和深度维度14(即,由于第二空间卷积层中的十四空间卷积滤波器而导致的十四特征图或十四通道)。第三空间特征图体积可以具有空间维度 109×109 (即,从 111×111 减小第三空间卷积层的 3×3 卷积)和深度维度

14(即,由于第三空间卷积层中的十四个空间卷积滤波器而导致的十四个特征图或十四个通道)。第四空间特征图体积可以具有空间维度 107×107 (即,从 109×109 减小第四空间卷积层的 3×3 卷积)和深度维度14(即,由于第四空间卷积层中的十四个空间卷积滤波器而导致的十四个特征图或十四个通道)。第五空间特征图体积可以具有空间维度 105×105 (即,从 107×107 减小第五空间卷积层的 3×3 卷积)和深度维度14(即,由于第五空间卷积层中的十四个空间卷积滤波器而导致的十四个特征图或十四个通道)。第六空间特征图体积可以具有空间维度 103×103 (即,从 105×105 减小第六空间卷积层的 3×3 卷积)和深度维度14(即,由于第六空间卷积层中的十四个空间卷积滤波器而导致的十四个特征图或十四个通道)。第七空间特征图体积可以具有空间维度 101×101 (即,从 103×103 减小第七空间卷积层的 3×3 卷积)和深度维度14(即,由于第七空间卷积层中的十四个空间卷积滤波器而导致的十四个特征图或十四个通道)。

[0325] 与图29所示的多循环示例类似,对于五个测序循环N-2、N-1、N、N+1和N+2以及五个每循环图像块2914、2924、2934、2944和2954以及状态数据2912的五个实例,空间逻辑106分别产生七个空间特征图体积的五个相应序列,其中图29中的空间图2916、2926、2936、2946和2956等同于最终空间特征图体积的五个独立实例。

[0326] 空间图2916、2926、2936、2946和2956由时间逻辑2928处理以生成针对当前测序循环N的碱基调用2930。

[0327] 每簇/每孔状态

[0328] 图30示出将每簇状态数据用于碱基调用簇的一个具体实施。与图29不同,在图30中,状态数据不与输入簇图像2914、2924、2934、2944和2954组合。相反,状态数据与空间逻辑106的输出组合。需注意,簇图像2914、2924、2934、2944和2954包含簇像素和非簇像素两者。因此,空间图3016、3026、3036、3046和3056包含从簇像素和非簇像素两者发出的空间卷积特征。

[0329] 在一些具体实施中,簇特征滤波逻辑3030使用簇中心位置3002来从空间图3016、3026、3036、3046和3056中滤出对应于非簇像素的那些特征。所得的经滤波的每簇空间图3054、3064、3074、3084和3094仅包含对应于簇像素的那些特征,统称为簇特征3004。在一个具体实施中,空间图3016、3026、3036、3046和3056的维数为 101×101 ,并且经滤波的每簇空间图3054、3064、3074、3084和3094的维数为 25×25 。在此类具体实施中, $25 \times 25 \times k$ 张量包括作为对与簇像素相对应的数据应用连续卷积运算的结果而产生的空间特征,这些簇像素包括原始簇像素和簇像素的连续生成的卷积特征。

[0330] 与其中状态数据2940在到基于神经网络的碱基调用器2900的起始输入处与簇图像2950组合的图29不同,在图30中,用于当前测序循环N的每簇状态3042与基于神经网络的碱基调用器2900的中间输出(即,经滤波的每簇空间图3054、3064、3074、3084和3094)组合。

[0331] 在一些具体实施中,每簇状态3042的维数被修改以符合经滤波的每簇空间图3054、3064、3074、3084和3094的维数(例如,通过修整维度、添加维度、填充(例如,零填充)、克隆等)。在不同的具体实施中,每簇状态3042的相应五个实例和经滤波的每簇空间图3054、3064、3074、3084和3094可以使用上文所论述的技术来分别组合,诸如级联、求和、逐元素乘法、逐元素乘法和求和(卷积)等。

[0332] 然后,根据图30所示的示例,时间逻辑2928处理相应经滤波的每簇空间图3054、

3064、3074、3084和3094与每簇状态3042的相应五个实例的相应五个组合以生成针对当前测序循环N的碱基调用3060。碱基调用3060是针对对应于簇像素且因此对应于簇特征3004的簇进行的。

[0333] 论述现在转向每簇状态3042的不同变化以及如何根据所公开的技术的不同具体实施来确定这些变化。

[0334] 图31示出使用对应簇像素的历史强度值来生成3100每簇状态的一个具体实施。在图31中,通过累加来自先前测序循环和当前测序循环的簇像素的历史强度值的概括统计值来确定每簇状态3042,如上文所论述,例如,使用最小值选择函数、最大值选择函数、平均函数、指数加权平均函数等。然后,每簇状态3042与簇像素的空间卷积特征以及当前碱基调用迭代/操作的空间卷积特征相组合,以生成针对对应簇的碱基调用。这些步骤由图31的动作3102、3112、3122和3132示出。

[0335] 图32示出使用对应于簇像素的空间卷积特征的历史特征值来生成3200每簇状态的一个具体实施。在图31中,通过累加对应于来自先前测序循环和当前测序循环的簇像素的空间卷积特征的历史特征值的概括统计值来确定每簇状态3042,如上文所论述,例如,使用最小值选择函数、最大值选择函数、平均函数、指数加权平均函数等。然后,每簇状态3042与簇像素的空间卷积特征以及当前碱基调用迭代/操作的空间卷积特征相组合,以生成针对对应簇的碱基调用。这些步骤由图32的动作3202、3212、3222和3232示出。

[0336] 图33示出通过内插像素强度来生成3300簇强度并且使用经内插簇强度来生成用于碱基调用的每簇状态的一个具体实施。在一个具体实施中,可以对测序图像中的像素群组的强度值进行插值(例如,通过使用线性、双线性、三次插值)以估计簇强度。在一些具体实施中,簇中心位置和簇形状数据可以用于选择给定像素群组以内插对应簇的簇强度。在其它具体实施中,簇形状数据可以用于使用部分像素的加权强度来确定簇强度。

[0337] 可以累加在先前测序循环和当前测序循环处内插的簇强度的概括统计值以生成针对当前测序循环的每簇状态3042,如上文所论述,例如,使用最小值选择函数、最大值选择函数、平均函数、指数加权平均函数等。然后,每簇状态3042与簇像素的空间卷积特征以及当前碱基调用迭代/操作的空间卷积特征相组合,以生成针对对应簇的碱基调用。这些步骤由图32的动作3302、3312、3322和3332示出。

[0338] 作为每簇状态的RTA系数

[0339] 图34示出使用与基于神经网络的碱基调用器分开的实时分析(RTA)碱基调用器2900来生成3400每簇状态的一个具体实施。

[0340] 关于RTA碱基调用器,RTA为可以使用线性强度提取器从测序图像提取特征以用于碱基调用的碱基调用器。以下论述描述通过RTA进行的强度提取和碱基调用的一个具体实施。在该具体实施中,RTA可以执行模板生成步骤以产生模板图像,该模板图像使用来自被称为模板循环的一定数量的初始测序循环的测序图像来标识区块上的簇的位置。模板图像可以用作随后的配准和强度提取步骤的参考。可以通过检测和合并模版循环的每个测序图像中的亮点来生成模版图像,这又涉及锐化测序图像(例如,使用拉普拉斯卷积)、通过空间隔离的大津方法确定“开”阈值,以及随后的具有子像素位置插值的五像素局部最大值检测。在另一个示例中,可以使用基准标记来标识区块上的簇的位置。生物学标本成像在其上的固体载体可包括此类基准标记物,以便于确定标本或其图像相对于附接到固体载体的探

针的取向。示例性基准可以包括但不限于小珠(具有或不具有荧光部分或诸如带标记的探针可以与其结合的核酸之类的部分)、以已知或可以确定的特征附接的荧光分子或将形态学形状与荧光部分组合的结构。示例性基准阐述于美国专利公布号2002/0150909中,该美国专利公布以引用方式并入本文。

[0341] RTA然后可以相对于模板图像配准当前测序图像。这可以通过使用图像相关性来将当前测序图像与子区上的模板图像对准,或者通过使用非线性变换(例如,全六参数线性仿射变换)来实现。

[0342] RTA可以生成颜色矩阵以校正测序图像的颜色通道之间的串扰。RTA可以实现经验相位校正以补偿由相位误差引起的测序图像中的噪声。

[0343] 在将不同的校正应用于测序图像之后,RTA可以提取测序图像中每个斑点位置的信号强度。例如,对于给定斑点位置,可以通过确定斑点位置中的像素的强度的加权平均值来提取信号强度。例如,可以使用双线性或双三次插值来执行中心像素和相邻像素的加权平均。在一些具体实施中,图像中的每个斑点位置可以包括几个像素(例如,1-5个像素)。

[0344] RTA然后可以对所提取的信号强度进行空间归一化,以考虑跨所采样的图像的照明变化。例如,强度值可以被归一化,使得第5百分位和第95百分位分别具有值0和1。图像的归一化信号强度(例如,每个通道的归一化强度)可以用于计算图像中的多个斑点的平均简洁度。

[0345] 在一些具体实施中,RTA可以使用均衡器来最大化所提取的信号强度的信噪比。均衡器可以被训练(例如,使用最小二乘估计、自适应均衡算法)以最大化测序图像中的簇强度数据的信噪比。在一个具体实施中,均衡器可以包括经训练的系数以校正空间串扰以及其它形式的串扰和噪声。在一些具体实施中,均衡器可以为单个查找表(LUT)。在其它具体实施中,均衡器可以为具有子像素分辨率的多个LUT的LUT组,也被称为“均衡器滤波器”或“卷积内核”。在一个具体实施中,均衡器中的LUT的数量可以取决于测序图像的像素可以被划分成的子像素的数量。例如,如果像素可被划分成 $n \times n$ 个子像素(例如, 5×5 个子像素),则均衡器可以生成 n^2 个LUT(例如,25个LUT)。在又一些具体实施中,均衡器可以基于快速傅立叶变换(FFT)。在又一些具体实施中,均衡器可以基于Winograd卷积。

[0346] 在训练均衡器的一个具体实施中,来自测序图像的数据通过孔子像素位置进行分箱。例如,对于 5×5 的LUT,1/25的孔的中心在箱(1,1)中(例如,传感器像素的左上角),1/25的孔在箱(1,2)中,依此类推。在一个具体实施中,使用对来自对应于相应箱的孔的数据子集的最小二乘估计来确定每个箱的均衡器系数。这样,所得的估计均衡器系数对于每个箱来说是不同的。

[0347] 每个LUT/均衡器滤波器/卷积内核具有从训练学习的多个系数。在一个具体实施中,LUT中系数的数量对应于用于对簇进行碱基调用的像素的数量。例如,如果用于对簇进行碱基调用的像素(图像或像素块)的局部网格的大小为 $p \times p$ (例如, 9×9 像素块),则每个LUT具有 p^2 个系数(例如,81个系数)。

[0348] 在一个具体实施中,训练产生均衡器系数,这些均衡器系数被配置为以最大化信噪比的方式混合/组合像素的强度值,所述像素描绘来自进行碱基调用的目标簇的强度发射和来自一个或多个相邻簇的强度发射。信噪比被最大化的信号是来自目标簇的强度发射,而信噪比被最小化的噪声是来自相邻簇的强度发射,即空间串扰,加上一些随机噪声

(例如,以考虑背景强度发射)。均衡器系数用作权重,并且混合/组合包括执行均衡器系数与像素的强度值之间的逐元素乘法,以计算这些像素的强度值的加权和,即,卷积运算。

[0349] RTA可以通过将数学模型拟合到优化的强度数据来执行碱基调用。可以使用的合适的数学模型包括例如k均值分簇算法、类k均值分簇算法、预期最大化分簇算法、基于直方图的方法等。四个高斯分布可以拟合到双通道强度数据集,使得一个分布应用于数据集中表示的四个核苷酸中的每个核苷酸。在一个特定具体实施中,可以应用预期最大化(EM)算法。作为EM算法的结果,对于每个X、Y值(分别指两个通道强度的每个通道强度),可以生成表示某个X、Y强度值属于数据所拟合的四个高斯分布中的一个高斯分布的可能性的值。在四个碱基给出四个独立分布的情况下,X、Y强度值也将具有四个相关联的可能性值,四个碱基中的每个碱基对应一个可能性值。四个可能性值中的最大值指示碱基调用。例如,如果簇在两个通道中均为“关”,则碱基调用为G。如果簇在一个通道中为“关”且在另一个通道中为“开”,则碱基调用为C或T(取决于哪一个通道为开),并且如果簇在两个通道中均为“开”,则碱基调用为A。

[0350] 关于RTA的更多细节可以在以下专利申请中找到:2018年3月1日提交的标题为“Optical Distortion Correction For Imaged Samples”的美国非临时专利申请号15/909,437;2014年10月31日提交的标题为“Image Analysis Useful For Patterned Objects”的美国非临时专利申请号14/530,299;2014年12月3日提交的标题为“Methods And Systems For Analyzing Image Data”的美国非临时专利申请号15/153,953;2011年1月13日提交的标题为“Data Processing System And Methods”的美国非临时专利申请号13/006,206;以及2021年5月4日提交的标题为“Equalization-Based Image Processing And Spatial Crosstalk Attenuator”的美国非临时专利申请号17/308,035(代理人案卷号ILLM 1032-2/IP-1991-US),所有这些申请均以引用的方式并入,如在本文完全阐述一样。

[0351] 簇群体中簇的大量(例如,数千个、数百万个、数十亿个等)强度分布中的簇间强度分布变化导致碱基调用吞吐量下降和碱基调用误差率增加。为了校正簇间强度分布变化,RTA在逐簇的基础上生成每簇变化校正系数。

[0352] 在双通道具体实施中,每簇变化校正系数包括说明簇间强度分布变化中的尺度变化的放大系数,以及分别说明簇间强度分布变化中沿第一强度通道和第二强度通道的移位变化的两个通道特异性偏移系数。在另一种具体实施中,通过使用针对不同强度通道(例如,第一强度通道和第二强度通道)的公共偏移系数来解释移位变化。

[0353] 对于本领域的技术人员来说显而易见的是,RTA的变化校正逻辑可以类似地应用于使用单通道具体实施、四通道具体实施等生成的测序图像。例如,在四通道具体实施的情况下,确定四个通道特异性偏移系数,以分别校正四个强度通道中的移位变化。

[0354] 基于将在测序运行的先前测序循环处针对目标簇确定的历史强度统计值的分析与在当前测序循环处针对目标簇确定的当前强度统计值的分析相结合,在测序运行的当前测序循环处生成目标簇的每簇变化校正系数。每簇变化校正系数用于在测序运行的下一测序循环处校正针对目标簇登记的下一强度读数。经过校正的下一个强度读数用于在下一个测序循环处碱基调用目标簇。在测序运行的连续测序循环处,将相应每簇变化校正系数重复地应用于相应簇的相应强度分布的结果是,强度分布变得一致并锚固到原点(例如,在梯

形的底部下角处)。

[0355] 当前测序循环

[0356] 在当前测序循环*i*处,测序仪生成测序图像。测序图像包含在当前测序循环*i*处针对目标簇登记的当前强度数据,以及包含针对簇群体中的多个簇登记的当前强度数据。

[0357] 当前强度数据被提供给RTA。RTA处理当前强度数据并且在当前测序循环*i*处生成针对目标簇的当前碱基调用。

[0358] 在当前测序循环*i*处,目标簇的强度分布包括当前强度数据,以及在当前测序循环*i*之前,即,在先前测序循环1至*i*-1之前的测序运行的那些测序循环处针对目标簇登记的当前历史强度数据。我们将当前强度数据和当前历史强度数据统称为当前可用强度数据。

[0359] 在强度分布中,四个强度分布对应于四个碱基A、C、T和G。在一个具体实施中,通过确定当前强度数据属于四个强度分布中的哪个强度分布来进行当前碱基调用。在一些具体实施中,这是通过使用期望最大化算法来实现的。期望值最大化算法迭代地最大化最拟合当前可用强度数据的观察平均值(质心)和分布(协方差)的似然性。

[0360] 一旦通过使用期望最大化算法在当前测序循环*i*处确定了四个强度分布,就计算属于四个强度分布中的每个强度分布的当前强度数据的可能性。最大可能性给出了当前碱基调用。例如,考虑“*m*、*n*”分别是第一强度通道和第二强度通道中的当前强度数据的强度值。期望值最大化算法生成四个值,这四个值表示属于四个强度分布中的每个强度分布的“*m*、*n*”个强度值的似然性。四个值中的最大值表示被调用的碱基。

[0361] 在其它具体实施中,*k*均值分簇算法、类*k*均值分簇算法、基于直方图的方法等可以用于碱基调用。

[0362] 下一个测序循环

[0363] 在下一测序循环*i*+1处,强度校正参数确定器基于当前碱基调用而确定用于目标簇的强度校正参数。在双通道具体实施中,强度校正参数包括第一强度通道中的分布强度、第二强度通道中的分布强度、第一强度通道中的强度误差、第二强度通道中的强度误差、分布质心到原点的距离以及分布强度到强度误差的相似性度量。

[0364] 我们如下定义强度校正参数中的每个强度校正参数:

[0365] 1) 第一强度通道中的分布强度是在碱基特异性强度分布的质心处的第一强度通道中的强度值,目标簇在当前测序循环*i*处属于该碱基特异性强度分布。需注意,碱基特异性强度分布是用于调用当前碱基调用的基础。

[0366] 2) 第二强度通道中的分布强度是在碱基特异性强度分布的质心处的第二强度通道中的强度值。

[0367] 3) 第一强度通道中的强度误差是第一强度通道中的当前强度数据的测量强度值与第一强度通道中的分布强度之间的差异。

[0368] 4) 第二强度通道中的强度误差是第二强度通道中的当前强度数据的测量强度值与第二强度通道中的分布强度之间的差异。

[0369] 5) 分布质心到原点的距离是碱基特异性强度分布的质心与多维空间的原点之间的欧几里得距离,其中碱基特异性强度分布被拟合(例如,通过使用期望最大化算法)。在其他具体实施中,可使用距离度量,诸如马氏距离和最小协方差行列式(MCD)距离,以及其相关联的质心估计器。

[0370] 6) 分布强度到强度误差相似性度量是第一强度通道和第二强度通道中的分布强度和强度误差之间的逐通道点积的总和。

[0371] 累加强度校正参数确定器将强度校正参数与来自先前测序循环 $i-1$ 的历史累加强度校正参数累加,以确定累加强度校正参数。累加的示例包括求和与求平均。

[0372] 变化校正系数确定器基于确定的累加强度校正参数而确定变化校正系数。

[0373] 在下一测序循环 $i+1$ 处,测序仪生成测序图像。测序图像包含在下一测序循环 $i+1$ 处针对目标簇登记的下一强度数据,以及包含针对簇群体中的多个簇登记的下一强度数据。

[0374] 强度校正器将变化校正系数应用于下一强度数据,以生成经校正的下一强度数据。

[0375] 在下一测序循环 $i+1$ 处,目标簇的强度分布包括经校正的下一强度数据,以及在下一测序循环 $i+1$ 之前,即,在先前测序循环 1 至 i 之前的测序运行的那些测序循环中针对目标簇登记的下一历史强度数据。我们将经校正的下一强度数据和下一历史强度数据统称为下一可用强度数据。

[0376] 经校正的下一强度数据被提供给RTA。RTA处理经校正的下一强度数据,并且可以在下一测序循环 $i+1$ 处针对目标簇生成下一碱基调用。为了生成下一碱基调用,预期最大化算法可以基于经校正的下一强度数据而观察均值(质心)和分布(协方差),以最佳拟合下一可用强度数据。

[0377] 一旦通过使用预期最大化算法在下一测序循环 $i+1$ 处确定了四个强度分布,就计算属于四个强度分布中的每个强度分布的经校正的下一强度数据的可能性。最大可能性给出了下一碱基调用。

[0378] 需注意,碱基调用流水线是在逐簇的基础上执行的,并且是针对簇群体中的多个簇并行执行的。而且,对于测序运行的连续测序循环(例如,针对双端测序运行中读段1的连续150个测序循环和读段2的另外连续150个测序循环),重复执行碱基调用流水线。

[0379] 最小二乘解

[0380] 最小二乘解确定针对累加强度校正参数和变化校正系数的闭合形式表达式。最小二乘解确定器包括强度建模器和最小化器。

[0381] 强度建模器根据以下表达式针对目标簇的测量强度与变化校正系数之间的关系进行建模:

[0382] $y_{C,i} = ax_{C,i} + d_i + n_{C,i}$ 等式(1)

[0383] 其中:

[0384] a 为用于目标簇的放大系数

[0385] d_i 为用于强度通道 i 的通道特异性偏移系数

[0386] $x_{C,i}$ 为在当前测序循环 C 处在强度通道 i 中针对目标簇的分布强度

[0387] $y_{C,i}$ 为在当前测序循环 C 处在强度通道 i 中针对目标簇的测量强度

[0388] $n_{C,i}$ 为在当前测序循环 C 处在强度通道 i 中针对目标簇的附加噪声

[0389] 最小化器使用最小二乘解来最小化以下表达式:

$$[0390] \quad \text{errorf}(\hat{a}, \hat{d}_i) = \sum_{c=1}^C \sum_{i=1}^2 (\hat{a}x_{c,i} + \hat{d}_i - y_{c,i})^2 \quad \text{等式 (2)}$$

[0391] 其中:

[0392] errorf为误差函数

[0393] $\hat{a}$ 为用于目标簇的放大系数

[0394] $\hat{d}_i$ 为强度通道i的通道特异性偏移系数

[0395] C为当前测序循环

[0396] 使用链式法则,最小化器计算误差函数相对于放大系数 $\hat{a}$ 和通道特异性偏移系数 $\hat{d}_i$ 的两个偏导数。偏导数将等式2设置为零,以最小化误差函数:

$$[0397] \quad \frac{\partial \text{errorf}}{\partial \hat{a}} = \frac{\partial}{\partial \hat{a}} \left[\sum_{c=1}^C \sum_{i=1}^2 (\hat{a}x_{c,i} + \hat{d}_i - y_{c,i})^2 \right] = 0 \quad \text{等式 (3)}$$

$$[0398] \quad \frac{\partial \text{errorf}}{\partial \hat{d}_i} = \frac{\partial}{\partial \hat{d}_i} \left[\sum_{c=1}^C \sum_{i=1}^2 (\hat{a}x_{c,i} + \hat{d}_i - y_{c,i})^2 \right] = 0 \quad \text{等式 (4)}$$

[0399] 通道特异性强度误差 $e_{c,i}$ 定义如下:

$$[0400] \quad e_{c,i} = y_{c,i} - x_{c,i} \quad \text{等式 (5)}$$

[0401] 闭合形式表达式

[0402] 第一偏导数确定针对放大系数 $\hat{a}$ 的闭合形式表达式如下:

$$\frac{\partial \text{errorf}}{\partial \hat{a}} = \sum_{c=1}^C \sum_{i=1}^2 2x_{c,i} (\hat{a}x_{c,i} + \hat{d}_i - y_{c,i}) = 0 \quad \text{等式 (6)}$$

$$= \sum_{c=1}^C \sum_{i=1}^2 2x_{c,i} (\hat{a}x_{c,i} - x_{c,i} + \hat{d}_i - y_{c,i} + x_{c,i}) = 0 \quad \text{等式 (7)}$$

$$[0403] \quad = \sum_{c=1}^C \sum_{i=1}^2 x_{c,i}^2 (\hat{a} - 1) + \hat{d}_i x_{c,i} - e_{c,i} x_{c,i} = 0 \quad \text{等式 (8)}$$

$$= (\hat{a} - 1) \sum_{c=1}^C (x_{c,1}^2 + x_{c,2}^2) + \hat{d}_1 \sum_{c=1}^C x_{c,1} + \hat{d}_2 \sum_{c=1}^C x_{c,2} - \sum_{c=1}^C e_{c,1} x_{c,1} - \sum_{c=1}^C e_{c,2} x_{c,2} = 0 \quad \text{等式 (9)}$$

[0404] 累加强度校正参数的闭合形式表达式 $\bar{x}_1$ 、 $\bar{x}_2$ 、 $\bar{e}_1$ 、 $\bar{e}_2$ 、 $\overline{xx}$ 和 $\overline{xe}$ 将等式9重新表征如下:

$$[0405] \quad = C(\hat{a} - 1)\overline{xx} + C\bar{x}_1\hat{d}_1 + C\bar{x}_2\hat{d}_2 - C\overline{xe} = 0 \quad \text{等式 (10)}$$

$$[0406] \quad = (\hat{a} - 1)\overline{xx} + \bar{x}_1\hat{d}_1 + \bar{x}_2\hat{d}_2 - \overline{xe} = 0 \quad \text{等式 (11)}$$

[0407] 其中:

$$[0408] \quad \bar{x}_1 = \sum_{c=1}^C x_{c,1} \quad \text{中间术语 (1)}$$

$$[0409] \quad \bar{x}_2 = \sum_{c=1}^C x_{c,2} \quad \text{中间术语 (2)}$$

$$[0410] \quad \bar{e}_1 = \sum_{c=1}^C e_{c,1} \quad \text{中间术语 (3)}$$

$$[0411] \quad \bar{e}_2 = \sum_{c=1}^C e_{c,2} \quad \text{中间术语 (4)}$$

$$[0412] \quad \overline{xx} = \sum_{c=1}^C (x_{c,1}^2 + x_{c,2}^2) \quad \text{中间术语 (5)}$$

$$[0413] \quad \overline{xe} = \sum_{c=1}^C (x_{c,1}e_{c,1} + x_{c,2}e_{c,2}) \quad \text{中间术语 (6)}$$

[0414] 我们如下定义累加强度校正参数中的每个累加强度校正参数：

[0415] 1) 第一累加强度校正参数 $\bar{x}_1$ 为在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇测量的第一强度通道中的分布强度的总和。

[0416] 2) 第二累加强度校正参数 $\bar{x}_2$ 为在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇测量的第二强度通道中的分布强度的总和。

[0417] 3) 第三累加强度校正参数 $\bar{e}_1$ 为在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的第一强度通道中的强度误差的总和。

[0418] 4) 第四累加强度校正参数 $\bar{e}_2$ 为在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的第二强度通道中的强度误差的总和。

[0419] 5) 第五累加强度校正参数 $\overline{xx}$ 为在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的分布质心到原点的距离的总和。

[0420] 6) 第六累加强度校正参数 $\overline{xe}$ 为在先前测序循环1至i-1和当前测序循环i的每一者处针对目标簇计算的分布强度-强度误差相似性度量的总和。

[0421] 第二偏导数确定针对偏移系数 $\hat{d}_i$ 的闭合形式表达式如下：

$$\frac{\partial errorf}{\partial \hat{d}_i} = \sum_{c=1}^C \sum_{i=1}^2 2 (\hat{a}x_{c,i} + \hat{d}_i - y_{c,i}) = 0 \quad \text{等式 (12)}$$

$$[0422] \quad = \sum_{c=1}^C \sum_{i=1}^2 (\hat{a}x_{c,i} - x_{c,i} + \hat{d}_i - y_{c,i} + x_{c,i}) = 0 \quad \text{等式 (13)}$$

$$= \sum_{c=1}^C \sum_{i=1}^2 x_{c,i}^2 \left((\hat{a} - 1) + \sum_{i=1}^2 \hat{d}_i + \sum_{c=1}^C x_{c,i} \right) = 0 \quad \text{等式 (14)}$$

[0423] 然后,针对每个强度通道：

$$[0424] \quad = C(\hat{a} - 1)\bar{x}_i + C\hat{d}_i - C\bar{e}_i = 0 \quad \text{等式 (15)}$$

$$[0425] \quad = (\hat{a} - 1)\bar{x}_i + \hat{d}_i - \bar{e}_i = 0 \quad \text{等式 (16)}$$

[0426] 针对第一强度通道,即i=1：

$$[0427] \quad \hat{d}_1 = \bar{e}_1 + (1 - \hat{a})\bar{x}_1 \quad \text{等式 (17)}$$

[0428] 其中：

[0429] $\hat{d}_1$ 为针对第一强度通道的偏移系数

[0430] 针对第二强度通道,即i=2：

[0431] $\hat{d}_2 = \bar{e}_2 + (1 - \hat{a})\bar{x}_2$ 等式 (18)

[0432] 其中:

[0433] $\hat{d}_2$ 为针对第二强度通道的偏移系数

[0434] 将等式17和等式18代入等式11:

[0435] $\frac{\partial errorf}{\partial \hat{a}} = (\hat{a} - 1)\overline{xx} + \bar{x}_1[\bar{e}_1 + (1 - \hat{a})\bar{x}_1] + \bar{x}_2[\bar{e}_2 + (1 - \hat{a})\bar{x}_2] - \overline{xe} = 0$ 等式 (19)

$$= (\hat{a} - 1)\overline{xx} + (1 - \hat{a})\bar{x}_1^2 + (1 - \hat{a})\bar{x}_2^2 + \bar{x}_1\bar{e}_1 + \bar{x}_2\bar{e}_2 - \overline{xe} = 0 \quad \text{等式 (20)}$$

$$= (\hat{a} - 1)(\overline{xx} - \bar{x}_1^2 - \bar{x}_2^2) + \bar{x}_1\bar{e}_1 + \bar{x}_2\bar{e}_2 - \overline{xe} = 0 \quad \text{等式 (21)}$$

[0436] $\hat{a} - 1 = \frac{\bar{x}_1\bar{e}_1 + \bar{x}_2\bar{e}_2 - \overline{xe}}{\overline{xx} - \bar{x}_1^2 - \bar{x}_2^2}$ 等式 (22)

[0437] $\hat{a} = 1 + \frac{\bar{x}_1\bar{e}_1 + \bar{x}_2\bar{e}_2 - \overline{xe}}{\overline{xx} - \bar{x}_1^2 - \bar{x}_2^2}$ 等式 (23)

[0438] 其中:

[0439] $\hat{a}$ 为用于目标簇的放大系数

[0440] 在另一个具体实施中,为了减少每簇的存储器需求,通过引入约束 $\hat{d}_1 = \hat{d}_2$,如下确定不同强度通道(例如,第一强度通道和第二强度通道)的公共偏移系数:

[0441] $\bar{x} = \frac{1}{2C} \sum_{c=1}^C (x_{c,1} + x_{c,2})$ 中间术语 (1.1)

[0442] $\overline{xx} = \frac{1}{C} \sum_{c=1}^C (x_{c,1}^2 + x_{c,2}^2)$ 中间术语 (2.1)

[0443] $\bar{e} = \frac{1}{2C} \sum_{c=1}^C (e_{c,1} + e_{c,2})$ 中间术语 (3.1)

[0444] $\overline{xe} = \frac{1}{C} \sum_{c=1}^C (x_{c,1}e_{c,1} + x_{c,2}e_{c,2})$ 中间术语 (4.1)

[0445] $\hat{a} = 1 + \frac{\overline{xe} - 2\bar{x}\bar{e}}{\overline{xx} - 2\bar{x}^2}$ 等式 (24)

[0446] $\hat{d}_1 = \bar{e} + \bar{x}(1 - \hat{a})$ 等式 (25)

[0447] $\hat{d}_2 = \bar{e} + \bar{x}(1 - \hat{a})$ 等式 (26)

[0448] 对本领域的技术人员来说显而易见的是,在测序运行之前执行最小二乘解以确定闭合形式表达式。一旦确定,闭合形式表达式被应用于在逐簇测序运行期间生成的强度值,并在测序运行的每个测序循环处迭代。

[0449] 关于每簇变化校正系数以及如何确定它们的额外细节可以在2021年10月27日提交的标题为“Systems And Methods For Per-Cluster Intensity Correction And Base Calling”的美国临时专利申请号63/106,256(代理人案卷号ILLM 1034-1/IP-2026-PRV)中找到,该申请以全文引用的方式并入本文。

[0450] 在图34中,根据一个具体实施,每簇状态3042为由RTA针对当前测序循环确定的每簇变化校正系数。然后,每簇状态3042与簇像素的空间卷积特征以及当前碱基调用迭代/操作的空间卷积特征相组合,以生成针对对应簇的碱基调用。这些步骤由图34的动作3402、3412、3422和3432示出。

[0451] 压缩网络

[0452] 如上所述,基于神经网络的碱基调用器2900的专门化架构为对应的测序循环处理图像块的滑动窗口。在后续滑动窗口的测序循环之间存在重叠。这使得基于神经网络的碱基调用器2900对于重叠的测序循环冗余地处理图像块。这继而导致计算资源的浪费。例如,在一个具体实施中,基于神经网络的碱基调用器2900的每个空间卷积层具有近亿次乘法运算。然后,对于5个测序循环的窗口和7个空间卷积层的级联(或序列),空间卷积神经网络执行约6.2亿次乘法运算。此外,时间卷积神经网络执行大约1000万次乘法运算。

[0453] 由于当前滑动窗口(或碱基调用的当前迭代)中循环N-1的图像数据在先前的滑动窗口(或碱基调用的先前迭代)中被处理为循环N,因此有机会存储在当前滑动窗口中完成的处理的中间结果和它们在后续滑动窗口中的中间结果,从而绕过(或避免)对输入图像数据的冗余处理(或重新处理),以用于后续滑动窗口之间的重叠的测序循环。

[0454] 然而,中间结果是需要不切实际的存储量的数TB的数据。为了克服该技术问题,所公开的技术提出了在基于神经网络的碱基调用器2900第一次生成中间结果时压缩中间结果,并且在随后的滑动窗口中重新利用压缩的中间结果以避免冗余计算,从而不重新生成(或仅生成一次)中间结果。

[0455] 在图35中,根据一个具体实施,压缩逻辑3530(或压缩网络或压缩子网络或压缩层或挤压层)处理簇特征滤波逻辑3030的输出并且生成输出的压缩表示。在一个具体实施中,压缩网络3530包括减小由簇特征滤波逻辑3030生成的特征图的深度维数的压缩卷积层。

[0456] 例如,认为每簇空间图3054、3064、3074、3084和3094的深度维数为14(即,每个空间输出十四特征图或十四通道)。压缩网络3530针对相应测序循环N-2、N-1、N、N+1和N+2将经滤波的每簇空间图3054、3064、3074、3084和3094衰减为相应压缩的经滤波的每簇空间图3554、3564、3574、3584和3594,统称为压缩簇特征3504。在一个具体实施中,压缩的经滤波每簇空间图3554、3564、3574、3584和3594中的每个压缩的经滤波每簇空间图具有深度维数2(即,每个压缩空间输出两个特征图或两个通道)。在其它具体实施中,压缩的经滤波每簇空间图3554、3564、3574、3584和3594可以具有深度维数3或4(即,每个压缩空间输出三个或第四特征图或三个或第四通道)。在其它具体实施中,压缩的经滤波每簇空间图3554、3564、3574、3584和3594可以具有深度维数1(即,每个压缩的空间输出一个特征图或一个通道)。在一个具体实施中,压缩层3530不包括类似于ReLU的激活功能。在其他具体实施中,它可以包括激活功能。在其它具体实施中,压缩逻辑3530可以将对应的压缩空间图集配置为各自具有多于四个特征图。

[0457] 现在论述压缩逻辑3530如何生成压缩输出。

[0458] 在一个具体实施中,压缩逻辑3530使用 1×1 卷积来减少特征图的数量(即,深度维度或通道的数量),同时引入非线性。 1×1 卷积的内核大小为1。 1×1 卷积可以在不改变空间维度的情况下将体积深度变换为另一种挤压或扩展的表示。 1×1 卷积运算类似于跨输入通道的完全连接的线性层。这对于从具有许多通道的特征图映射到较少的特征图是有用的。单个 1×1 卷积可以应用于具有两个特征图的输入张量。 1×1 卷积将双通道输入压缩为单通道输出。

[0459] 由压缩层3530生成的压缩输出(或压缩特征图或压缩的空间图或压缩的时间图)的数量是在压缩层108中配置的 1×1 卷积滤波器(或压缩卷积滤波器或压缩滤波器)的数量的函数。在一个具体实施中,压缩层3530可以具有两个 1×1 卷积滤波器。第一 1×1 卷积滤波

器可以处理具有十四个特征图的空间特征体积,并且生成第一特征图,同时保持 101×101 的空间维数。第二 1×1 卷积滤波器也可以处理具有十四个特征图的空间特征体积,并且生成第二特征图,同时保持 101×101 的空间维数。因此,压缩层3530将具有十四个特征图的空间特征体积减小为具有两个空间特征图的压缩输出(即,压缩比=7)。

[0460] 在一些具体实施中,所公开的技术在基于神经网络的碱基调用器2900的空间网络中节省了约80%的卷积。在一个具体实施中,当压缩逻辑和随后的测序循环中压缩的特征图的重新利用用于五个测序循环的输入窗口(例如,循环N、循环N+1、循环N-1、循环N+2、循环N-2)时,在空间卷积中可以观察到80%的节省。在另一个具体实施中,当压缩逻辑和随后的测序循环中压缩的特征图的重新利用用于10个测序循环的输入窗口(例如,循环N、循环N+1、循环N-1、循环N+2、循环N-2、循环N+3以及循环N-3)时,在空间卷积中可以观察到90%的节省。即,窗口尺寸越大,使用压缩逻辑和重新利用压缩的特征图所带来的节省就越大,并且窗口尺寸越大,由于结合了来自附加侧翼循环的更大上下文,碱基调用性能就越好。因此,对于给定的计算能力,较大窗口的较大节省提高了整体性能。

[0461] 由压缩逻辑带来的计算效率和紧凑的计算占用面积有助于基于神经网络的碱基调用器2900在资源受限的处理器上的硬件实现,这些资源受限的处理器如中央处理单元(CPU)、图形处理单元(GPU)、现场可编程门阵列(FPGA)、粗粒度可重新配置架构(CGRA)、专用集成电路(ASIC)、专用指令集处理器(ASIP)和数字信号处理器(DSP)。

[0462] 由压缩逻辑保存的计算允许在基于神经网络的碱基调用器2900中结合更多的卷积算子。示例包括在空间和时间卷积层中增加更多的卷积滤波器、增加卷积滤波器的尺寸,以及增加空间卷积层和时间卷积层的数量。附加卷积运算提高了基于神经网络的碱基调用器2900的强度模式检测和整体碱基调用准确性。

[0463] 关于压缩逻辑和其压缩输出的更多细节可以在2021年2月18日提交的标题为“Data Compression For Artificial Intelligence-Based Base Calling”的美国非临时专利申请号17/179,395(代理人案卷号ILLM 1029-2/IP-1964-US)中找到,其以全文引用的方式并入本文。

[0464] 在图35中,用于当前测序循环N的每簇状态3042与基于神经网络的碱基调用器2900的中间输出(即,压缩的经滤波每簇空间图3554、3564、3574、3584和3594)组合。

[0465] 在一些具体实施中,每簇状态3042的维数被修改以符合压缩的经滤波每簇空间图3554、3564、3574、3584和3594的维数(例如,通过修整维度、添加维度、填充(例如,零填充)、克隆等)。在不同的具体实施中,每簇状态3042的相应五个实例和压缩的经滤波每簇空间图3554、3564、3574、3584和3594可以使用上文所论述的技术来分别组合,诸如级联、求和、逐元素乘法、逐元素乘法和求和(卷积)等。

[0466] 然后,根据图35所示的示例,时间逻辑2928处理相应压缩的经滤波每簇空间图3554、3564、3574、3584和3594与每簇状态3042的相应五个实例的相应五个组合以生成针对当前测序循环N的碱基调用3080。碱基调用3080是针对对应于簇像素且因此对应于簇特征3004的簇进行的。

[0467] 论述现在转向每簇状态3042的不同变化以及如何根据所公开的技术的不同具体实施来确定这些变化。

[0468] 图36示出使用对应于簇像素的压缩空间卷积特征的历史特征值来生成3600每簇

状态的一个具体实施。在图36中,通过累加对应于来自先前测序循环和当前测序循环的簇像素的压缩空间卷积特征的历史特征值的概括统计值来确定每簇状态3042,如上文所论述,例如,使用最小值选择函数、最大值选择函数、平均函数、指数加权平均函数等。然后,每簇状态3042与簇像素的压缩空间卷积特征以及当前碱基调用迭代/操作的压缩空间卷积特征相组合,以生成针对对应簇的碱基调用。这些步骤由图36的动作3602、3612、3622、3632和3642示出。

[0469] 在图37中,根据一个具体实施,每簇状态3042为由RTA针对当前测序循环确定的每簇变化校正系数。然后,每簇状态3042与簇像素的压缩空间卷积特征以及当前碱基调用迭代/操作的压缩空间卷积特征相组合,以生成针对对应簇的碱基调用。这些步骤由图37的动作3702、3712、3722、3732和3742示出。

[0470] 图38示出使用对应簇像素的历史强度值来生成3800每簇状态3042以用于将每簇状态3042与压缩特征合并的一个具体实施。在图38中,通过累加来自先前测序循环和当前测序循环的簇像素的历史强度值的概括统计值来确定每簇状态3042,如上文所论述,例如,使用最小值选择函数、最大值选择函数、平均函数、指数加权平均函数等。然后,每簇状态3042与簇像素的压缩空间卷积特征以及当前碱基调用迭代/操作的压缩空间卷积特征相组合,以生成针对对应簇的碱基调用。这些步骤由图38的动作3802、3812、3822、3832和3842示出。

[0471] 孔状态数据到像素状态数据的变换

[0472] 图39示出从稀疏的每孔状态生成密集的每像素状态的一个具体实施。在图39中,每孔状态3902为稀疏的,并且可以使用转置卷积或插值(例如,线性、双线性、三次插值)3904变换为密集的每像素状态3906。

[0473] “稀疏”和“密集”是指阵列(例如,矢量或矩阵)中零元素相对于非零元素的数量。稀疏阵列是包含大部分零和少数非零条目的阵列。密集阵列包含大部分非零。在一个具体实施中,在密集阵列中,每个每像素状态可以由至少两个、四个或八个相邻每像素状态围绕。在稀疏阵列中,在一个具体实施中,每个每孔状态不被甚至两个、四个或八个相邻每孔状态围绕。如果第一行与最后一行之间的所有行被定义为非零元素并且被给予值,则状态的集合被称为密集的。如果行中存在间隙或零元素,则状态的集合被称为稀疏的。稀疏性可一意味着许多元素(例如,每隔一个、每三个或每四个)为零或非常接近零。

[0474] 图40示出提供稀疏的每孔状态、密集的每像素状态和每像素强度值作为到碱基调用器的输入以用于执行碱基调用操作的一个具体实施。在图40中,稀疏的每孔状态3902可以被提供作为到碱基调用器144(例如,基于神经网络的碱基调用器2900)的输入,作为直接输入以及使用例如残余连接和/或跳过连接的冗余输入。在一些具体实施中,稀疏的每孔状态3902连同密集的每像素状态3906以及密集的每像素强度值4012(即,簇图像)一起被提供作为补充输入,该补充输入由碱基调用器144处理以生成碱基调用。

[0475] 图41示出使用稀疏的每孔状态、密集的每像素状态和每像素强度值作为到基于神经网络的碱基调用器2900的输入以用于执行碱基调用操作的一个具体实施。到基于神经网络的碱基调用器2900的输入为密集的每像素状态3906和密集的每像素强度值4012(即,簇图像)。

[0476] 测序系统

[0477] 图42A和图42B描绘测序系统4200A的一个具体实施。测序系统4200A包括可配置处理器4246。可配置处理器4246实现本文所公开的碱基调用技术。测序系统也称为“测序仪”。

[0478] 测序系统4200A可以进行操作以获得与生物物质或化学物质中的至少一者相关的任何信息或数据。在一些具体实施中,测序系统4200A为可类似于台式设备或台式计算机的工作站。例如,用于进行期望反应的大部分(或全部)系统和部件可以位于共同的外壳4202内。

[0479] 在特定具体实施中,测序系统4200A为被配置用于各种应用的核酸测序系统,各种应用包括但不限于从头测序、全基因组或靶基因组区域的重测序以及宏基因组学。测序仪也可用于DNA或RNA分析。在一些具体实施中,测序系统4200A还可以被配置为在生物传感器中生成反应位点。例如,测序系统4200A可以被配置为接收样品并生成来源于样品的克隆扩增核酸的表面附着簇。每个簇可构成生物传感器中的反应位点或作为其一部分。

[0480] 示例性测序系统4200A可以包括被配置为与生物传感器4212相互作用以在生物传感器4212内进行期望反应的系统插座或接口4210。在以下关于图42A的描述中,将生物传感器4212装载到系统插座4210中。然而,应当理解,可以将包括生物传感器4212的卡盒插入到系统插座4210中,并且在一些状态下,可以暂时或永久地移除卡盒。如上所述,除了别的以外,卡盒还可包括流体控制部件和流体储存部件。

[0481] 在特定具体实施中,测序系统4200A被配置为在生物传感器4212内进行大量平行反应。生物传感器4212包括可以发生期望反应的一个或多个反应位点。反应位点可例如固定至生物传感器的固体表面或固定至位于生物传感器的对应反应室内的小珠(或其他可移动基板)。反应位点可包括,例如,克隆扩增核酸的簇。生物传感器4212可以包括固态成像设备(例如,CCD或CMOS成像器件)和安装到其上的流通池。流通池可以包括一个或多个流动通道,该一个或多个流动通道从测序系统4200A接收溶液并将溶液引向反应位点。任选地,生物传感器4212可以被配置为接合热元件,以用于将热能传递到流动通道中或从流动通道传递出去。

[0482] 测序系统4200A可以包括彼此相互作用以执行用于生物或化学分析的预定方法或测定协议的各种部件、组件和系统(或子系统)。例如,测序系统4200A包括系统控制器4206,该系统控制器可以与测序系统4200A的各种部件、组件和子系统以及生物传感器4212通信。例如,除了系统插座4210之外,测序系统4200A还可以包括:流体控制系统4208,该流体控制系统用于控制通过测序系统4200A的整个流体网络和生物传感器4212的流体流动;流体存储系统4214,该流体存储系统被配置为容纳可生物测定系统可使用的所有流体(例如,气体或液体);温度控制系统4204,该温度控制系统可以调节流体网络、流体存储系统4214和/或生物传感器4212中的流体的温度;以及照明系统4216,该照明系统被配置为照亮生物传感器4212。如上所述,如果将具有生物传感器4212的卡盒装载到系统插座4210中,则该卡盒还可以包括流体控制部件和流体存储部件。

[0483] 而且如图所示,测序系统4200A可以包括与用户交互的用户界面4218。例如,用户界面4218可以包括用于显示或请求来自用户的信息的显示器4220和用于接收用户输入的用户输入设备4222。在一些具体实施中,显示器4220和用户输入设备4222为同一设备。例如,用户界面4218可以包括触敏显示器,该触敏显示器被配置为检测个体触摸的存在并且还识别触摸在显示器上的位置。然而,可以使用其它用户输入设备4222,诸如鼠标、触摸板、

键盘、小键盘、手持扫描仪、语音辨识系统、运动辨识系统等。如将在下文更详细地论述,测序系统4200A可以与包括生物传感器4212(例如,呈卡盒的形式)的各种部件通信,以进行期望反应。测序系统4200A还可以被配置为分析从生物传感器获得的数据以向用户提供期望信息。

[0484] 系统控制器4206可以包括任何基于处理器或基于微处理器的系统,包括使用微控制器、精简指令集计算机(RISC)、专用集成电路(ASIC)、现场可编程门阵列(FPGA)、粗粒度可重新配置架构(CGRA)、逻辑电路以及能够执行本文所述的功能的任何其它电路或处理器。上述示例仅是示例性的,因此不旨在以任何方式限制术语系统控制器的定义和/或含义。在示例性具体实施中,系统控制器4206执行存储在一个或多个存储元件、存储器或模块中的指令集,以便进行获得检测数据和分析检测数据中的至少一者。检测数据可包括多个像素信号序列,使得可在许多碱基调用循环内检测来自数百万个传感器(或像素)中的每一个传感器(或像素)的像素信号序列。存储元件可以呈测序系统4200A内的信息源或物理存储元件的形式。

[0485] 指令集可以包括指示测序系统4200A或生物传感器4212执行特定操作(诸如本文所述的各种具体实施的方法和过程)的各种命令。指令集可为软件程序的形式,该软件程序可形成有形的一个或多个非暂态计算机可读介质的一部分。如本文所用,术语“软件”和“固件”是可互换的,并且包括存储在存储器中以供计算机执行的任何计算机程序,包括RAM存储器、ROM存储器、EPROM存储器、EEPROM存储器和非易失性RAM(NVRAM)存储器。上述存储器类型仅是示例性的,因此不限制可用于存储计算机程序的存储器类型。

[0486] 软件可为各种形式,诸如系统软件或应用软件。此外,软件可以是独立程序的集合的形式,或者是较大程序内的程序模块或程序模块的一部分的形式。软件还可包括面向对象编程形式的模块化编程。在获得检测数据之后,检测数据可以由测序系统4200A自动处理,响应于用户输入而处理,或者响应于另一个处理机器提出的请求(例如,通过通信链路的远程请求)而处理。在示出的具体实施中,系统控制器4206包括分析模块4244。在其它具体实施中,系统控制器4206不包括分析模块4244,而是具有对分析模块4244的访问权限(例如,分析模块4244可以单独地托管在云上)。

[0487] 系统控制器4206可以经由通信链路连接到生物传感器4212和测序系统4200A的其它部件。系统控制器4206还可以通信地连接到非现场系统或服务器。通信链路可以是硬连线的、有线的或无线的。系统控制器4206可以从用户界面4218和用户输入设备4222接收用户输入或命令。

[0488] 流体控制系统4208包括流体网络,并且被配置为引导和调节一种或多种流体通过流体网络的流动。流体网络可以与生物传感器4212和流体存储系统4214流体连通。例如,选择的流体可以从流体存储系统4214抽吸并以受控方式引导到生物传感器4212,或者流体可以从生物传感器4212抽吸并朝向例如流体存储系统4214中的废物存储器引导。虽然未示出,但流体控制系统4208可以包括检测流体网络内的流体的流速或压力的流量传感器。传感器可以与系统控制器4206通信。

[0489] 温度控制系统4204被配置为调节流体网络、流体存储系统4214和/或生物传感器4212的不同区处的流体的温度。例如,温度控制系统4204可以包括热循环仪,该热循环仪与生物传感器4212对接并且控制沿着生物传感器4212中的反应位点流动的流体的温度。温度

控制系统4204还可以调节测序系统4200A或生物传感器4212的固体元件或部件的温度。尽管未示出,但温度控制系统4204可以包括用于检测流体或其它部件的温度的传感器。传感器可以与系统控制器4206通信。

[0490] 流体存储系统4214与生物传感器4212流体连通,并且可以存储用于在其中进行期望反应的各种反应组分或反应物。流体存储系统4214还可以存储用于洗涤或清洁流体网络和生物传感器4212以及用于稀释反应物的流体。例如,流体存储系统4214可以包括各种贮存器,以存储样品、试剂、酶、其它生物分子、缓冲溶液、水性溶液和非极性溶液等。此外,流体存储系统4214还可以包括废物存储器,用于接收来自生物传感器4212的废物。在包括卡盒的具体实施中,卡盒可包括流体储存系统、流体控制系统或温度控制系统中的一者或多者。因此,本文所述的与那些系统有关的一个或多个部件可容纳在卡盒外壳内。例如,卡盒可具有各种储存器,以储存样品、试剂、酶、其他生物分子、缓冲溶液、水性溶液和非极性溶液、废物等。因此,流体储存系统、流体控制系统或温度控制系统中的一者或多者可经由卡盒或其他生物传感器与生物测定系统可移除地接合。

[0491] 照明系统4216可以包括光源(例如,一个或多个LED)和用于照明生物传感器的多个光学部件。光源的示例可包括激光器、弧光灯、LED或激光二极管。光学部件可以是例如反射器、二向色镜、分束器、准直器、透镜、滤光器、楔镜、棱镜、反射镜、检测器等。在使用照明系统的具体实施中,照明系统4216可以被配置为将激发光引导到反应位点。作为一个示例,荧光团可以由绿色波长的光激发,因此激发光的波长可以为大约4232nm。在一个具体实施中,照明系统4216被配置为产生平行于生物传感器4212的表面的表面法线的照明。在另一个具体实施中,照明系统4216被配置为产生相对于生物传感器4212的表面的表面法线成偏角的照明。在又一个具体实施中,照明系统4216被配置为产生具有多个角度的照明,包括一些平行照明和一些偏角照明。

[0492] 系统插座或接口4210被配置为以机械方式、电气方式和流体方式中的至少一种方式接合生物传感器4212。系统插座4210可以将生物传感器4212保持在期望方向,以有利于流体流过生物传感器4212。系统插座4210还可以包括电触点,这些电触点被配置为接合生物传感器4212,使得测序系统4200A可以与生物传感器4212通信和/或向生物传感器4212提供功率。此外,系统插座4210可以包括被配置为接合生物传感器4212的流体端口(例如,喷嘴)。在一些具体实施中,生物传感器4212以机械方式、电气方式以及流体方式可移除地耦接到系统插座4210。

[0493] 此外,测序系统4200A可以与其它系统或网络或其它生物测定系统4200A远程通信。由生物测定系统4200A获得的检测数据可以存储在远程数据库中。

[0494] 图42B是可以在图42A的系统中使用的系统控制器4206的框图。在一个具体实施中,系统控制器4206包括可以彼此通信的一个或多个处理器或模块。处理器或模块中的每一者可以包括用于执行特定过程的算法(例如,存储在有形和/或非暂态计算机可读存储介质上的指令)或子算法。系统控制器4206在概念上被示出为模块的集合,但可以利用专用硬件板、DSP、处理器等的任何组合来实现。另选地,系统控制器4206可以利用具有单个处理器或多个处理器的现成PC来实现,其中功能操作分布在处理器之间。作为进一步的选择,下文所述的模块可利用混合配置来实现,其中某些模块化功能利用专用硬件来执行,而其余模块化功能利用现成PC等来执行。模块还可被实现为处理单元内的软件模块。

[0495] 在操作期间,通信端口4250可以将信息(例如,命令)传输到生物传感器4212(图42A)和/或子系统4208、4214、4204(图42A)或从其接收信息(例如,数据)。在具体实施中,通信端口4250可以输出多个像素信号序列。通信链路4234可以从用户界面4218(图42A)接收用户输入并将数据或信息传输到用户界面4218。来自生物传感器4212或子系统4208、4214、4204的数据可以在生物测定会话期间由系统控制器4206实时处理。除此之外或另选地,数据可在生物测定会话期间临时存储在系统存储器中,并且以比实时或脱机操作更慢的速度进行处理。

[0496] 如图42B所示,系统控制器4206可以包括与主控制模块4224以及中央处理单元(CPU)4252通信的多个模块4226-4248。主控制模块4224可以与用户界面4218(图42A)通信。尽管模块4226-4248被示出为与主控制模块4224直接通信,但模块4226-4248也可以彼此直接通信,与用户界面4218和生物传感器4212直接通信。而且,模块4226-4248可以通过其它模块与主控制模块4224通信。

[0497] 多个模块4226-4248包括分别与子系统4208、4214、4204和4216通信的系统模块4228-4232、4226。流体控制模块4228可以与流体控制系统4208通信,以控制流体网络的阀和流量传感器,从而控制一种或多种流体通过流体网络的流动。流体存储模块4230可以在流体量低时或在废物存储器处于或接近容量时通知用户。流体存储模块4230还可以与温度控制模块4232通信,使得流体可以存储在期望温度下。照明模块4226可以与照明系统4216通信,以在协议期间的指定时间照亮反应位点,诸如在已发生期望反应(例如,结合事件)之后。在一些具体实施中,照明模块4226可以与照明系统4216通信,从而以指定角度照明反应位点。

[0498] 多个模块4226-4248还可以包括与生物传感器4212通信的设备模块4236和确定与生物传感器4212相关的标识信息的标识模块4238。设备模块4236可以例如与系统插座4210通信以确认生物传感器已与测序系统4200A建立电连接和流体连接。标识模块4238可以接收标识生物传感器4212的信号。标识模块4238可以使用生物传感器4212的身份来将其它信息提供给用户。例如,标识模块4238可以确定并然后显示批号、制造日期或建议与生物传感器4212一起运行的协议。

[0499] 多个模块4226-4248还包括接收和分析来自生物传感器4212的信号数据(例如,图像数据)的分析模块4244(也称为信号处理模块或信号处理器)。分析模块4244包括用于存储检测/图像数据的存储器(例如,RAM或闪存)。检测数据可包括多个像素信号序列,使得可在许多碱基调用循环内检测来自数百万个传感器(或像素)中的每一个传感器(或像素)的像素信号序列。信号数据可以被存储用于后续分析,或者可以被传输到用户界面4218以向用户显示期望信息。在一些具体实施中,信号数据可以在分析模块4244接收到信号数据之前由固态成像器件(例如,CMOS图像传感器)处理。

[0500] 分析模块4244被配置为在多个测序循环中的每个测序循环处从光检测器获得图像数据。图像数据来源于由光检测器检测到的发射信号,并且通过基于神经网络的碱基调用器2900处理该多个测序循环中的每个测序循环的图像数据,并且在多个测序循环中的每个测序循环处针对簇中的至少一些簇产生碱基调用。光检测器可以为一个或多个俯视相机(例如,从顶部拍摄生物传感器4212上的簇的图像的Illumina的GAIIX的CCD相机)的一部分,或者可以为生物传感器4212本身(例如,位于生物传感器4212上的簇下面并且从底部拍

摄簇的图像的Illumina的iSeq的CMOS图像传感器)的一部分。

[0501] 光检测器的输出是测序图像,每个测序图像描绘了簇及其周围背景的信号发射。测序图像描绘了由于在测序期间核苷酸掺入序列中而产生的信号发射。信号发射来自相关联的簇及其周围背景。测序图像存储在存储器4248中。

[0502] 协议模块4240和4242与主控制模块4224通信,以在进行预先确定的测定协议时控制子系统4208、4214和4204的操作。协议模块4240和4242可以包括用于指示测序系统4200A根据预先确定的协议执行特定操作的指令集。如图所示,协议模块可以为边合成边测序(SBS)模块4240,该模块被配置为发出用于执行边合成边测序过程的各种命令。在SBS中,监测核酸引物沿核酸模板的延伸,以确定模板中核苷酸的序列。基础化学过程可以是聚合(例如,由聚合酶催化)或连接(例如,由连接酶催化)。在特定的基于聚合酶的SBS具体实施中,以依赖于模板的方式将荧光标记的核苷酸添加至引物(从而使引物延伸),使得对添加至引物的核苷酸的顺序和类型的检测可用于确定模板的序列。例如,为了启动第一SBS循环,可发出命令以将一个或多个标记的核苷酸、DNA聚合酶等递送至/通过容纳有核酸模板阵列的流通池。核酸模板可位于对应的反应位点。其中引物延伸导致标记的核苷酸掺入的那些反应位点可通过成像事件来检测。在成像事件期间,照明系统4216可以将激发光提供到反应位点。任选地,核苷酸还可以包括一旦将核苷酸添加到引物就终止进一步的引物延伸的可逆终止属性。例如,可以将具有可逆终止子部分的核苷酸类似物添加到引物,使得后续的延伸直到递送解封闭剂以去除该部分才发生。因此,对于使用可逆终止的具体实施,可发出命令以将解封闭剂递送到流通池(在检测发生之前或之后)。可发出一个或多个命令以实现各个递送步骤之间的洗涤。然后可重复该循环n次,以将引物延伸n个核苷酸,从而检测长度为n的序列。示例性测序技术描述于:例如Bentley等人,Nature 456:53-59(2008)、WO 04/018497;US 7,057,026;WO 91/06678;WO 07/123744;US 7,329,492;US 7,211,414;US 7,315,019;US 7,405,281和US 2008/014708082,这些文献中的每一篇以引用方式并入本文。

[0503] 对于SBS循环的核苷酸递送步骤,可以一次递送单一类型的核苷酸,或者可以递送多种不同的核苷酸类型(例如,A、C、T和G一起)。对于一次仅存在单一类型的核苷酸的核苷酸递送构型,不同的核苷酸不需要具有不同的标记,因为它们可基于个体化递送中固有的时间间隔来区分。因此,测序方法或装置可使用单色检测。例如,激发源仅提供单个波长或单个波长范围内的激发。对于其中递送导致多种不同核苷酸同时存在于流通池中的核苷酸递送构型,可基于附着到混合物中相应核苷酸类型的不同荧光标记来区分掺入不同核苷酸类型的位点。例如,可使用四种不同的核苷酸,每种核苷酸具有四种不同荧光团中的一种。在一个具体实施中,可使用在光谱的四个不同区域中的激发来区分四种不同的荧光团。例如,可使用四种不同的激发辐射源。另选地,可使用少于四种不同的激发源,但来自单个源的激发辐射的光学过滤可用于在流通池处产生不同范围的激发辐射。

[0504] 在一些具体实施中,可在具有四种不同核苷酸的混合物中检测到少于四种不同颜色。例如,核苷酸对可在相同波长下检测,但基于对中的一个成员相对于另一个成员的信号差异,或基于对中的一个成员的导致与检测到的该对的另一个成员的信号相比明显的信号出现或消失的变化(例如,通过化学改性、光化学改性或物理改性)来区分。用于使用少于四种颜色的检测来区分四个不同核苷酸的示例性装置和方法描述于例如美国专利申请序列号61/538,294和61/619,878,这些专利申请全文以引用方式并入本文。2012年9月21日提交

的美国申请号13/624,200也全文以引用方式并入。

[0505] 多个协议模块还可以包括样品制备(或生成)模块4242,该模块被配置为向流体控制系统4208和温度控制系统4204发出命令,以扩增生物传感器4212内的产物。例如,生物传感器4212可以接合到测序系统4200A。扩增模块4242可以向流体控制系统4208发出指令,以将必要的扩增组分递送到生物传感器4212内的反应室。在其他具体实施中,反应位点可能已包含一些用于扩增的组分,诸如模板DNA和/或引物。在将扩增组分递送到反应室之后,扩增模块4242可以指示温度控制系统4204根据已知的扩增协议循环通过不同的温度阶段。在一些具体实施中,扩增和/或核苷酸掺入等温进行。

[0506] SBS模块4240可以发出命令以执行桥式PCR,其中克隆扩增子的簇形成在流通池的通道内的局部区域上。通过桥式PCR产生扩增子后,可将扩增子“线性化”以制备单链模板DNA或sstDNA,并且可将测序引物杂交至侧接感兴趣的区域的通用序列。例如,可如上所述或如下使用基于可逆终止子的边合成边测序方法。

[0507] 每个碱基调用或测序循环可通过单个碱基延伸sstDNA,这可例如通过使用经修饰的DNA聚合酶和四种类型的核苷酸的混合物来完成。不同类型的核苷酸可具有独特的荧光标记,并且每个核苷酸还可具有可逆终止子,该可逆终止子仅允许在每个循环中发生单碱基掺入。在将单个碱基添加到sstDNA之后,激发光可入射到反应位点上并且可检测荧光发射。在检测后,可从sstDNA化学切割荧光标记和终止子。接下来可为另一个类似的碱基调用或测序循环。在此类测序协议中,SBS模块4240可以指示流体控制系统4208引导试剂和酶溶液流过生物传感器4212。可与本文所述的设备和方法一起使用的基于可逆终止子的示例性SBS方法描述于美国专利申请公布号2007/0166705A1、美国专利申请公布号2006/0188901A1、美国专利号7,057,026、美国专利申请公布号2006/0240439A1、美国专利申请公布号2006/02814714709A1、PCT公布号W0 05/065814、美国专利申请公布号2005/014700900A1、PCT公布号W0 06/08B199和PCT公布号W0 07/01470251,这些专利中的每一篇均全文以引用方式并入本文。用于基于可逆终止子的SBS的示例性试剂描述于US 7,541,444;US 7,057,026;US 7,414,14716;US 7,427,673;US 7,566,537;US 7,592,435和W0 07/14835368,这些专利申请中的每一篇均全文以引用方式并入本文。

[0508] 在一些具体实施中,扩增模块和SBS模块可在单个测定协议中操作,其中例如扩增模板核酸并随后将其在同一盒内测序。

[0509] 测序系统4200A还可以允许用户重新配置测定协议。例如,测序系统4200A可以通过用户界面4218将用于修改所确定的协议的选项提供给用户。例如,如果确定生物传感器4212将用于扩增,则测序系统4200A可以请求退火循环的温度。此外,如果用户已提供对于所选测定协议通常不可接受的用户输入,则测序系统4200A可以向用户发出警告。

[0510] 在具体实施中,生物传感器4212包括数百万个传感器(或像素),每个传感器(或像素)在后续的碱基调用循环内生成多个像素信号序列。分析模块4244根据传感器阵列上传感器的逐行和/或逐列位置来检测多个像素信号序列并将它们归属于对应的传感器(或像素)。

[0511] 图42C是用于分析来自测序系统4200A的传感器数据(诸如碱基调用传感器输出)的系统的简化框图。在图42C的示例中,系统包括可配置处理器4246。可配置处理器4246可以与由中央处理单元(CPU)4252(即,主机处理器)执行的运行时程序/逻辑4280协调地执行

碱基调用器(例如,基于神经网络的碱基调用器2900)。测序系统4200A包括生物传感器4212和流通池。流通池可以包括一个或多个区块,其中遗传物质的簇暴露于簇流的序列,该簇流的序列用于引起簇中的反应以识别遗传物质中的碱基。传感器感测流通池的每个区块中该序列的每个循环的反应以提供区块数据。遗传测序是数据密集型操作,其将碱基调用传感器数据转换为在碱基调用操作期间感测到的遗传物质的每个簇的碱基调用序列。

[0512] 该示例中的系统包括执行运行时程序/逻辑4280以协调碱基调用操作的CPU 4252、用于存储区块数据阵列的序列的存储器4248B、由碱基调用操作产生的碱基调用读段,以及碱基调用操作中使用的其它信息。另外,在该图示中,系统包括存储器4248A,以存储配置文件(或多个文件)诸如FPGA位文件和用于配置和重新配置可配置处理器4246的神经网络的模型参数,并且执行神经网络。测序系统4200A可以包括用于配置可配置处理器以及在一些具体实施中的可重新配置处理器的程序,以执行神经网络。

[0513] 测序系统4200A通过总线4289耦接到可配置处理器4246。总线4289可以使用高通量技术来实现,诸如在一个示例中,总线技术与当前由PCI-SIG(PCI特别兴趣小组)维护和开发的PCIe标准(快速外围部件互连)兼容。而且在该示例中,存储器4248A通过总线4293耦接到可配置处理器4246。存储器4248A可以为设置在具有可配置处理器4246的电路板上的板上存储器。存储器4248A用于由可配置处理器4246高速访问在碱基调用操作中使用的的数据。总线4293还可以使用高通量技术诸如与PCIe标准兼容的总线技术来实现。

[0514] 可配置处理器,包括现场可编程门阵列FPGA、粗粒度可重构阵列CGRA以及其他可配置和可重构的设备,可被配置为比使用执行计算机程序的通用处理器可能实现的更有效或更快地实现各种功能。可配置处理器的配置涉及编译功能描述以产生有时称为位流或位文件的配置文件,以及将配置文件分发到处理器上的可配置元件。该配置文件通过将电路配置为设置数据流模式、分布式存储器和其他片上存储器资源的使用、查找表内容、可配置逻辑块和可配置执行单元(如乘法累加单元、可配置互连和可配置阵列的其他元件)的操作,来定义要由可配置处理器执行的逻辑功能。如果配置文件可在现场通过改变加载的配置文件而改变,则可配置处理器是可重构的。例如,配置文件可存储在易失性SRAM元件中、非易失性读写存储器元件中以及它们的组合中,分布在可配置或可重构处理器上的可配置元件阵列中。多种可商购获得的可配置处理器适用于如本文所述的碱基调用操作。示例包括Google的Tensor Processing Unit(TPU)TM、机架解决方案(如GX4 Rackmount SeriesTM、GX9 Rackmount SeriesTM)、NVIDIA DGX-1TM、Microsoft的Stratix V FPGATM、Graphcore的Intelligent Processor Unit(IPU)TM、Qualcomm的具有Snapdragon processorsTM的Zeroth PlatformTM、NVIDIA的VoltaTM、NVIDIA的DRIVE PXTM、NVIDIA的JETSON TX1/TX2 MODULETM、Intel的NirvanaTM、Movidius VPUTM、Fujitsu DPITM、ARM的DynamicIQTM、IBM TrueNorthTM、具有Testa V100sTM的Lambda GPU服务器、Xilinx AlveoTMU200、Xilinx AlveoTMU2190、Xilinx AlveoTMU280、Intel/Altera StratixTMGX2800、Intel/Altera StratixTMGX2800和Intel StratixTMGX10M。在一些示例中,主机CPU可在与可配置处理器相同的集成电路上实现。

[0515] 本文所述的具体实施使用可配置处理器4246来实现基于神经网络的碱基调用器2900。可配置处理器4246的配置文件可以通过使用高级描述语言HDL或寄存器传输级RTL语言规范指定要执行的逻辑功能来实现。可使用被设计用于所选择的可配置处理器的资源来

编译规范以生成配置文件。为了生成可能不是可配置处理器的专用集成电路的设计,可编译相同或相似的规范。

[0516] 因此,在本文所述的所有具体实施中,可配置处理器4246的另选方案包括配置的处理器,该配置的处理器包括专用ASIC或专用集成电路或集成电路组,或片上系统SOC器件,或图形处理单元(GPU)处理器或粗粒度可重新配置架构(CGRA)处理器,该配置的处理器被配置为执行如本文所述的基于神经网络的碱基调用操作。

[0517] 一般来讲,如被配置为执行神经网络的运行的本文所述的可配置处理器和配置的处理器在本文中称为神经网络处理器。

[0518] 在该示例中,可配置处理器4246由使用CPU 4252执行的程序所加载的配置文件配置,或由在可配置处理器上配置可配置元件4291(例如,配置逻辑块(CLB),诸如查找表(LUT)、触发器、计算处理单元(PMU)和计算存储器单元(CMU)、可配置I/O块、可编程互连)的阵列的其它源配置,以执行碱基调用功能。在该示例中,该配置包括数据流逻辑4297,该数据流逻辑耦接到总线4289和总线4293,并且执行用于在碱基调用操作中使用的元件之间分发数据和控制参数的功能。

[0519] 此外,可配置处理器4246配置有数据流逻辑4297以执行基于神经网络的碱基调用器2900。逻辑4297包括多循环执行簇(例如,4279),在该示例中,该多循环执行簇包括执行簇1至执行簇X。可以根据涉及操作的期望通量和可配置处理器4246上的可用资源的权衡来选择多循环执行簇的数量。

[0520] 多循环执行簇通过使用可配置处理器4246上的可配置互连和存储器资源实现的数据流路径4299耦接到数据流逻辑4297。另外,多循环执行簇通过使用例如可配置处理器4246上的可配置互连和存储器资源实现的控制路径4295耦接到数据流逻辑4297,该控制路径提供指示可用执行簇、准备好向可用执行簇提供用于执行基于神经网络的碱基调用器2900的运行的输入单元、准备好向基于神经网络的碱基调用器2900提供经训练的参数、准备好提供碱基调用分类数据的输出块以及用于执行基于神经网络的碱基调用器2900的其它控制数据。

[0521] 可配置处理器4246被配置为使用经训练的参数来执行基于神经网络的碱基调用器2900的运行,以产生用于碱基调用操作的感测循环的分类数据。执行基于神经网络的碱基调用器2900的运行,以产生用于碱基调用操作的受试者感测循环的分类数据。基于神经网络的碱基调用器2900的运行对序列(包括来自N个感测循环的相应感测循环的区块数据的数字N个阵列)进行操作,其中N个感测循环在本文所述的示例中针对时间序列中每个操作的一个碱基位置提供用于不同碱基调用操作的传感器数据。任选地,如果需要,根据正在执行的特定神经网络模型,N个感测循环中的一些可能会失序。数字N可以是大于一的任何数字。在本文所述的一些示例中,N个感测循环中的感测循环表示时间序列中受试者感测循环之前的至少一个感测循环和受试者循环(subject cycle)之后的至少一个感测循环的一组感测循环。本文描述了其中数字N为等于或大于五的整数的示例。

[0522] 数据流逻辑4297被配置为使用用于给定运行的输入单元将区块数据和模型参数的至少一些经训练的参数从存储器4248A移动到用于基于神经网络的碱基调用器2900的运行的可配置处理器4246,该输入单元包括用于N个阵列的空间对准的块的区块数据。输入单元可通过一个DMA操作中的直接存储器存取操作来移动,或者在可用时隙期间与所部署的

神经网络的执行相协调地移动的较小单元中移动。

[0523] 如本文所述的用于感测循环的区块数据可包括具有一个或多个特征的传感器数据阵列。例如,传感器数据可包括两个图像,对这两个图像进行分析以识别在DNA、RNA或其他遗传物质的遗传序列中的碱基位置处的四种碱基中的一种。区块数据还可包括关于图像和传感器的元数据。例如,在碱基调用操作的具体实施中,区块数据可包括关于图像与簇的对准的信息,诸如距中心距离的信息,该距离指示传感器数据阵列中的每个像素距区块上遗传物质的簇的中心的距离。

[0524] 在如下所述的基于神经网络的碱基调用器2900的执行期间,区块数据还可以包括在执行基于神经网络的碱基调用器2900期间产生的数据,该数据被称为中间数据,可被重复使用而不是在基于神经网络的碱基调用器2900的运行期间重新计算。例如,在基于神经网络的碱基调用器2900的执行期间,数据流逻辑4297可以将中间数据代替用于区块数据阵列的给定块的传感器数据写入到存储器4248A。下文更详细地描述了类似于此的具体实施。

[0525] 如图所示,描述了用于分析碱基调用传感器输出的系统,该系统包括可以由运行时程序/逻辑4280访问的存储器(例如,4248A),该存储器存储区块数据,该区块数据包括来自碱基调用操作的感测循环的区块的传感器数据。另外,该系统包括神经网络处理器,诸如对存储器具有访问权限的可配置处理器4246。神经网络处理器被配置为使用经训练参数来执行神经网络的运行,以产生用于感测循环的分类数据。如本文所述,神经网络的运行对来自N个感测循环的相应感测循环(包括受试者循环)的区块数据的N个阵列的序列进行操作,以产生受试者循环的分类数据。提供数据流逻辑4297以使用输入单元(包括来自N个感测循环的相应感测循环的N个阵列的空间对准块的数据)将区块数据和经训练参数从存储器移动到神经网络处理器以用于神经网络的运行。

[0526] 另外,描述了一种系统,其中神经网络处理器可访问存储器,并且包括多个执行簇,所述多个执行簇中的执行簇被配置为执行神经网络。数据流逻辑1997可访问存储器和多个执行簇中的执行簇,以将区块数据的输入单元提供至该多个执行簇中的可用执行簇,该输入单元包括来自相应感测循环(包括受试者感测循环)的区块数据阵列的数字N个空间对准块,并且使执行簇将N个空间对准块应用于神经网络以产生用于受试者感测循环的空间对准块的分类数据的输出块,其中N大于1。

[0527] 图43A是示出碱基调用操作的各方面的简化图,该操作包括由主机处理器执行的运行时程序(例如,运行时逻辑4280)的功能。在该图中,来自流通池的图像传感器的输出在线4300上提供给图像处理线程4301,这些图像处理线程可以对图像执行处理,诸如各个区块的传感器数据阵列中的对准和布置以及图像的重采样,并且可以由为流通池中的每个区块计算区块簇掩膜的过程使用,该过程标识与流通池的对应区块上的遗传物质的簇对应的传感器数据阵列中的像素。根据碱基调用操作的状态,图像处理线程4301的输出在线4302上提供给CPU中的调度逻辑4303,该调度逻辑将区块数据阵列在高速总线4304上路由到数据高速缓存4305(例如,SSD存储装置),或者在高速总线4306上路由到神经网络处理器硬件4307,诸如图19C的可配置处理器1946。经处理和变换的图像可以存储在数据高速缓存4305上,以用于先前使用的感测循环。硬件4307将由神经网络输出的分类数据返回到调度逻辑4303,该调度逻辑将信息传递到数据高速缓存4305,或者在线4308上传递到使用分类数据执行碱基调用和质量分数计算的线程4309,并且可以以标准格式布置用于碱基调用读段的

数据。在线4310上将执行碱基调用和质量分数计算的线程4309的输出提供给线程4311,该线程聚合碱基调用读段,执行其它操作诸如数据压缩,并且将所得的碱基调用输出写入指定目的地以供客户利用。

[0528] 在一些具体实施中,主机可以包括执行硬件4307的输出的最终处理以支持神经网络的线程(未示出)。例如,硬件4307可以提供来自多簇神经网络的最终层的分类数据的输出。主机处理器可以对分类数据执行输出激活功能诸如softmax功能,以配置供碱基调用和质量评分线程4302使用的数据。而且,主机处理器可以执行输入操作(未示出),诸如在输入到硬件4307之前对区块数据进行批量归一化。

[0529] 图43B是可配置处理器1946(诸如图19C的可配置处理器)的配置的简化图。在图43B中,可配置处理器1946包括具有多个高速PCIe接口的FPGA。FPGA配置有封装器4390,该封装器包括参考图19C描述的数据流逻辑1997。封装器4390通过CPU通信链路4377来管理与CPU中的运行时程序的接口和协调,并且经由DRAM通信链路4397来管理与板上DRAM 4399(例如,存储器1448A)的通信。封装器4390中的数据流逻辑1997将通过遍历板上DRAM 4399上的数字N个循环的区块数据阵列而检索到的块数据提供到簇4385,并且从簇4385检索过程数据4387以递送回板上DRAM 4399。封装器4390还管理板上DRAM 4399与主机存储器之间的数据传输,以用于区块数据的输入阵列和分类数据的输出块两者。封装器将线4383上的块数据传输到分配的簇4385。封装器在线4381上将经训练的参数诸如权重和偏置提供到从板上DRAM 4399检索到的簇4385。封装器在线4379上将配置和控制数据提供到簇4385,该簇经由CPU通信链路4377从主机上的运行时程序提供或响应于该运行时程序而生成。簇还可以在线4389上将状态信号提供到封装器4390,该状态信号与来自主机的控制信号协作使用,以管理区块数据阵列的遍历,从而提供空间对准的块数据,并且使用簇4385的资源对块数据执行多循环神经网络。

[0530] 如上所述,在由封装器4390管理的单个可配置处理器上可以存在多个簇,该多个簇被配置用于在区块数据的多个块中的对应块上执行。每个簇可被配置为使用本文所述的多个感测循环的区块数据来提供受试者感测循环中的碱基调用的分类数据。

[0531] 在系统的示例中,可将模型数据(包括内核数据,如过滤器权重和偏置)从主机CPU发送到可配置处理器,使得模型可根据循环数进行更新。举一个代表性示例,碱基调用操作可包括大约数百个感测循环。在一些具体实施中,碱基调用操作可包括双端读段。例如,模型训练参数可以每20个循环(或其他数量的循环)更新一次,或者根据针对特定系统和神经网络模型实现的更新模式来更新。在包括双端读段的一些具体实施中,其中区块上的遗传簇中的给定字符串的序列包括从第一末端沿字符串向下(或向上)延伸的第一部分和从第二末端沿字符串向上(或向下)延伸的第二部分,可在从第一部分到第二部分的过渡中更新经训练参数。

[0532] 在一些示例中,可以将区块的感测数据的多个循环的图像数据从CPU发送到封装器4390。封装器4390可以任选地对感测数据进行一些预处理和变换,并且将信息写入板上DRAM 4399。每个感测循环的输入区块数据可包括传感器数据阵列,包括每个感测循环每个区块大约 4000×3000 个像素或更多,其中两个特征表示区块的两个图像的颜色,并且每个特征每个像素一个或两个字节。对于其中数字N为要在多循环神经网络的每个运行中使用的三个感测循环的具体实施,用于多循环神经网络的每个运行的区块数据阵列可消耗每个

区块大约数百兆字节。在系统的一些具体实施中,区块数据还包括每个区块存储一次的距簇中心的距离(DFC)数据的阵列,或关于传感器数据和区块的其他类型的元数据。

[0533] 在操作中,当多循环簇可用时,封装器将块分配给簇。封装器在区块的遍历中获取区块数据的下一个块,并将其连同适当的控制和配置信息一起发送到所分配的簇。簇可被配置为在可配置处理器上具有足够的存储器,以保存包括来自一些系统中的多个循环的块且正被就地处理的数据块,以及当在各种具体实施中使用乒乓缓冲技术或光栅扫描技术完成对当前块的处理时将被处理的数据块。

[0534] 当分配的簇完成其对当前块的神经网络的运行并产生输出块时,其将发信号通知封装器。封装器将从分配的簇读取输出块,或者另选地,分配的簇将数据推送到封装器。然后,封装器将对DRAM 4399中的经处理的区块组装输出块。当整个区块的处理已完成并且数据的输出块已传输到DRAM时,封装器将区块的经处理输出阵列以指定格式发送回主机/CPU。在一些具体实施中,板上DRAM 4399由封装器4390中的存储器管理逻辑管理。运行时程序可控制测序操作,以连续流的方式完成运行中所有循环的区块数据的所有阵列的分析,从而提供实时分析。

[0535] 图44示出确定CPU上的状态数据并且将状态数据从CPU加载到FPGA以用于碱基调用的一个具体实施。

[0536] 术语和附加具体实施

[0537] 碱基调用包括将荧光标记的标签与分析物结合或连接。分析物可为核苷酸或寡核苷酸,并且标签可用于特定核苷酸类型(A、C、T或G)。激发光被导向具有标签的分析物,并且标签发射可检测的荧光信号或强度发射。强度发射指示由化学附着到分析物的激发标签发射的光子。

[0538] 在包括权利要求书的整个本申请中,当使用诸如或类似于“描绘分析物及其周围背景的强度发射的图像、图像数据或图像区域”的短语时,它们是指附着到分析物的标签的强度发射。本领域的技术人员将会知道,附着的标签的强度发射代表或等同于标签所附着到的分析物的强度发射,因此可互换使用。类似地,分析物的属性是指附着到分析物的标签的属性或来自所连接标签的强度发射的属性。例如,分析物的中心是指由附着到分析物的标签发射的强度发射的中心。在另一个示例中,分析物的周围背景是指由附着到分析物的标签发射的强度发射的周围背景。

[0539] 本申请中引用的所有文献和类似材料,包括但不限于专利、专利申请、文章、书籍、论文和网页,无论这些文献和类似材料的格式如何,都明确地全文以引用方式并入。如果所并入文献和类似材料中的一者或多者与本申请不同或矛盾,包括但不限于所定义的术语、术语用法、所描述的技术等,则以本申请为准。

[0540] 所公开的技术使用神经网络来改善可从核酸样品(诸如核酸模板或其互补序列,例如DNA或RNA多核苷酸或其他核酸样品)获得的核酸序列信息的质量和数量。因此,相对于先前可用的方法,所公开的技术的某些具体实施提供更高通量的多核苷酸测序,例如更高的DNA或RNA序列数据收集速率、更高的序列数据收集效率和/或获得这种序列数据的更低成本。

[0541] 所公开的技术使用神经网络来识别固相核酸簇的中心并且分析在此类簇的测序期间生成的光信号,以明确地区分相邻、邻接或重叠的簇,以便将测序信号分配给单个离散

源簇。因此,这些和相关具体实施允许从高密度簇阵列的区域检索有意义的信息诸如序列数据,其中由于重叠或非常紧密间隔的相邻簇的混杂效应,包括从其发出的重叠信号(例如,如在核酸测序中所用)的效应,先前无法从此类区域获得可用信息。

[0542] 如下文更详细地描述,在某些具体实施中,提供了包含固体载体的组合物,该固体载体具有固定到其上的一个或多个如本文提供的核酸簇。每个簇包含多个相同序列的固定化核酸并且具有可识别中心,该可识别中心具有如本文所提供的可检测中心标记,通过该可检测中心标记可将可识别中心与簇中周围区域中的固定化核酸区分开。本文还描述了用于制造和使用具有可识别中心的此类簇的方法。

[0543] 本发明所公开的具体实施将在许多情况下使用,其中从识别、确定、注释、记录或以其他方式分配簇内基本上中心位置的位置的能力中获得优势,诸如高通量核酸测序、用于将光学或其他信号分配给离散源簇的图像分析算法的开发,以及其中识别固定化核酸簇中心是期望的且有益的其他应用。

[0544] 在某些具体实施中,本发明设想了涉及高通量核酸分析诸如核酸序列测定(例如,“测序”)的方法。示例性的高通量核酸分析包括但不限于从头测序、重新测序、全基因组测序、基因表达分析、基因表达监测、表观遗传分析、基因组甲基化分析、等位基因特异性引物延伸(APSE)、遗传多样性分析、全基因组多态性发现和分析、单核苷酸多态性分析、基于杂交的序列测定方法等。本领域的技术人员将会知道,可使用本发明的方法和组合物分析多种不同的核酸。

[0545] 虽然关于核酸测序描述了本发明的具体实施,但它们适用于分析在不同时间点、空间位置或其他时间或物理视角采集的图像数据的任何领域。例如,本文所述的方法和系统可用于分子和细胞生物学领域,其中在不同的时间点或视角采集来自微阵列、生物标本、细胞、生物体等的图像数据并进行分析。图像可使用本领域已知的任何数量的技术获得,包括但不限于荧光显微镜法、光学显微镜法、共聚焦显微镜法、光学成像、磁共振成像、断层扫描等。又如,可应用本文所述的方法和系统,其中在不同的时间点或视角采集通过监测、空中或卫星成像技术等获得的图像数据并进行分析。该方法和系统尤其可用于分析针对视野获得的图像,其中被观察的分析物在视野中相对于彼此保持在相同位置。然而,该分析物可能具有在单独图像中不同的特征,例如,该分析物可能在视野的单独图像中看起来不同。例如,就在不同图像中检测到的给定分析物的颜色、在不同图像中检测到的给定分析物的信号强度的变化、或者甚至在一个图像中检测到的给定分析物的信号出现以及在另一个图像中检测到的该分析物的信号消失而言,分析物可能看起来不同。

[0546] 本文所述的示例可用于学术或商业分析的各种生物或化学过程和系统。更具体地,本文所述的示例可用于期望检测指示指定反应的事件、属性、质量或特征的各种过程和系统中。例如,本文所述的示例包括光检测设备、生物传感器及其部件,以及与生物传感器一起操作的生物测定系统。在一些示例中,设备、生物传感器和系统可包括流通池和一个或多个光传感器,它们以基本上一体的结构(可移除地或固定地)耦接在一起。

[0547] 这些设备、生物传感器和生物测定系统可被配置为执行可单独或共同检测的多个指定反应。这些设备、生物传感器和生物测定系统可被配置为执行多个循环,其中所述多个指定反应同步发生。例如,这些设备、生物传感器和生物测定系统可用于通过酶操纵和光或图像检测/采集的迭代循环对DNA特征的密集阵列进行测序。因此,这些设备、生物传感器和

生物测定系统(例如,经由一个或多个盒)可包括一个或多个微流体通道,所述一个或多个微流体通道将反应溶液中的试剂或其他反应组分递送到这些设备、生物传感器和生物测定系统的反应位点。在一些示例中,反应溶液可为基本上酸性的,诸如具有小于或等于约5,或小于或等于约4,或小于或等于约3的pH。在一些其他示例中,反应溶液可为基本上碱的/碱性的,诸如具有大于或等于约8,或大于或等于约9,或大于或等于约10的pH。如本文所用,术语“酸度”及其语法变型是指小于约7的pH值,并且术语“碱度”、“碱性”及其语法变型是指大于约7的pH值。

[0548] 在一些示例中,反应位点以预定方式提供或间隔开,诸如以均匀或重复的模式提供或间隔开。在一些其他示例中,反应位点是随机分布的。反应位点中的每个位点可与一个或多个光导以及检测来自相关联反应位点的光的一个或多个光传感器相关联。在一些示例中,反应位点位于反应凹槽或反应室中,这可至少部分地分隔其中的指定反应。

[0549] 如本文所用,“指定反应”包括感兴趣的化学或生物物质(例如,感兴趣的分析物)的化学、电、物理或光学性质(或质量)中的至少一者的变化。在特定示例中,指定反应为阳性结合事件,例如,将荧光标记的生物分子与感兴趣的分析物结合。更一般地,指定反应可以是化学转化、化学变化或化学相互作用。指定反应也可电属性的变化。在特定示例中,指定反应包括将荧光标记的分子与分析物结合。分析物可为寡核苷酸,并且荧光标记的分子可为核苷酸。当激发光被导向具有标记核苷酸的寡核苷酸,并且荧光团发出可检测的荧光信号时,可检测到指定反应。在另选的示例中,检测到的荧光是化学发光或生物发光的结果。指定反应还可例如通过使供体荧光团接近受体荧光团来增加荧光(或Förster)共振能量转移(FRET),通过分离供体荧光团和受体荧光团来降低FRET,通过分离淬灭基团与荧光团来增加荧光,或通过共定位淬灭基团和荧光团来减少荧光。

[0550] 如本文所用,“反应溶液”、“反应组分”或“反应物”包括可用于获得至少一种指定反应的任何物质。例如,可能的反应组分包括例如试剂、酶、样品、其他生物分子和缓冲液。可将反应组分递送至溶液中的反应位点和/或固定在反应位点处。反应组分可直接或间接地与另一种物质相互作用,诸如固定在反应位点的感兴趣的分析物。如上所述,反应溶液可以是基本上酸性的(即,包括相对高的酸度)(例如,具有小于或等于约5的pH,小于或等于约4的pH,或小于或等于约3的pH)或基本上碱的/碱性的(即,包括相对高的碱性/碱度)(例如,具有大于或等于约8的pH,大于或等于约9的pH,或大于或等于约10的pH)。

[0551] 如本文所用,术语“反应位点”是可发生至少一个指定反应的局部区域。反应位点可包括其上可固定物质的反应结构或基底的支撑表面。例如,反应位点可包括其上具有反应组分(诸如其上的核酸群体)的反应结构的表面(可位于流通池的通道中)。在一些此类示例中,群体中的核酸具有相同的序列,例如为单链或双链模板的克隆拷贝。然而,在一些示例中,反应位点可仅包含单个核酸分子,例如单链或双链形式。

[0552] 多个反应位点可沿着反应结构随机分布或以预定方式布置(例如,在矩阵中并排布置,诸如在微阵列中)。反应位点还可包括反应室或反应凹槽,其至少部分地限定了被配置为分隔指定反应的空间区域或体积。如本文所用,术语“反应室”或“反应凹槽”包括支撑结构的限定空间区域(其通常与流动通道流体连通)。反应凹槽可至少部分地与其他或空间区域的周围环境隔开。例如,多个反应凹槽可通过共用壁诸如检测表面彼此隔开。作为更具体的示例,反应凹槽可以是包括由检测表面的内表面限定的凹痕、凹坑、孔、槽、腔或凹陷的

纳米孔,并且具有开口或孔隙(即,为敞开的),使得纳米孔可与流动通道流体连通。

[0553] 在一些示例中,反应结构的反应凹槽的尺寸和形状相对于固体(包括半固体)被设定成使得固体可完全或部分地插入其中。例如,反应凹槽的尺寸和形状可被设定成容纳捕获小珠。该捕获小珠可在其上具有克隆扩增的DNA或其他物质。或者,反应凹槽的尺寸和形状可被设定成接纳大约数量的小珠或固体基板。又如,反应凹槽可填充有多孔凝胶或物质,该多孔凝胶或物质被配置为控制扩散或过滤可流入反应凹槽的流体或溶液。

[0554] 在一些示例中,光传感器(例如,光电二极管)与对应的反应位点相关联。与反应位点相关联的光传感器被配置为当在相关联的反应位点处已发生指定反应时,经由至少一个光导检测来自相关联的反应位点的光发射。在一些情况下,多个光传感器(例如,光检测或相机设备的若干像素)可与单个反应位点相关联。在其他情况下,单个光传感器(例如,单个像素)可与单个反应位点或与一组反应位点相关联。光传感器、反应位点和生物传感器的其他特征可被配置为使得光中的至少一些被光传感器直接检测到而不被反射。

[0555] 如本文所用,“生物或化学物质”包括生物分子、感兴趣的样品、感兴趣的分析物和其他化合物。生物或化学物质可用于检测、识别或分析其他化合物,或者用作研究或分析其他化合物的中间物。在特定示例中,生物或化学物质包括生物分子。如本文所用,“生物分子”包括生物聚合物、核苷、核酸、多核苷酸、寡核苷酸、蛋白质、酶、多肽、抗体、抗原、配体、受体、多糖、碳水化合物、多磷酸盐、细胞、组织、生物体或它们的片段中的至少一种,或任何其他生物活性化合物诸如前述物质的类似物或模拟物。在另一个示例中,生物或化学物质或生物分子包括在偶联反应中用于检测另一反应的产物的酶或试剂,诸如酶或试剂,诸如用于在焦磷酸测序反应中检测焦磷酸盐的酶或试剂。可用于焦磷酸盐检测的酶和试剂描述于例如美国专利公布号2005/0244870A1中,其全文以引用方式并入。

[0556] 生物分子、样品以及生物或化学物质可为天然存在的或合成的,并且可悬浮在反应凹槽或区域内的溶液或混合物中。生物分子、样品以及生物或化学物质也可结合至固相或凝胶材料。生物分子、样品以及生物或化学物质也可包括药物组合物。在一些情况下,感兴趣的生物分子、样品以及生物或化学物质可称为靶标、探针或分析物。

[0557] 如本文所用,“生物传感器”包括具有多个反应位点的反应结构的设备,该设备被配置为检测在反应位点处或附近发生的指定反应。生物传感器可包括固态光检测或“成像”设备(例如,CCD或CMOS光检测设备)以及任选地安装到其上的流通池。流通池可包括与反应位点流体连通的至少一个流动通道。作为一个具体示例,生物传感器被配置为流体耦接和电耦接到生物测定系统。生物测定系统可根据预定方案(例如,边合成边测序)将反应溶液递送到反应位点,并且执行多个成像事件。例如,生物测定系统可引导反应溶液沿反应位点流动。反应溶液中的至少一种可包含四种类型的具有相同或不同荧光标记的核苷酸。核苷酸可结合至反应位点,诸如结合至反应位点处的对应寡核苷酸。然后,生物测定系统可使用激发光源(例如,固态光源,诸如发光二极管(LED))照亮反应位点。激发光可具有预定的一个或多个波长,包括一个波长范围。由入射激发光激发的荧光标记可提供可由光传感器检测的发射信号(例如,与激发光不同并且可能彼此不同的一个或多个波长的光)。

[0558] 如本文所用,当关于生物分子或生物或化学物质使用时,术语“固定的”包括在分子水平上基本上将生物分子或生物或化学物质连接到表面,诸如连接到光检测设备或反应结构的检测表面。例如,可使用吸附技术将生物分子或生物或化学物质固定到反应结构的

表面,所述吸附技术包括非共价相互作用(例如,静电力、范德华力以及疏水界面的脱水)和共价结合技术,其中官能团或接头有利于将生物分子连接到表面。将生物分子或生物或化学物质固定到表面可基于表面的属性、携带生物分子或生物或化学物质的液体介质以及生物分子或生物或化学物质本身的属性。在一些情况下,表面可被官能化(例如,化学或物理改性),以有利于将生物分子(或生物或化学物质)固定到表面。

[0559] 在一些示例中,核酸可固定到反应结构,诸如固定到其反应凹槽的表面。在特定示例中,本文所述的设备、生物传感器、生物测定系统和方法可包括使用天然核苷酸以及被配置为与天然核苷酸相互作用的酶。天然核苷酸包括例如核糖核苷酸或脱氧核糖核苷酸。天然核苷酸可为单磷酸盐、二磷酸盐或三磷酸盐形式,并且可具有选自腺嘌呤(A)、胸腺嘧啶(T)、尿嘧啶(U)、鸟嘌呤(G)或胞嘧啶(C)的碱基。然而,应当理解,可使用非天然核苷酸、经修饰的核苷酸或前述核苷酸的类似物。

[0560] 如上所述,生物分子或生物或化学物质可固定在反应结构的反应凹槽中的反应位点处。此类生物分子或生物物质可通过过盈配合、粘附、共价键或截留物理地保持或固定在反应凹槽内。可设置在反应凹槽内的物品或固体的示例包括聚合物小珠、微丸、琼脂糖凝胶、粉末、量子点或可被压缩和/或保持在反应室内的其他固体。在某些具体实施中,反应凹槽可涂覆或填充有能够共价结合DNA寡核苷酸的水凝胶层。在特定示例中,核酸超结构(诸如DNA球)可例如通过附接至反应凹槽的内表面或通过停留在反应凹槽内的液体中而设置在反应凹槽中或反应凹槽处。可进行DNA球或其他核酸超结构,然后将其设置在反应凹槽中或反应凹槽处。另选地,DNA球可在反应凹槽处原位合成。固定在反应凹槽中的物质可为固态、液态或气态。

[0561] 如本文所用,术语“分析物”旨在表示图案中的点或区域,其可根据相对位置区别于其他点或区域。单个分析物可包括一种或多种特定类型的分子。例如,分析物可包括具有特定序列的单个靶核酸分子,或者分析物可包括具有相同序列(和/或其互补序列)的若干核酸分子。位于图案的不同分析物处的不同分子可根据分析物在图案中的位置而彼此区分。示例性分析物包括但不限于基板中的孔、基板中或基板上的小珠(或其他颗粒)、基板的突出部、基板上的脊、基板上的凝胶材料垫或基板中的通道。

[0562] 待检测、表征或鉴定的多种靶分析物中的任一种可用于本文所述的设备、系统或方法中。示例性分析物包括但不限于核酸(例如,DNA、RNA或它们的类似物)、蛋白质、多糖、细胞、抗体、表位、受体、配体、酶(例如,激酶、磷酸酶或聚合酶)、小分子候选药物、细胞、病毒、生物体等。

[0563] 术语“分析物”、“核酸”、“核酸分子”和“多核苷酸”在本文中可互换使用。在各种具体实施中,核酸可用作如本文所提供的模板(例如,核酸模板,或与核酸模板互补的核酸互补序列)以用于特定类型的核酸分析,包括但不限于核酸扩增、核酸表达分析和/或核酸序列测定或它们的合适组合。在某些具体实施中,核酸包括例如3'-5'磷酸二酯或其他键中的脱氧核糖核苷酸的线性聚合物,诸如脱氧核糖核酸(DNA),例如单链和双链DNA、基因组DNA、拷贝DNA或互补DNA(cDNA)、重组DNA或任何形式的合成或修饰DNA。在其他具体实施中,核酸包括例如3'-5'磷酸二酯或其他键中的核糖核苷酸的线性聚合物,诸如核糖核酸(RNA),例如单链和双链RNA、信使(mRNA)、拷贝RNA或互补RNA(cRNA)、选择性剪接的mRNA、核糖体RNA、核仁小RNA(snoRNA)、微小RNA(miRNA)、小干扰RNA(siRNA)、piwi RNA(piRNA)或任何形式的

合成或修饰的RNA。用于本发明的组合物和方法中的核酸的长度可变化,并且可以是完整的或全长的分子或片段或较大核酸分子的较小部分。在特定具体实施中,核酸可具有一种或多种可检测标记,如本文别处所述。

[0564] 术语“分析物”、“簇”、“核酸簇”、“核酸群体”和“DNA簇”可互换使用,指附着到固体载体的核酸模板和/或其互补序列的多个拷贝。通常并且在某些优选的具体实施中,核酸簇包含模板核酸和/或其互补序列的多个拷贝,这些多个拷贝通过它们的5'末端附着到固体载体。构成核酸簇的核酸链的拷贝可为单链或双链形式。簇中存在的核酸模板的拷贝可在例如由于标记部分的存在而彼此不同的对应位置处具有核苷酸。对应的位置还可包含具有不同化学结构但具有相似Watson-Crick碱基配对属性的类似结构,诸如尿嘧啶和胸腺嘧啶的情况。

[0565] 核酸群体也可称为“核酸簇”。核酸群体可任选地通过簇扩增或桥式扩增技术产生,如本文其他地方进一步详细阐述的。靶序列的多个重复可存在于单个核酸分子中,诸如使用滚环扩增程序产生的多联体。

[0566] 根据所用的条件,本发明的核酸簇可具有不同的形状、大小和密度。例如,簇可具有基本上圆形、多边形、圆环形或环形的形状。核酸簇的直径可被设计成约0.2 μm 至约6 μm 、约0.3 μm 至约4 μm 、约0.4 μm 至约3 μm 、约0.5 μm 至约2 μm 、约0.75 μm 至约1.5 μm 或任何介于其间的直径。在特定具体实施中,核酸簇的直径为约0.5 μm 、约1 μm 、约1.5 μm 、约2 μm 、约2.5 μm 、约3 μm 、约4 μm 、约5 μm 或约6 μm 。核酸簇的直径可受到多个参数的影响,包括但不限于产生簇时进行的扩增循环数、核酸模板的长度或附着到其上形成有簇的表面的引物的密度。核酸簇的密度可被设计成通常在0.1/mm²、1/mm²、10/mm²、100/mm²、1,000/mm²、10,000/mm²至100,000/mm²的范围内。本发明还部分地设想了较高密度的核酸簇,例如100,000/mm²至1,000,000/mm²和1,000,000/mm²至10,000,000/mm²。

[0567] 如本文所用,“分析物”是标本或视野内的感兴趣区域。当与微阵列设备或其他分子分析设备结合使用时,分析物是指由类似或相同分子占据的区域。例如,分析物可以是扩增的寡核苷酸或具有相同或相似序列的任何其他组的多核苷酸或多肽。在其他具体实施中,分析物可为占据标本上的物理区域的任何元素或元素组。例如,分析物可以是一片土地、一片水域等。当对分析物成像时,每种分析物都会有一定的面积。因此,在许多具体实施中,分析物不仅仅是一个像素。

[0568] 分析物之间的距离可用任意多种方式描述。在一些具体实施中,分析物之间的距离可被描述为从一种分析物的中心到另一种分析物的中心。在其他具体实施中,距离可被描述为从一种分析物的边缘到另一种分析物的边缘,或者在每种分析物的最外侧可识别点之间。分析物的边缘可被描述为芯片上的理论或实际物理边界,或分析物边界内的某个点。在其他具体实施中,距离可相对于标本上的固定点或标本的图像中的固定点来描述。

[0569] 一般来讲,本文将关于分析方法来描述若干具体实施。应当理解,还提供了用于以自动化或半自动化的方式执行方法的系统。因此,本公开提供了基于神经网络的模板生成和碱基调用系统,其中该系统可包括处理器;存储设备;以及用于图像分析的程序,该程序包括用于执行本文所述方法中的一种或多种方法的指令。因此,本文所述的方法可在例如具有本文所述或本领域已知的部件的计算机上进行。

[0570] 本文示出的方法和系统可用于分析多种对象中的任一种。特别有用的对象是具有

连接的分析物的固体载体或固相表面。当用于在xy平面中具有分析物的重复图案的对象时,本文示出的方法和系统提供优点。一个示例是具有细胞、病毒、核酸、蛋白质、抗体、碳水化合物、小分子(诸如候选药物)、生物活性分子或其他感兴趣的分析物的连接集合的微阵列。

[0571] 已经开发出越来越多的应用,用于具有生物分子(诸如核酸和多肽)的分析物的阵列。此类微阵列通常包括脱氧核糖核酸(DNA)或核糖核酸(RNA)探针。这些探针对人类和其他生物体中存在的核苷酸序列为特异性的。在某些应用中,例如,可将单个DNA或RNA探针连接到阵列的单个分析物上。可将诸如来自已知的人或生物体的试样暴露于阵列,使得靶核酸(例如,基因片段、mRNA或其扩增子)与阵列中相应分析物处的互补探针杂交。探针可在靶标特异性过程中进行标记(例如,由于存在于靶核酸上的标记或由于在分析物中以杂交形式存在的探针或靶标的酶标记)。然后可通过扫描分析物上特定频率的光来检查阵列,以识别样品中存在哪些靶核酸。

[0572] 生物微阵列可用于基因测序和类似应用。一般来讲,基因测序包括确定靶核酸(诸如DNA或RNA的片段)的长度中的核苷酸顺序。相对短的序列通常在每种分析物处进行测序,并且所得的序列信息可用于各种生物信息学方法中以将序列片段逻辑地拟合在一起,从而可靠地确定从其衍生出片段的遗传物质的长度范围更广的序列。已开发出用于特征性片段的自动化的、基于计算机的算法,并且最近已将其用于基因组作图、基因及其功能的鉴定等。微阵列尤其可用于表征基因组含量,因为存在大量变体,并且这取代了对单个探针和靶标进行许多实验的选择方案。微阵列是用于以实际方式进行此类研究的理想形式。

[0573] 本领域已知的多种分析物阵列(也称为“微阵列”)中的任一种均可用于本文所述的方法或系统中。典型的阵列包含分析物,每种分析物具有单独的探针或探针群。在后一种情况下,每种分析物处的探针群通常是同质的,具有单一种类的探针。例如,就核酸阵列而言,每种分析物可具有多个核酸分子,每个核酸分子具有共同的序列。然而,在一些具体实施中,阵列的每种分析物处的探针群可以是异质的。类似地,蛋白质阵列可具有含单个蛋白质或蛋白质群的分析物,单个蛋白质或蛋白质群通常但不总是具有相同的氨基酸序列。探针可例如通过探针与表面的共价键合或通过探针与表面的非共价相互作用附着到阵列的表面。在一些具体实施中,探针诸如核酸分子可经由凝胶层附着到表面,如例如在以下专利申请中所述:美国专利申请序列号13/784,368和美国专利申请公布号2011/0059865A1,这些文献中的每一篇文献均以引用方式并入本文。

[0574] 示例性阵列包括但不限于得自Illumina公司(San Diego, Calif.)的BeadChip阵列或其它阵列,诸如其中探针附着到存在于表面上小珠(例如,表面上的孔中的小珠)的那些阵列,诸如美国专利号6,266,459;6,355,431;6,770,441;6,859,570;或7,622,294;或PCT公布号W0 00/63437,这些专利申请中的每一篇均以引用方式并入本文。可使用的可商购获得的微阵列的其他示例包括,例如, Affymetrix® GeneChip®微阵列或根据有时称为VLSIPS™(超大规模固定化聚合物合成)技术合成的其他微阵列。点状微阵列也可用于根据本发明的一些具体实施的方法或系统中。示例性点状微阵列是得自Amersham Biosciences的CodeLink™阵列。可用的另一个微阵列是使用喷墨印刷方法(诸如得自Agilent Technologies的SurePrint™技术)制造的微阵列。

[0575] 其他可用的阵列包括用于核酸测序应用的阵列。例如,具有基因组片段扩增子的

阵列(通常称为簇)是特别有用的,诸如秒数于以下各者中的那些情况:在Bentley等人, Nature 456:53-59(2008), WO 04/018497; WO91/06678; WO 07/123744; U.S. Pat. No. 7,329,492; 7,211,414; 7,315,019; 7,405,281, or 7,057,026; 或美国专利申请公布号2008/0108082A1, 这些文献中的每一篇文献均以引用方式并入本文。可用于核酸测序的另一种类型的阵列是由乳液PCR技术产生的颗粒的阵列。示例描述于Dressman等人, Proc. Natl. Acad. Sci. USA 100:8817-8822(2003), WO 05/010145、美国专利申请公布号2005/0130173, 或美国专利申请公布号2005/0064460中, 这些文献中的每一篇均全文以引用方式并入本文。

[0576] 用于核酸测序的阵列通常具有核酸分析物的随机空间模式。例如, 得自Illumina公司(San Diego, Calif.)的HiSeq或MiSeq测序平台利用流通池, 在该流通池上通过随机接种然后桥式扩增形成核酸阵列。然而, 图案化阵列也可用于核酸测序或其他分析应用。示例性图案化阵列、其制造方法及其使用方法阐述于美国序列号13/787,396; 美国序列号13/783,043; 美国序列号13/784,368和美国专利申请公布号2013/0116153A1; 和美国专利申请公布号2012/0316086A1; 这些文献中的每一篇文献均以引用方式并入本文。此类图案化阵列的分析物可用于捕获单个核酸模板分子以进行接种, 随后例如经由桥式扩增形成同质群体。此类图案化阵列尤其可用于核酸测序应用。

[0577] 可选择阵列(或本文的方法或系统中使用的其他对象)上的分析物的尺寸以适合特定应用。例如, 在一些具体实施中, 阵列的分析物可具有仅容纳单个核酸分子的尺寸。具有多个在该尺寸范围内的分析物的表面可用于构造分子阵列, 从而以单分子分辨率进行检测。该尺寸范围内的分析物也可用于具有分析物的阵列中, 这些分析物各自包含核酸分子的群体。因此, 阵列的分析物可各自具有不大于约 1mm^2 、不大于约 $500\mu\text{m}^2$ 、不大于约 $100\mu\text{m}^2$ 、不大于约 $10\mu\text{m}^2$ 、不大于约 $1\mu\text{m}^2$ 、不大于约 500nm^2 、或不大于约 100nm^2 、不大于约 10nm^2 、不大于约 5nm^2 或不大于约 1nm^2 的面积。另选地或除此之外, 阵列的分析物将不小于约 1mm^2 、不小于约 $500\mu\text{m}^2$ 、不小于约 $100\mu\text{m}^2$ 、不小于约 $10\mu\text{m}^2$ 、不小于约 $1\mu\text{m}^2$ 、不小于约 500nm^2 、不小于约 100nm^2 、不小于约 10nm^2 、不小于约 5nm^2 或不小于约 1nm^2 。实际上, 分析物可具有在选自上文所例示的那些上限和下限之间的范围内的尺寸。尽管已关于核酸和核酸的尺度举例说明了表面的分析物的若干尺寸范围, 但应当理解, 这些尺寸范围内的分析物可用于不包括核酸的应用。还应当理解, 分析物的尺寸不必局限于核酸应用中使用的尺度。

[0578] 对于包括具有多种分析物的对象(诸如, 分析物的阵列)的具体实施而言, 分析物可以是离散的, 彼此之间有间隔。可用于本发明的阵列可具有由至多 $100\mu\text{m}$ 、 $50\mu\text{m}$ 、 $10\mu\text{m}$ 、 $5\mu\text{m}$ 、 $1\mu\text{m}$ 、 $0.5\mu\text{m}$ 或更小的边缘到边缘距离隔开的分析物。另选地或除此之外, 阵列可具有由至少 $0.5\mu\text{m}$ 、 $1\mu\text{m}$ 、 $5\mu\text{m}$ 、 $10\mu\text{m}$ 、 $50\mu\text{m}$ 、 $100\mu\text{m}$ 或更大的边缘到边缘距离隔开的分析物。这些范围可适用于分析物的平均边缘到边缘间距以及最小或最大间距。

[0579] 在一些具体实施中, 阵列的分析物不必是离散的, 相反相邻的分析物可彼此邻接。无论分析物是否为离散的, 分析物的尺寸和/或分析物的间距都可变化, 使得阵列可具有期望的密度。例如, 规则图案中的平均分析物间距可为至多 $100\mu\text{m}$ 、 $50\mu\text{m}$ 、 $10\mu\text{m}$ 、 $5\mu\text{m}$ 、 $1\mu\text{m}$ 、 $0.5\mu\text{m}$ 或更小。另选地或除此之外, 规则图案中的平均分析物间距可为至少 $0.5\mu\text{m}$ 、 $1\mu\text{m}$ 、 $5\mu\text{m}$ 、 $10\mu\text{m}$ 、 $50\mu\text{m}$ 、 $100\mu\text{m}$ 或更大。这些范围也可适用于规则图案的最大或最小间距。例如, 规则图案的最大分析物间距可为至多 $100\mu\text{m}$ 、 $50\mu\text{m}$ 、 $10\mu\text{m}$ 、 $5\mu\text{m}$ 、 $1\mu\text{m}$ 、 $0.5\mu\text{m}$ 或更小; 并且/或者规则图案中的

最小分析物间距可为至少 $0.5\mu\text{m}$ 、 $1\mu\text{m}$ 、 $5\mu\text{m}$ 、 $10\mu\text{m}$ 、 $50\mu\text{m}$ 、 $100\mu\text{m}$ 或更大。

[0580] 阵列中分析物的密度也可根据每单位面积存在的分析物数量来理解。例如,阵列的平均分析物密度可为至少约 1×10^3 个分析物/ mm^2 、 1×10^4 个分析物/ mm^2 、 1×10^5 个分析物/ mm^2 、 1×10^6 个分析物/ mm^2 、 1×10^7 个分析物/ mm^2 、 1×10^8 个分析物/ mm^2 或 1×10^9 个分析物/ mm^2 或更高。另选地或除此之外,阵列的平均分析物密度可为至多约 1×10^9 个分析物/ mm^2 、 1×10^8 个分析物/ mm^2 、 1×10^7 个分析物/ mm^2 、 1×10^6 个分析物/ mm^2 、 1×10^5 个分析物/ mm^2 、 1×10^4 个分析物/ mm^2 或 1×10^3 个分析物/ mm^2 或更低。

[0581] 上述范围可适用于规则图案的全部或部分,包括例如分析物阵列的全部或部分。

[0582] 图案中的分析物可具有多种形状中的任一种。例如,当在二维平面内(诸如在阵列的表面上)观察时,分析物可看起来为圆形、环形、椭圆形、矩形、正方形、对称的、不对称的、三角形、多边形等。分析物可以规则的重复图案布置,包括例如六边形或直线形图案。可选择图案以实现所需的填充水平。例如,圆形分析物以六边形布置进行最佳填充。当然,其他填充布置也可用于圆形分析物,并且反之亦然。

[0583] 图案可根据在形成该图案的最小几何单元的子集中存在的分析物的数量来表征。该子集可包括例如至少约2、3、4、5、6、10种或更多种分析物。根据分析物的尺寸和密度,几何单元可占据小于 1mm^2 、 $500\mu\text{m}^2$ 、 $100\mu\text{m}^2$ 、 $50\mu\text{m}^2$ 、 $10\mu\text{m}^2$ 、 $1\mu\text{m}^2$ 、 500nm^2 、 100nm^2 、 50nm^2 、 10nm^2 或更小的面积。另选地或除此之外,几何单元可占据大于 10nm^2 、 50nm^2 、 100nm^2 、 500nm^2 、 $1\mu\text{m}^2$ 、 $10\mu\text{m}^2$ 、 $50\mu\text{m}^2$ 、 $100\mu\text{m}^2$ 、 $500\mu\text{m}^2$ 、 1mm^2 或更大的面积。几何单元中的分析物的特征(诸如形状、尺寸、间距等)可选自本文针对阵列或图案中的分析物更一般地阐述的那些。

[0584] 具有分析物的规则图案的阵列关于分析物的相对位置可为有序的,但关于每种分析物的一个或多个其他特征为随机的。例如,就核酸阵列而言,核酸分析物关于其相对位置可为有序的,但关于人们对存在于任何特定分析物处的核酸物质的序列的了解为随机的。作为更具体的示例,通过用模板核酸接种分析物的重复图案,并在每种分析物处扩增模板以在分析物处形成模板的拷贝(例如,通过簇扩增或桥式扩增)形成的核酸阵列将具有核酸分析物的规则图案,但关于核酸序列在整个阵列中的分布将是随机的。因此,通常在阵列上检测核酸材料的存在可产生分析物的重复图案,而序列特异性检测可产生信号在整个阵列上的非重复分布。

[0585] 应当理解,本文对图案、顺序、随机性等的描述不仅涉及对象上的分析物,诸如阵列上的分析物,而且还涉及图像中的分析物。因此,图案、顺序、随机性等可以用于存储、操纵或传送图像数据的多种格式中的任一种存在,包括但不限于计算机可读介质或计算机部件,诸如图形用户界面或其他输出设备。

[0586] 如本文所用,术语“图像”旨在表示对象的全部或部分的表示。该表示可以是光学检测的再现。例如,可从荧光、发光、散射或吸收信号获得图像。存在于图像中的对象的部分可以是对象的表面或其他xy平面。通常,图像为二维表示,但在一些情况下,图像中的信息可来源于3个或更多个维度。图像不需要包括光学检测的信号。相反,可存在非光信号。图像可以计算机可读格式或介质提供,诸如在本文其他地方阐述的那些中的一者或多者。

[0587] 如本文所用,“图像”是指标本或其他对象的至少一部分的再现或表示。在一些具体实施中,再现是例如由相机或其他光学检测器产生的光学再现。该再现可以是非光学再现,例如,从纳米孔分析物阵列获得的电信号的表示或从离子敏感型CMOS检测器获得的电

信号的表示。在特定具体实施中,非光学再现可从本文阐述的方法或设备中排除。图像可具有能够区分以多种间距中的任一种存在的标本的分析物的分辨率,所述间距包括例如间隔小于100 μm 、50 μm 、10 μm 、5 μm 、1 μm 或0.5 μm 的那些间距。

[0588] 如本文所用,“获取”、“采集”等术语是指获得图像文件的过程的任何部分。在一些具体实施中,数据采集可包括生成标本的图像,查找标本中的信号,指示检测设备查找或生成信号的图像,给出用于进一步分析或转换图像文件的指令,以及图像文件的任何数量的转换或操纵。

[0589] 如本文所用,术语“模板”是指信号或分析物之间的位置或关系的表示。因此,在一些具体实施中,模板是具有对应于标本中的分析物的信号的表示的物理网格。在一些具体实施中,模板可以是指示与分析物对应的位置的图表、表格、文本文件或其他计算机文件。在本文呈现的具体实施中,生成模板以便跟踪在不同参考点处捕获的标本的图像集上标本的分析物的位置。例如,模板可以是一组x,y坐标或描述一种分析物相对于另一种分析物的方向和/或距离的一组值。

[0590] 如本文所用,术语“标本”可指捕获图像的对象或对象的区域。例如,在拍摄地球表面的图像的具体实施中,一片土地可为标本。在流通池中进行生物分子的分析的其他具体实施中,流通池可被分成任何数量的子部分,每个子部分可为标本。例如,流通池可被分成各种流动通道或槽道,并且每个槽道可被进一步分成2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、110、120、140、160、180、200、400、600、800、1000或更多个单独的成像区域。流通池的一个示例具有8个槽道,其中每个槽道被分成120个标本或区块。在另一个具体实施中,标本可由多个区块或甚至整个流通池组成。因此,每个标本的图像可表示较大表面的成像区域。

[0591] 应当理解,本文所述的对范围和序数列表的引用不仅包括所枚举的数字,而且还包括所枚举的数字之间的所有实数。

[0592] 如本文所用,“参考点”是指图像之间的任何时间或物理区别。在优选的具体实施中,参考点为时间点。在更优选的具体实施中,参考点是测序反应期间的时间点或循环。然而,术语“参考点”可包括区分或分离图像的其他方面,诸如角度方面、旋转方面、时间方面或可区分或分离图像的其他方面。

[0593] 如本文所用,“图像子集”是指集合内的图像集。例如,子集可包含1个、2个、3个、4个、6个、8个、10个、12个、14个、16个、18个、20个、30个、40个、50个、60个图像或选自图像集的任意数量个图像。在特定具体实施中,子集可包含不超过1个、2个、3个、4个、6个、8个、10个、12个、14个、16个、18个、20个、30个、40个、50个、60个图像或选自图像集的任意数量个图像。在优选的具体实施中,从一个或多个测序循环获得图像,其中四个图像与每个循环相关。因此,例如,子集可以是通过四个循环获得的一组16个图像。

[0594] 碱基是指如下核苷酸碱基或核苷酸,A(腺嘌呤)、C(胞嘧啶)、T(胸腺嘧啶)或G(鸟嘌呤)。本专利申请可互换使用“碱基”和“核苷酸”。

[0595] 术语“染色体”是指活细胞的携带遗传性的基因载体,其来源于包含DNA和蛋白质组分(尤其是组蛋白)的染色质链。本文采用了常规的国际公认的个体人类基因组染色体编号系统。

[0596] 术语“位点”是指参考基因组上的独特位置(例如,染色体ID、染色体位置和取向)。

在一些具体实施中,位点可为残基、序列标签或片段在序列上的位置。术语“基因位点”可用于指核酸序列或多态性在参考染色体上的特定位置。

[0597] 本文的术语“样本”是指通常来源于生物流体、细胞、组织、器官或生物体且包含核酸或核酸混合物的样本,该核酸或核酸混合物包含待测序和/或定相的至少一种核酸序列。此类样本包括但不限于痰/口腔液、羊水、血液、血液级分、细针活检样本(例如,外科活检、细针活检等)、尿液、腹膜液、胸膜液、组织外植体、器官培养物和其他任何组织或细胞制剂,或其级分或衍生物,或从其分离的级分或衍生物。虽然样本通常取自人类受试者(例如,患者),但样本可取自具有染色体的任何生物体,包括但不限于狗、猫、马、山羊、绵羊、牛、猪等。样本可按从生物来源获得的原样直接使用,或者经过预处理以改变样本的性质后使用。例如,此类预处理可包括由血液制备血浆、稀释粘性流体等。预处理的方法还可涉及但不限于过滤、沉淀、稀释、蒸馏、混合、离心、冷冻、冻干、浓缩、扩增、核酸片段化、干扰组分的灭活、添加试剂、裂解等。

[0598] 术语“序列”包括或代表彼此偶联的核苷酸链。核苷酸可基于DNA或RNA。应当理解,一个序列可包括多个子序列。例如,单个序列(例如,PCR扩增子的序列)可具有350个核苷酸。样本读段可包括这350个核苷酸内的多个子序列。例如,样本读段可包括具有例如20-50个核苷酸的第一侧翼子序列和第二侧翼子序列。第一侧翼子序列和第二侧翼子序列可位于具有对应子序列(例如,40个-100个核苷酸)的重复片段的任一侧上。每个侧翼子序列可包括引物子序列(例如,10个-30个核苷酸)(或包括其部分)。为了便于阅读,术语“子序列”将被称为“序列”,但应当理解,两个序列不一定在共同链上彼此分开。为了区分本文所述的各种序列,可给序列赋予不同的标记(例如,靶序列、引物序列、侧翼序列、参考序列等)。其他术语诸如“等位基因”可被赋予不同的标记,以区分相似的对象。本申请可互换使用“读段”和“序列读段”。

[0599] 术语“双端测序”是指对靶片段的两端进行测序的测序方法。双端测序可有助于检测基因组重排和重复片段,以及基因融合和新转录本。用于双端测序的方法描述于PCT公布W007010252、PCT申请序列号PCTGB2007/003798和美国专利申请公布号US2009/0088327中,这些专利中的每一篇均以引用方式并入本文。在一个示例中,可如下执行一系列操作;(a)生成核酸簇;(b)使核酸线性化;(c)使第一测序引物杂交,并如上所述进行延伸、扫描和解封闭的重复循环;(d)通过合成互补拷贝“反转”流通池表面上的靶核酸;(e)使重新合成的链线性化;以及(f)使第二测序引物杂交,并如上所述进行延伸、扫描和解封闭的重复循环。反转操作可通过将如上所述的试剂递送用于桥式扩增的单个循环来进行。

[0600] 术语“参考基因组”或“参考序列”是指可用于参考来自受试者的已鉴定序列的任何生物体的任何特定已知基因组序列,无论是部分的还是完整的。例如,可在ncbi.nlm.nih.gov的美国国家生物技术信息中心(National Center for Biotechnology Information)找到用于人类受试者以及许多其他生物体的参考基因组。“基因组”是指以核酸序列表达的生物体或病毒的完整遗传信息。基因组既包括基因又包括DNA的非编码序列。参考序列可大于与其比对的读段。例如,参考序列可为比对读段的至少约100倍大、或至少约1000倍大、或至少约10,000倍大、或至少约105倍大、或至少约106倍大、或至少约107倍大。在一个示例中,参考基因组序列是全长人类基因组的序列。在另一个示例中,参考基因组序列限于特定的人类染色体,诸如13号染色体。在一些具体实施中,参考染色体是来自人

类基因组版本hg19的染色体序列。此类序列可称为染色体参考序列,但术语参考基因组旨在涵盖此类序列。参考序列的其他示例包括其他物种的基因组,以及任何物种的染色体、亚染色体区域(诸如链)等。在各种具体实施中,参考基因组是来源于多个个体的共有序列或其他组合。然而,在某些应用中,参考序列可取自特定个体。在其他具体实施中,“基因组”还涵盖所谓的“图形基因组”,其使用基因组序列的特定存储格式和表示。在一个具体实施中,图形基因组将数据存储在在线性文件中。在另一个具体实施中,图形基因组是指其中另选序列(例如,具有小差异的染色体的不同拷贝)作为不同路径存储在图中的表示。关于图形基因组具体实施的其他详细信息可见于<https://www.biorxiv.org/content/biorxiv/early/2018/03/20/194530.full.pdf>,其内容据此全文以引用方式并入本文。

[0601] 术语“读段”是指描述核苷酸样本或参考的片段的序列数据的集合。术语“读段”可指样本读段和/或参考读段。通常,尽管不是必须的,读段表示样本或参考中邻接碱基对的短序列。读段可由样本或参考片段的碱基对序列(ACTG形式)象征性地表示。读段可存储在存储器设备中并视情况进行处理,以确定该读段是匹配参考序列还是满足其他标准。读段可直接从测序装置获得或间接从所存储的关于样本的序列信息获得。在一些情况下,读段为足够长度(例如,至少约25bp)的DNA序列,其可用于鉴定更大的序列或区域,例如,其可被比对并特异性地分配给染色体或基因组区域或基因。

[0602] 下一代测序方法包括例如边合成边测序技术(Illumina)、焦磷酸测序(454)、离子半导体技术(Ion Torrent测序)、单分子实时测序和边连接边测序(SOLiD测序)。根据测序方法,每个读段的长度可在约30bp至超过10,000bp之间变化。例如,使用SOLiD测序仪的DNA测序方法产生约50bp的核酸读段。又如,Ion Torrent测序产生最多400bp的核酸读段并且454焦磷酸测序产生约700bp的核酸读段。又如,单分子实时测序方法可产生10,000bp至15,000bp的读段。因此,在某些具体实施中,核酸序列读段的长度为30bp-100bp、50bp-200bp或50bp-400bp。

[0603] 术语“样本读段”、“样本序列”或“样本片段”是指来自样本的感兴趣的基因组序列的序列数据。例如,样本读段包含来自具有正向和反向引物序列的PCR扩增子的序列数据。序列数据可从任何选择的序列方法中获得。样本读段可例如来自边合成边测序(SBS)反应、边连接边测序反应或需要确定重复元件的长度和/或同一性的任何其他合适的测序方法。样本读段可以是来源于多个样本读段的共有(例如,平均或加权)序列。在某些具体实施中,提供参考序列包括基于PCR扩增子的引物序列鉴定感兴趣的基因位点。

[0604] 术语“原始片段”是指感兴趣的基因组序列的一部分的序列数据,该序列数据与样本读段或样本片段中感兴趣的指定位置或次要位置至少部分地重叠。原始片段的非限制性示例包括双重拼接片段、单重拼接片段、双重未拼接片段和单重未拼接片段。术语“原始”用于表示原始片段包括与样本读段中的序列数据具有一些关系的序列数据,而不管原始片段是否表现出对应于并验证或确认样本读段中的潜在变异的支持变异。术语“原始片段”不表示该片段必然包括在样本读段中验证变异调用的支持变异。例如,当变异调用应用程序确定样本读段表现出第一变异时,变异调用应用程序可确定一个或多个原始片段缺少对应类型的“支持”变异,原本在给定样本读段中的变异的情况下可能预计会出现该“支持”变异。

[0605] 术语“映射”、“比对(aligned、alignment或aligning)”是指将读段或标签与参考序列进行比较,从而确定参考序列是否包含该读段序列的过程。如果参考序列包含该读段,

则该读段可映射到参考序列,或者在某些具体实施中,映射到参考序列中的特定位置。在一些情况下,比对简单地告知读段是否为特定参考序列的成员(即,该读段是否存在于该参考序列中)。例如,读段与人类13号染色体的参考序列的比对将告知该读段是否存在于13号染色体的参考序列中。提供该信息的工具可被称为集合成员资格测试仪(set membership tester)。在一些情况下,比对另外指示读段或标签映射到的参考序列中的位置。例如,如果参考序列是人类全基因组序列,则比对可指示读段存在于13号染色体上,并且还可指示该读段存在于13号染色体的特定链和/或位点上。

[0606] 术语“插入缺失”指生物体DNA中碱基的插入和/或缺失。微插入缺失表示导致1至50个核苷酸的净变化的插入缺失。在基因组的编码区中,除非插入缺失的长度是3的倍数,否则会产生移码突变。插入缺失可与点突变形成对比。插入缺失插入片段从序列中插入和缺失核苷酸,而点突变是置换其中一个核苷酸而不改变DNA中的核苷酸总数的置换形式。插入缺失也可与串联碱基突变(TBM)形成对比,该串联碱基突变可被定义为相邻核苷酸处的置换(主要是两个相邻核苷酸处的置换,但已观察到三个相邻核苷酸处的置换)。

[0607] 术语“变异”是指与核酸参照物不同的核酸序列。典型的核酸序列变异包括但不限于单核苷酸多态性(SNP)、短缺失和插入多态性(Indel)、拷贝数变异(CNV)、微卫星标记或短串联重复序列和结构变异。体细胞变异调用是识别以低频率存在于DNA样本中的变异的工作。体细胞变异调用在癌症治疗的背景下是令人感兴趣的。癌症是由DNA中突变的积聚引起的。来自肿瘤的DNA样本通常是异质的,包括一些正常细胞、癌症进展早期的一些细胞(具有较少突变)和一些晚期细胞(具有较多突变)。由于这种异质性,当对肿瘤(例如,来自FFPE样本)测序时,体细胞突变将通常以低频率出现。例如,可在覆盖给定碱基的读段的仅10%中看到SNV。待由变异分类器分类为体细胞或种系的变异在本文中也称为“待测变异”。

[0608] 术语“噪声”是指由测序过程和/或变异调用应用程序中的一个或多个错误引起的错误变异调用。

[0609] 术语“变异频率”表示等位基因(基因变异)在群体中特定基因位点处的相对频率,以分数或百分比表示。例如,分数或百分比可以是群体中所有携带该等位基因的染色体的分数。以举例的方式,样本变异频率表示等位基因/变异在沿感兴趣的基因组序列的特定基因位点/位置处相对于“群体”的相对频率,群体对应于从个体获得的感兴趣的基因组序列的读段和/或样本的数量。又如,基线变异频率表示等位基因/变异在沿一个或多个基线基因组序列的特定基因位点/位置处的相对频率,其中“群体”对应于从正常个体的群体获得的一个或多个基线基因组序列的读段和/或样本的数量。

[0610] 术语“变异等位基因频率(VAF)”是指观察到的匹配变异的测序读段的百分比除以靶位置处的总体覆盖率。VAF是携带变异的测序读段的比例的量度。

[0611] 术语“位置”、“指定位置”和“基因位点”是指核苷酸序列内的一个或多个核苷酸的位置或坐标。术语“位置”、“指定位置”和“基因位点”也指核苷酸序列中的一个或多个碱基对的位置或坐标。

[0612] 术语“单倍型”指染色体上相邻位点处一起遗传的等位基因的组合。单倍型可为一个基因位点、多个基因位点或整个染色体,具体取决于在给定的一组基因位点之间发生的重组事件的数目(如果有的话)。

[0613] 术语“阈值”在本文中是指用作表征样本、核酸或其部分(例如,读段)的截止值的

数字或非数字值。阈值可基于经验分析而改变。可将阈值与测量值或计算值进行比较,以确定是否应以特定方式对产生此类值的源进行分类。阈值可根据经验或分析来识别。阈值的选择取决于用户希望必须进行分类的置信水平。阈值可被选择用于特定目的(例如,以平衡灵敏度和选择性)。如本文所用,术语“阈值”指示可改变分析过程的点和/或可触发动作的点。阈值不需要是预定数量。相反,阈值可以是例如基于多个因素的函数。阈值可根据情况进行调整。此外,阈值可指示上限、下限或限值之间的范围。

[0614] 在一些具体实施中,可将基于测序数据的度量或分数与阈值进行比较。如本文所用,术语“度量”或“分数”可包括由测序数据确定的值或结果,或者可包括基于由测序数据确定的值或结果的函数。与阈值一样,度量或分数可根据情况进行调整。例如,度量或分数可以是归一化值。作为分数或度量的示例,一个或多个具体实施可在分析数据时使用计数分数。计数分数可基于样本读段的数量。样本读段可能已经历一个或多个过滤阶段,使得样本读段具有至少一个共同的特征或质量。例如,用于确定计数分数的每个样本读段可能已经与参考序列比对,或者可能被分配为潜在等位基因。可对具有共同特征的样本读段的数量进行计数以确定读段计数。计数分数可基于读段计数。在一些具体实施中,计数分数可以是等于读段计数的值。在其他具体实施中,计数分数可基于读段计数和其他信息。例如,计数分数可基于基因位点的特定等位基因的读段计数和基因位点的读段总数。在一些具体实施中,计数分数可基于基因位点的读段计数和先前获得的数据。在一些具体实施中,计数分数可以是预定值之间的归一化分数。计数分数还可以是来自样本的其他基因位点的读段计数的函数或来自与感兴趣的样本并行运行的其他样本的读段计数的函数。例如,计数分数可以是特定等位基因的读段计数和样本中其他基因位点的读段计数和/或来自其他样本的读段计数的函数。例如,来自其他基因位点的读段计数和/或来自其他样本的读段计数可用于归一化特定等位基因的计数分数。

[0615] 术语“覆盖率”或“片段覆盖率”是指序列的相同片段的多个样本读段的计数或其他量度。读段计数可表示覆盖对应片段的读段数量的计数。另选地,可通过将读段计数乘以基于历史知识、样本知识、基因位点知识等的指定因子来确定覆盖率。

[0616] 术语“读段深度”(通常为数字后跟“×”)是指在靶位置处具有重叠比对的序列读段的数量。这通常表示为一组区间(诸如外显子、基因或组)内的平均值或超过截止值的百分比。例如,临床报告可能说组平均覆盖率为1,105×,其中98%的靶碱基覆盖率>100×。

[0617] 术语“碱基调用质量分数”或“Q分数”是指与单个测序碱基正确的概率成反比的0-50范围内的PHRED标度概率。例如,Q为20的T碱基调用被认为可能是正确的,概率为99.99%。Q<20的任何碱基调用应被认为是低质量的,并且在支持变异的相当大比例的测序读段具有低质量的情况下识别的任何变异应被认为是潜在假阳性的。

[0618] 术语“变异读段”或“变异读段数量”是指支持变异存在的测序读段的数量。

[0619] 关于“链型”(或DNA链型),DNA中的遗传信息可表示为字母A、G、C和T的字符串。例如,5'-AGGACA-3'。通常,序列沿此处所示的方向书写,即,5'端向左而3'端向右。DNA有时可以单链分子的形式出现(如在某些病毒中),但通常我们发现DNA为双链单元。它具有双螺旋结构,该结构具有两条逆平行链。在这种情况下,词语“逆平行”是指两条链平行延伸,但具有相反的极性。双链DNA通过碱基之间的配对保持在一起,并且配对总是使得腺嘌呤(A)与胸腺嘧啶(T)配对并且胞嘧啶(C)与鸟嘌呤(G)配对。这种配对被称为互补性,并且DNA的一

条链被称为另一条链的互补序列。因此,双链DNA可表示为两个字符串,像这样:5'-AGGACA-3'和3'-TCCTGT-5'。注意,两条链具有相反的极性。因此,两条DNA链的链型可称为参考链及其互补链、正向链和反向链、顶部链和底部链、有义链和反义链或沃森链和克里克链。

[0620] 读段比对(也称为读段映射)是找出序列在基因组中的位置的过程。一旦进行了比对,给定读段的“映射质量”或“映射质量分数(MAPQ)”对其在基因组上的位置正确的概率进行定量。映射质量以phred标度进行编码,其中P是比对不正确的概率。概率计算为: $P=10^{(-MAPQ/10)}$,其中MAPQ是映射质量。例如,40的映射质量=10的-4次方,这意味着读段有0.01%的机会被错误地比对。因此,映射质量与若干比对因素相关联,诸如读段的碱基质量、参考基因组的复杂性以及双端信息。关于第一个因素,如果读段的碱基质量低,则意味着观察到的序列可能是错误的,因此其比对是错误的。关于第二个因素,可映射性是指基因组的复杂性。重复区域更难以映射,并且落入这些区域中的读段通常获得低映射质量。在这种情况下,MAPQ反映了这样一个事实,即读段不是唯一比对的,它们的真正来源无法确定。关于第三个因素,在双端测序数据的情况下,一致的对更可能是良好比对的。映射质量越高,比对越好。以良好映射质量比对的读段通常意味着读段序列良好,并且在高可映射性区域中几乎没有错配地进行比对。MAPQ值可用作比对结果的质量控制。MAPQ高于20的所比对读段的比例通常用于下游分析。

[0621] 如本文所用,“信号”是指可检测的事件,诸如图像中的发射,优选地为光发射。因此,在优选的具体实施中,信号可表示在图像中捕获的任何可检测的光发射(即,“光点”)。因此,如本文所用,“信号”可指来自标本的分析物的实际发射,并且可指与实际分析物不相关的杂散发射。因此,信号可能由噪声产生,并且由于不代表标本的实际分析物而可能随后被丢弃。

[0622] 如本文所用,术语“丛”是指一组信号。在特定具体实施中,信号来源于不同的分析物。在优选的具体实施中,信号丛是聚集在一起的一组信号。在更优选的具体实施中,信号丛表示由一个扩增的寡核苷酸覆盖的物理区域。每个信号丛在理想情况下应被观察为若干信号(每个模板循环一个,并且可能由于串扰而更多)。因此,在来自同一信号丛的模板中包括两个(或更多个)信号的情况下,检测到重复信号。

[0623] 如本文所用,术语诸如“最小”、“最大”、“最小化”、“最大化”及其语法变型可包括不是绝对最大值或最小值的值。在一些具体实施中,这些值包括接近最大值和接近最小值。在其他具体实施中,这些值可包括局部最大值和/或局部最小值。在一些具体实施中,这些值仅包括绝对最大值或最小值。

[0624] 如本文所用,“串扰”是指在一个图像中检测到的信号也在单独的图像中检测到。在优选的具体实施中,当在两个单独的检测通道中检测到发射的信号时,可发生串扰。例如,在发射信号以一种颜色出现的情况下,该信号的发射光谱可与另一种颜色的另一个发射信号重叠。在优选的具体实施中,在单独的通道中检测用于指示核苷酸碱基A、C、G和T的存在的荧光分子。然而,因为A和C的发射光谱重叠,所以可在使用A颜色通道的检测期间检测C颜色信号中的一些。因此,A信号和C信号之间的串扰允许来自一个彩色图像的信号出现在另一个彩色图像中。在一些具体实施中,G和T串扰。在一些具体实施中,通道之间的串扰量为不对称的。应当理解,通道之间的串扰量可通过(除了别的以外)选择具有适当发射光谱的信号分子以及选择检测通道的尺寸和波长范围来控制。

[0625] 如本文所用,“配准(register、registering、registration)”和类似术语是指将来自第一时间点或视角的图像或数据集中的信号与来自另一时间点或视角的图像或数据集中的信号相关联的任何过程。例如,配准可用于对准来自图像集的信号以形成模板。又如,配准可用于将来自其他图像的信号与模板对准。一个信号可直接或间接配准到另一个信号。例如,来自图像“S”的信号可直接配准到图像“G”。又如,来自图像“N”的信号可直接配准到图像“G”,或者另选地,来自图像“N”的信号可配准到先前已配准到图像“G”的图像“S”。因此,来自图像“N”的信号间接配准到图像“G”。

[0626] 如本文所用,术语“基准”旨在表示对象中或对象上的可区分的参考点。参考点可为例如标记、第二对象、形状、边缘、区域、不规则性、通道、凹坑、柱等。参考点可存在于对象的图像中或存在于从检测对象导出的另一数据集中。参考点可由对象平面中的x和/或y坐标来指定。另选地或除此之外,参考点可由正交于xy平面的z坐标指定,例如由对象和检测器的相对位置限定。可相对于对象或图像或从对象导出的其他数据集的一个或多个其他分析物来指定参考点的一个或多个坐标。

[0627] 如本文所用,术语“光信号”旨在包括例如荧光信号、发光信号、散射信号或吸收信号。可在电磁光谱的紫外(UV)范围(约200纳米至390纳米)、可见(VIS)范围(约391纳米至770纳米)、红外(IR)范围(约0.771微米至25微米)或其他范围内检测光信号。可以排除这些范围中的一个或多个的全部或部分的方式检测光信号。

[0628] 如本文所用,术语“信号电平”旨在表示具有期望或预定义特征的检测到的能量或编码信息的量或数量。例如,光信号可通过强度、波长、能量、频率、功率、亮度等中的一者或多者来量化。其他信号可根据特征诸如电压、电流、电场强度、磁场强度、频率、功率、温度等进行量化。信号缺失被理解为信号电平为零或信号电平与噪声没有明显区别。

[0629] 如本文所用,术语“模拟”旨在表示创建物理事物或动作的表示或模型,其预测该事物或动作的特征。在许多情况下,表示或模型可与事物或动作区分开来。例如,就一个或多个特征诸如颜色、从事物的全部或部分检测到的信号强度、尺寸或形状而言,表示或模型可与事物区分开来。在特定具体实施中,当与事物或动作相比时,表示或模型可以是理想化的、放大的、变暗的或不完整的。因此,在一些具体实施中,例如,就上述特征中的至少一个特征而言,模型的表示可与其所表示的事物或动作区分开来。该表示或模型可以计算机可读格式或介质提供,诸如在本文其他地方阐述的那些中的一者或多者。

[0630] 如本文所用,术语“特定信号”旨在表示相对于其他能量或信息(诸如背景能量或信息)选择性地观察到的检测到的能量或编码信息。例如,特定信号可以是在特定强度、波长或颜色下检测到的光信号;在特定频率、功率或场强下检测到的电信号;或本领域已知的与光谱和分析检测有关的其他信号。

[0631] 如本文所用,术语“长条”旨在表示对象的矩形部分。长条可为细长的条带,其通过在平行于该条带的最长维度的方向上对象与检测器之间的相对运动来扫描。一般来讲,矩形部分或条带的宽度沿其全长将是恒定的。对象的多个长条可彼此平行。对象的多个长条可彼此相邻、彼此重叠、彼此邻接或通过间隙区域彼此隔开。

[0632] 如本文所用,术语“方差”旨在表示预期值与观察值之间的差值或两个或更多个观察值之间的差值。例如,方差可以是期望值与测量值之间的差异。方差可使用统计函数来表示,诸如标准偏差、标准偏差的平方、变异系数等。

[0633] 如本文所用,术语“xy坐标”旨在表示指定xy平面中的位置、尺寸、形状和/或取向的信息。该信息可以是例如笛卡尔系统中的数字坐标。坐标可相对于x轴和y轴中的一者或两者提供,或者可相对于xy平面中的另一个位置提供。例如,对象的分析物的坐标可以指定分析物相对于对象的基准或其它分析物的位置的位置。

[0634] 如本文所用,术语“xy平面”旨在表示由直线轴x和y限定的二维区域。当参考检测器和检测器所观察的对象使用时,该区域可被进一步指定为与检测器和所检测的对象之间的观察方向正交。

[0635] 如本文所用,术语“z坐标”旨在表示指定沿着正交于xy平面的轴的点、线或区域的位置的信息。在特定具体实施中,z轴正交于检测器所观察的对象的区域。例如,光学系统的焦点方向可沿z轴指定。

[0636] 在一些具体实施中,使用仿射变换来变换所采集的信号数据。在一些此类具体实施中,模板生成利用颜色通道之间的仿射变换在运行之间是一致的这一事实。由于这种一致性,当确定标本中分析物的坐标时,可使用一组默认偏移量。例如,默认偏移文件可包含不同通道相对于一个通道(诸如A通道)的相对变换(移动、缩放、偏斜)。然而,在其他具体实施中,颜色通道之间的偏移在运行期间和/或运行之间漂移,使得偏移驱动的模板生成变得困难。在此类具体实施中,本文所提供的方法和系统可利用无偏移模板生成,这将在下文进一步描述。

[0637] 在上述具体实施中的一些具体实施中,该系统可包括流通池。在一些具体实施中,流通池包括区块的槽道或其他配置,其中至少一些区块包括一个或多个分析物阵列。在一些具体实施中,分析物包含多个分子诸如核酸。在某些方面,流通池被配置为将标记的核苷酸碱基递送至核酸阵列,从而延伸与分析物内的核酸杂交的引物,以便产生对应于包含核酸的分析物的信号。在优选的具体实施中,分析物内的核酸彼此相同或基本上相同。

[0638] 在本文所述的一些图像分析系统中,图像集中的每个图像包括颜色信号,其中不同的颜色对应于不同的核苷酸碱基。在一些具体实施中,该图像集中的每个图像包括具有选自至少四种不同颜色的单一颜色的信号。在一些具体实施中,该图像集中的每个图像包括具有选自四种不同颜色的单一颜色的信号。在本文所述的一些系统中,可通过向分子阵列提供四种不同的标记核苷酸碱基来对核酸进行测序,从而产生四种不同的图像,每种图像包含具有单一颜色的信号,其中信号颜色对于四种不同图像中的每一种都是不同的,从而产生对应于核酸中特定位置存在的四种可能核苷酸的四种颜色图像的循环。在某些方面,该系统包括流通池,该流通池被配置为将另外的标记核苷酸碱基递送到分子阵列,从而产生彩色图像的多个循环。

[0639] 在优选的具体实施中,本文提供的方法可包括确定处理器是否正在主动采集数据或处理器是否处于低活动状态。采集和存储大量高质量图像通常需要大量存储容量。另外,一旦采集和存储,对图像数据的分析可能变为资源密集型,并且可能干扰其他功能的处理能力,诸如正在进行的额外图像数据的采集和存储。因此,如本文所用,术语低活动状态是指处理器在给定时间的处理能力。在一些具体实施中,当处理器未采集和/或存储数据时,低活动状态发生。在一些具体实施中,当进行一些数据采集和/或存储时,低活动状态发生,但是额外的处理能力保持不变,使得图像分析可在不干扰其他功能的情况下同时发生。

[0640] 如本文所用,“识别冲突”是指识别多个进程竞争资源的情况。在一些此类具体实

施中,一个进程被赋予比另一个进程更高的优先级。在一些具体实施中,冲突可能涉及为时间分配、处理能力、存储容量或被赋予优先级的任何其他资源赋予优先级的需要。因此,在一些具体实施中,当处理时间或能力要分布在两个进程(诸如分析数据集和采集和/或存储数据集)之间时,两个进程之间存在冲突,并且可通过赋予其中一个进程优先级来解决冲突。

[0641] 本文还提供了用于执行图像分析的系统。该系统可包括处理器;存储容量;以及用于图像分析的程序,该程序包括用于处理用于存储的第一数据集和用于分析的第二数据集的指令,其中该处理包括在存储设备上采集和/或存储第一数据集,并且当处理器未采集第一数据集时分析第二数据集。在某些方面,该程序包括用于以下操作的指令:识别采集和/或存储第一数据集与分析第二数据集之间的冲突的至少一个实例;以及解决冲突以有利于采集和/或存储图像数据,使得采集和/或存储第一数据集被赋予优先级。在某些方面,第一数据集包括从光学成像设备获得的图像文件。在某些方面,系统还包括光学成像设备。在一些具体实施中,光学成像设备包括光源和检测设备。

[0642] 如本文所用,术语“程序”是指执行任务或进程的指令或命令。术语“程序”可与术语模块互换使用。在某些具体实施中,程序可以是在同一命令集下执行的各种指令的编译。在其他具体实施中,程序可指离散的批或文件。

[0643] 下文阐述了利用本文所述的用于执行图像分析的方法和系统的一些令人惊讶的效果。在一些测序具体实施中,测序系统效用的重要量度是其总体效率。例如,每天产生的可映射数据量以及安装和运行仪器的总成本是经济测序解决方案的重要方面。为了减少生成可映射数据的时间并提高系统的效率,可在仪器计算机上启用实时碱基调用,并且可与测序化学过程和成像同步运行。这允许许多数据处理和分析在测序化学过程结束之前完成。另外,它可减少中间数据所需的存储,并限制需要通过网络传输的数据量。

[0644] 虽然序列输出增加了,但从本文提供的系统传输到网络和辅助分析处理硬件的每次运行的数据已显著减少。通过在仪器计算机(采集计算机)上转换数据,显著降低了网络负载。如果没有这些机载的、离线的数据简化技术,一组DNA测序仪器的图像输出将使大多数网络瘫痪。

[0645] 高通量DNA测序仪器的广泛采用部分是由于易于使用、对各种应用的支持以及对几乎任何实验室环境的适用性。本文介绍的高效算法允许将重要的分析功能添加到可控制测序仪器的简单工作站中。对计算硬件要求的这种降低具有若干实际益处,这些益处将随着对测序输出水平的不断提高而变得甚至更重要。例如,通过在简单塔上执行图像分析和碱基调用,将热量产生、实验室占地面积和功耗保持在最低水平。相比之下,其他商业测序技术最近已将其计算基础结构扩展到用于主要分析,处理能力高达五倍以上,从而导致热量输出和功耗的相应增加。因此,在一些具体实施中,本文提供的方法和系统的计算效率使得客户能够增加其测序通量,同时将服务器硬件费用保持在最小。

[0646] 因此,在一些具体实施中,本文提出的方法和/或系统充当状态机,跟踪每个标本的单独状态,并且当其检测到标本准备前进到下一状态时,其进行适当的处理并使标本前进到该状态。根据优选的具体实施,状态机如何监控文件系统以确定标本何时准备好前进到下一状态的更详细示例在下面的实施例1中示出。

[0647] 在优选的具体实施中,本文提供的方法和系统是多线程的,并且可与可配置数量

的线程一起工作。因此,例如在核酸测序的情况下,本文提供的方法和系统能够在实时分析的实时测序运行期间在后台工作,或者其可使用预先存在的图像数据集运行以用于离线分析。在某些优选的具体实施中,方法和系统通过为每个线程赋予其自身所负责的标本的子集来处理多线程。这将线程争用的可能性降至最低。

[0648] 本公开的方法可包括使用检测装置获得对象的目标图像的步骤,其中图像包括对象上分析物的重复图案。能够对表面进行高分辨率成像的检测装置尤其有用。在特定具体实施中,检测装置将具有足够的分辨率,从而以本文所述的密度、间距和/或分析物尺寸来区分分析物。能够从表面获得图像或图像数据的检测装置特别有用。示例性检测器是被配置为在获得区域图像时使对象和检测器保持静态关系的那些检测器。也可使用扫描装置。例如,可使用获得顺序区域图像的装置(例如,所谓的“步进拍摄”检测器)。在对象的表面上连续扫描点或线以积累数据来构建表面图像的设备也是有用的。点扫描检测器可被配置为经由表面的x-y平面中的光栅运动来扫对象表面上的点(即,小的检测区域)。线扫描检测器可被配置为沿着对象表面的y维度扫描线,该线的最长维度沿着x维度出现。应当理解,可移动检测设备、对象或两者以实现扫描检测。特别适用于例如核酸测序应用的检测装置在以下专利中有所描述:公布号2012/0270305A1;2013/0023422A1;和2013/0260372A1;和美国专利号5,528,050;5,719,391;8,158,926和8,241,573,这些专利中的每一篇均以引用方式并入本文。

[0649] 本文所公开的具体实施可被实现为使用编程或工程技术来产生软件、固件、硬件或它们的任何组合的方法、装置、系统或制品。如本文所用,术语“制品”是指在硬件或计算机可读介质诸如光学存储设备和易失性或非易失性存储器设备中实现的代码或逻辑。此类硬件可包括但不限于现场可编程门阵列(FPGA)、粗粒度可重构架构(CGRA)、专用集成电路(ASIC)、复杂可编程逻辑器件(CPLD)、可编程逻辑阵列(PLA)、微处理器或其他类似的处理设备。在特定具体实施中,本文阐述的信息或算法存在于非暂态存储介质中。

[0650] 在特定具体实施中,本文阐述的计算机实现的方法可在获得对象的多个图像时实时发生。此类实时分析尤其可用于核酸测序应用,其中核酸阵列经受流体和检测步骤的重复循环。测序数据的分析通常可能是计算密集型的,使得在进行其他数据采集或分析算法时实时或在后台执行本文所述的方法可能是有益的。可与本发明方法一起使用的示例性实时分析方法是用于可从Illumina公司(San Diego, Calif.)商购获得和/或描述于美国专利申请公布号2012/0020537A1中的MiSeq和HiSeq测序设备的那些方法,该专利申请公布以引用方式并入本文。

[0651] 由一个或多个编程的计算机形成的示例性数据分析系统,其中编程存储在一个或多个机器可读介质上,其中执行代码以进行本文所述方法的一个或多个步骤。在一个具体实施中,例如,该系统包括被设计成允许系统联网到被配置为从目标对象采集数据的一个或多个检测系统(例如,光学成像系统)的接口。接口可在适当的地方接收和调节数据。在特定具体实施中,检测系统将输出数字图像数据,例如,代表单独的图片元素或像素的图像数据,它们共同形成阵列或其他对象的图像。处理器根据由处理代码定义的一个或多个例程来处理接收到的检测数据。处理代码可存储在各种类型的存储器电路中。

[0652] 根据当前设想的具体实施,对检测数据执行的处理代码包括数据分析例程,该例程被设计成分析检测数据以确定在数据中可见或编码的单独分析物的位置和元数据,以及

没有检测到分析物的位置(即,没有分析物的位置,或没有从现有分析物检测到有意义信号的位置)。在特定具体实施中,由于存在附着到成像分析物的荧光染料,阵列中的分析物位置通常将看起来比非分析物位置更亮。应当理解,例如,当探针在分析物处的靶标不存在于被检测的阵列中时,分析物不需要看起来比其周围区域更亮。单独分析物出现的颜色可取决于所采用的染料以及成像系统用于成像目的所用的光的波长。靶标不与之结合或以其他方式缺乏特定标记的分析物可根据其他特征(诸如它们在微阵列中的预期位置)进行鉴定。

[0653] 一旦数据分析例程已在数据中定位单独分析物,就可进行值分配。一般来讲,值分配将基于由对应位置处的检测器组件(例如,像素)表示的数据的特征,将数字值分配给每个分析物。也就是说,例如当处理成像数据时,值分配例程可被设计成识别在特定位置处检测到特定颜色或波长的光,如由该位置处的像素组或簇所指示的。例如,在典型的DNA成像应用中,四种常见核苷酸将由四种独立且可区分的颜色表示。然后,可为每种颜色分配对应于该核苷酸的值。

[0654] 如本文所用,术语“模块”、“系统”或“系统控制器”可包括操作以执行一个或多个功能的硬件和/或软件系统和电路。例如,模块、系统或系统控制器可包括基于存储在有形和非暂态计算机可读存储介质诸如计算机存储器上的指令来执行操作的计算机处理器、控制器或其他基于逻辑的设备。另选地,模块、系统或系统控制器可包括基于硬接线逻辑和电路来执行操作的硬接线设备。附图中所示的模块、系统或系统控制器可表示基于软件或硬接线指令来操作的硬件和电路、指导硬件执行操作的软件或它们的组合。模块、系统或系统控制器可包括或表示硬件电路或电路,该硬件电路或电路包括一个或多个处理器并且/或者与该一个或多个处理器连接,诸如一个或多个计算机微处理器。

[0655] 如本文所用,术语“软件”和“固件”是可互换的,并且包括存储在存储器中以供计算机执行的任何计算机程序,包括RAM存储器、ROM存储器、EPROM存储器、EEPROM存储器和非易失性RAM(NVRAM)存储器。上述存储器类型仅为示例,因此不限制可用于存储计算机程序的存储器类型。

[0656] 在分子生物学领域中,所使用的核酸测序方法中的一种方法是边合成边测序。该技术可应用于大规模并行测序项目。例如,通过使用自动化平台,可以同时进行成千上万的测序反应。因此,本发明的具体实施中的一个具体实施涉及用于采集、存储和分析在核酸测序期间生成的图像数据的仪器和方法。

[0657] 可采集和存储的数据量的巨大增益使得简化的图像分析方法甚至更加有益。例如,本文所述的图像分析方法允许设计者和最终用户有效地利用现有计算机硬件。因此,面对快速增加的数据输出,本文呈现了减少处理数据的计算负担的方法和系统。例如,在DNA测序领域中,在最近一年中产出已按比例缩放了15倍,并且现在可在DNA测序设备的单次运行中达到数百个十亿碱基。如果计算基础设施要求成比例增长,则大多数研究者将仍然无法实现大型基因组级实验。因此,生成更原始的序列数据将增加对二次分析和数据存储的需要,使得对数据传输和存储的优化极其有价值。本文呈现的方法和系统的一些具体实施可减少产生可用序列数据所需的时间、硬件、网络和实验室基础设施要求。

[0658] 本公开描述了用于执行这些方法的各种方法和系统。方法中一些示例被描述为一系列步骤。然而,应当理解,具体实施不限于本文所述的特定步骤和/或步骤顺序。可省略步骤,可修改步骤,并且/或者可添加其他步骤。此外,可组合本文所述的步骤,可同时执行

步骤,可并行执行步骤,可将步骤分成多个子步骤,可以不同顺序执行步骤,或者可以迭代方式重新执行步骤(或一系列步骤)。此外,尽管本文阐述了不同方法,但应当理解,在其他具体实施中可组合这些不同方法(或这些不同方法的步骤)。

[0659] 在一些具体实施中,“被配置为”执行任务或操作的处理单元、处理器、模块或计算系统可被理解为被特别地构造成执行任务或操作(例如,使存储在其上或与其结合使用的一个或多个程序或指令被定制成或旨在执行任务或操作,并且/或者使处理电路的布置被定制成或旨在执行任务或操作)。出于清楚和免生疑问的目的,除非或直到被特别编程或在结构上经修改以执行任务或操作,通用计算机(如果适当编程,则其可被“配置为”执行任务或操作)不被“配置为”执行任务或操作。

[0660] 此外,本文所述的方法的操作可足够复杂,使得在商业上合理的时间段内这些操作无法由普通技术人员或本领域的普通技术人员实际执行。例如,这些方法可依赖于相对复杂的计算,使得人员无法在商业上合理的时间内完成这些方法。

[0661] 在整个本申请中,已引用了各种公布、专利或专利申请。这些公布的公开内容据此全文以引用方式并入本申请中,以便更全面地描述与本发明有关的现有技术。

[0662] 术语“包含”在本文中旨在为开放式的,不仅包括所列举的要素,而且还涵盖任何附加要素。

[0663] 如本文所用,当参考项目的集合使用时,术语“每个”旨在识别集合中的单个项目,但不一定是指集合中的每个项目。如果明确公开或上下文另有明确规定,则可能会出现例外情况。

[0664] 尽管已参考上文提供的示例描述了本发明,但应当理解,在不脱离本发明的前提下,可进行各种修改。

[0665] 本申请中的模块可在硬件或软件中实现,并且不需要按如图所示那样精确地划分成相同的框。这些模块中的一些还可在不同的处理器或计算机上实现,或者在多个不同的处理器或计算机之间扩展。此外,应当理解,在不影响所实现的功能的情况下,可组合、同步操作或以与图中所示不同的序列操作模块中的一些。同样如本文所用,术语“模块”可包括“子模块”,“子模块”本身在本文中可被认为构成模块。图中被分配为模块的框也可被认为是方法中的流程图步骤。

[0666] 如本文所用,信息项的“识别”不一定需要直接指定该信息项。通过简单地利用一个或多个间接层参考实际信息,或者通过识别一起足以确定实际信息项的一个或多个不同信息项,可以在字段中“识别”信息。此外,术语“指定”在本文中用来指与“识别”相同的含义。

[0667] 如本文所用,给定信号、事件或值“依赖于”先导信号、该先导信号的事件或值、受给定信号、事件或值影响的事件或值。如果存在居间处理元件、步骤或时间段,则给定信号、事件或值仍然可以“依赖于”先导信号、事件或值。如果居间处理元件或步骤组合多于一个信号、事件或值,则处理元件或步骤的信号输出被认为“依赖于”信号、事件或值输入中的每一者。如果给定信号、事件或值与先导信号、事件或值相同,则这仅仅是其中给定信号、事件或值仍然被认为“依赖于”或“取决于”或“基于”先导信号、事件或值的简并情况。给定信号、事件或值对另一信号、事件或值的“响应性”以类似方式定义。

[0668] 如本文所用,“并行”或“同步”不需要精确的同时性。如果这些个体中的一个的评

估在这些个体中的另一个的评估完成之前开始,则就足够了。

[0669] 本申请可互换地指“测序图像”、“簇图像”和“簇强度图像”。

[0670] 其它具体实施

[0671] 在一个具体实施中,对于当前测序循环,状态数据可以使用例如到基于神经网络的碱基调用器2900的中间层的残余连接和/或跳过连接被冗余地提供作为到碱基调用器144的输入。在一些具体实施中,除了状态数据被提供作为到基于神经网络的碱基调用器2900的第一层的输入之外,残余连接和/或跳过连接还提供状态数据作为输入。

[0672] 在一个具体实施中,对于当前测序循环,可以从先前测序循环的子集确定状态数据,这与所有先前测序循环相反。

[0673] 在一个具体实施中,对于当前测序循环,可以确定状态数据的多个实例,使得从先前测序循环的不同子集确定每个实例。

[0674] 在一个具体实施中,对于当前测序循环,状态数据的多个副本可以被提供作为到碱基调用器144的输入,例如作为多个输入通道。

[0675] 在一些具体实施中,状态信息可以被提供作为到时间层的输入。

[0676] 性能结果作为创造性和非显而易见性的客观指标

[0677] 图45比较了非基于神经网络的碱基调用器RTA、无状态信息的基于神经网络的碱基调用器DeepRTA(5ci_k14)以及所公开的使用每像素MIN状态信息作为附加/补充输入的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC)的碱基调用性能。在图45中,x轴表示测序运行的测序循环的进展,并且y轴表示碱基调用错误率。而且,图45具有用于第一读取(读段1)的测序和第二读取(读段2)的测序的两个图表。

[0678] 如所展示,带状态的DeepRTA的碱基调用准确性在读段1中表现出优于无状态信息的DeepRTA的10%的改进,并且在读段2中表现出优于无状态信息的DeepRTA的8%的改进。带状态的DeepRTA也显著优于RTA。

[0679] 图46比较了所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC)跨不同状态通道的碱基调用性能。在图46中,不同的状态通道为DC(MIN)_blue、DC(MIN)_green、DC(MIN)_avg、DC(MIN)_min和DC(MIN)。DC(MIN)_blue状态通道为包括仅用于蓝色通道的每像素MIN状态值的单个通道。DC(MIN)_green状态通道为包括仅用于绿色通道的每像素MIN状态值的单个通道。DC(MIN)_avg状态通道为包括被确定为蓝色通道和绿色通道的MIN状态值的平均值的每像素和泛通道AVG状态值的单个通道。DC(MIN)_min状态通道为包括选自蓝色通道和绿色通道的MIN状态值的每像素和泛通道MIN状态值的单个通道。DC(MIN)状态通道为两个通道,一个通道包括用于蓝色通道的每像素和每通道MIN状态值,另一个通道包括用于绿色通道的每像素和每通道MIN状态值。

[0680] 在DC(MIN)状态通道作为补充输入的情况下,带状态的DeepRTA的碱基调用准确性在读段1中表现出优于DeepRTA的10%的改进。在DC(MIN)_blue状态通道作为补充输入的情况下,带状态的DeepRTA的碱基调用准确性在读段1中表现出优于DeepRTA的7%的改进。在DC(MIN)_green状态通道作为补充输入的情况下,带状态的DeepRTA的碱基调用准确性在读段1中表现出优于DeepRTA的4%的改进。在DC(MIN)_avg状态通道作为补充输入的情况下,带状态的DeepRTA的碱基调用准确性在读段1中表现出优于DeepRTA的8%的改进。在DC(MIN)_min状态通道作为补充输入的情况下,带状态的DeepRTA的碱基调用准确性在读段1

中表现出优于DeepRTA的7%的改进。

[0681] 图47示出所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (DC) 对k-mer (特别是5mer) 的碱基调用性能。5mer是指五个碱基位置的重复碱基模式(例如, ACGCG、GGGGG、TCGCG)。在图47中,x轴表示5mer的所有1024种组合,并且y轴表示所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (DC) 的碱基调用性能。

[0682] 具体地,y轴将所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (DC) 的碱基调用错误与无状态信息的DeepRTA的碱基调用错误进行比较,使得所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (DC) 的碱基调用错误越小,沿着y轴的负值越大。如图所示,Gs的高度易错k-mer (由于缺失启用状态(即,仅Gs的暗状态))在所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (DC) 的情况下经历小的碱基调用错误率4702。

[0683] 图48比较了非基于神经网络的碱基调用器RTA、无状态信息的基于神经网络的碱基调用器DeepRTA (5ci_k14)、所公开的使用每像素MIN状态信息作为附加/补充输入的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN) 以及所公开的使用每像素MIN和MAX状态信息作为附加/补充输入的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 的碱基调用性能。

[0684] 图48示出了无状态信息的DeepRTA (5ci_k14) 展示优于RTA的32%的改进,所公开的使用每像素MIN状态信息作为附加/补充输入的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN) 展示优于RTA的39%的改进,并且所公开的使用每像素MIN和MAX状态信息作为附加/补充输入的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 展示优于RTA的53%的改进。

[0685] 图49比较了非基于神经网络的碱基调用器RTA、无状态信息和5循环侧翼输入窗口的基于神经网络的碱基调用器DeepRTA (5ci_k14)、所公开的使用每像素MIN状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN)、所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同3循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (3ci_k14_DC_MIN_MAX) 以及所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 的碱基调用性能。

[0686] 图49示出了所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同3循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (3ci_k14_DC_MIN_MAX) 展示优于RTA的45%的改进。这表明,即使与无状态信息的DeepRTA (5ci_k14) (以及带状态的DeepRTA (5ci_k14_DC_MIN) (都具有5循环侧翼输入窗口) 相比具有较小的输入窗口,带状态的DeepRTA (3ci_k14_DC_MIN_MAX) (3循环侧翼输入窗口) 也表现得更好。即,利用状态信息,通过处理较少的图像数据并由此节省计算,我们得到更好的碱基调用准确性。

[0687] 图50将所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA (5ci_k14_DC_MIN_MAX) 对k-mer (特别是5mer) 的碱基调用性能与无状态信息和5循环侧翼输入窗口的基于神经网络的碱基调用器DeepRTA (5ci_k14) 的碱基调用性能进行比较。5mer是指五个碱基位置

的重复碱基模式(例如,ACGCG、GGGGG、TCGCG)。在图50中,x轴表示5mer的所有1024种组合,并且y轴表示所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(DC)的碱基调用性能。

[0688] 具体地,在图50中,所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX)具有比无状态信息和5循环侧翼输入窗口(5ci_k14)的基于状态和基于神经网络的碱基调用器DeepRTA更长的负线,这指示所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX)对于5mer具沿x轴有更低的碱基调用错误率。

[0689] 图51比较了具有均衡器具体实施的非基于神经网络的碱基调用器RTA、所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX)以及所公开的使用每像素AVG状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_AVG)的碱基调用性能。

[0690] 图51示出了所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX)展示优于具有均衡器具体实施的RTA的36.7%的改进。图51也示出了所公开的使用每像素AVG状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_AVG)展示优于具有均衡器具体实施的RTA的34.6%的改进。

[0691] 图52比较了具有均衡器具体实施的非基于神经网络的碱基调用器RTA以及所公开的使用每像素MIN、MAX和AVG状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX_AVG)的碱基调用性能。

[0692] 图52示出了所公开的使用每像素分钟,MIN、MAX和AVG状态信息作为附加/补充输入连同5循环侧翼输入窗口的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX_AVG)展示优于具有均衡器具体实施的RTA的36.9%的改进。

[0693] 图53比较了具有均衡器具体实施的非基于神经网络的碱基调用器RTA、所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口和大小为14(K=14)的滤波器组的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX)以及所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口和大小为32(K=32)的滤波器组的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX)的碱基调用性能。

[0694] 图53示出了所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口和大小为14(K=14)的滤波器组的带状态的基于状态和基于神经网络的碱基调用器DeepRTA(5ci_k14_DC_MIN_MAX)展示对于读段1优于具有均衡器具体实施的RTA的36.7%的改进,并且对于读段2优于具有均衡器具体实施的RTA的27.2%的改进。图53还示出了所公开的使用每像素MIN和MAX状态信息作为附加/补充输入连同5循环侧翼输入窗口和大小为32(K=32)的滤波器组的带状态的基于状态和基于神经网络的碱基调用器

DeepRTA (5ci_k14_DC_MIN_MAX) 展示对于读段1优于具有均衡器具体实施的RTA的44.8%的改进,并且对于读段2优于具有均衡器具体实施的RTA的36.9%的改进。

[0695] 图54比较了所公开的带状态的基于状态和基于神经网络的碱基调用器DeepRTA针对指数加权平均具体实施的不同 α 参数配置(例如,0.05、0.07、0.10、0.12)的碱基调用性能,如上文所论述。如图所示, α 参数0.12对于读段1和读段2都具有最好的碱基调用准确性。

[0696] 图55和图56是根据所公开的技术的一个具体实施追踪由指数加权平均值确定的状态值的图表。在图55和图56中,蓝线是原始强度,并且使用来自循环1-20的平均强度和 $\alpha=0.02$ 在循环21处初始化指数加权平均估计值。x轴是循环指数,并且y轴是所测量的振幅(蓝色)加上所估计的平均强度(橙色)。所估计的橙色振幅根据需要遵循强度/平均强度的包络。由于需要对开/关强度求平均以恢复平均强度,所以它是适度嘈杂的。在下一示例中,振幅在循环75处突然下降。指数移动平均值能够非常快速地估计该振幅变化。像素强度平均值是非常有噪声的统计值,因为它必须在大的开/关步长以及较小的振幅变化上求平均。

[0697] 图57从次级分析任务和度量的角度比较了具有均衡器具体实施的RTA (RTA+Eq)、无状态信息的DeepRTA (DeepRTA (k14_1m)) 以及所公开的带状态的DeepRTA (DeepRTA_extrachannels (k14_1,_onoff_expavg)),这些次级分析任务和度量如调用读数的数量、读段1和读段1以及其平均值的读取错配错误率、单核苷酸多态性 (SNP) 重新调用、SNO精度、SNP调用准确性 (F1分数)、插入/缺失 (Indel) 重新调用、Indel精度和Indel调用准确性 (F1分数)。所公开的DeepRTA_extrachannels (k14_1,_onoff_expavg) 模型使用从上文所论述的指数加权平均值具体实施确定的状态信息。如图所示,所公开的DeepRTA_extrachannels (k14_1,_onoff_expavg) 模型在多种次级分析任务和度量方面优于其它模型。

[0698] 条款

[0699] 本发明所公开的技术可以作为系统、方法或制品来实践。具体实施的一个或多个特征可与基本具体实施组合。不互相排斥的具体实施被教导为可组合的。具体实施的一个或多个特征可与其他具体实施组合。本公开周期性地提醒用户这些选项。从一些具体实施中省略重复这些选项的表述不应被视为限制前述部分中教导的组合,这些表述将据此以引用方式并入以下具体实施中的每个具体实施中。

[0700] 可以以计算机产品的形式实现所公开的技术或其元素的一个或多个具体实施和条款,该计算机产品包括具有用于执行所指示的方法步骤的计算机可用程序代码的非暂态计算机可读存储介质。此外,所公开的技术或其元素的一个或多个具体实施和条款可以包括存储器的装置和耦合到存储器的至少一个处理器的形式来实现,并且操作以执行示例性方法步骤。又进一步地,在另一个方面中,可以用于执行本文所述的方法步骤中的一个或多个方法步骤的装置的形式来实现所公开的技术或其元素的一个或多个具体实施和条款;该装置可包括 (i) 硬件模块, (ii) 在一个或多个硬件处理器上执行的软件模块,或 (iii) 硬件和软件模块的组合; (i) 至 (iii) 中的任何一个实施本文所阐述的特定技术,并且软件模块被存储在计算机可读存储介质(或多个此类介质)中。

[0701] 本章节所述的条款可以组合为特征。为了简洁起见,特征的组合不是单独列举的,并且不与每个基本特征组一起重复。读者应当理解在本章节中所述的条款中标识的特征可如何容易地与在本申请的其它章节中被标识为具体实施的基本特征集合组合。这些条款并不意味着相互排斥、穷尽或限制;所公开的技术不限于这些条款,而是包括在所要求保护的

技术和其等同物的范围内的所有可能的组合、修改和变化。

[0702] 在本章节中所述的条款的其它具体实施可以包括存储指令的非暂态计算机可读存储介质,这些指令可由处理器执行以执行上述条款中的任一项条款。在本章节中所述的条款的又一个具体实施可以包括一种系统,该系统包括存储器和一个或多个处理器,该一个或多个处理器可操作以执行存储在存储器中的指令,以执行上述条款中的任一项条款。

[0703] 本发明公开了以下条款:

[0704] 条款集1-每像素状态通道

[0705] 1.一种系统,包括:

[0706] 存储器,该存储器针对测序运行的当前测序循环存储多个像素中的相应像素的相应当前强度值;

[0707] 存储器,该存储器针对测序运行的在当前测序循环之前的一个或多个先前测序循环存储相应像素的相应先前强度值的序列;

[0708] 状态发生器,该状态发生器具有对存储器的访问权限,被配置为根据以下各者生成相应像素的相应当前状态值:

[0709] (i) 相应当前强度值,以及

[0710] (ii) 相应先前强度值的序列中的相应先前强度值,

[0711] 并且被配置为将相应当前状态值存储在存储器中;以及碱基调用器,该碱基调用器具有对存储器的访问权限,被配置为响应于处理以下各者而生成针对当前测序循环的碱基调用:

[0712] (i) 相应当前强度值,以及

[0713] (ii) 相应当前状态值。

[0714] 2.根据条款1所述的系统,其中相应像素的相应强度值各自由多个通道的通道特异性强度值表征。

[0715] 3.根据条款2所述的系统,其中相应像素的相应状态值各自由多个通道的通道特异性状态值表征。

[0716] 4.根据条款3所述的系统,其中相应当前状态值是利用相应当前强度值进行逐像素编码的。

[0717] 5.根据条款3所述的系统,其中多个通道中的通道子集的通道特异性状态值是利用相应当前强度值进行逐像素编码的。

[0718] 6.根据条款3所述的系统,其中通道特异性状态值跨多个通道中的通道被平均以生成相应泛通道当前状态值,其中相应泛通道当前状态值是利用相应当前强度值进行逐像素编码的。

[0719] 7.根据条款4所述的系统,其中逐像素编码包括逐像素级联。

[0720] 8.根据条款4所述的系统,其中逐像素编码包括逐像素求和。

[0721] 9.根据条款1所述的系统,其中相应当前状态值为从以下各者在当

[0722] 前测序循环处针对相应像素确定的相应当前平均强度:

[0723] (i) 相应先前强度值,以及

[0724] (ii) 相应当前强度值。

[0725] 10.根据条款9所述的系统,其中相应当前平均强度各自由通道特异性当前平均强

度表征。

[0726] 11. 根据条款10所述的系统,其中相应当前状态值为从以下各者在当

[0727] 前测序循环处针对相应像素确定的相应当前最大强度:

[0728] (i) 相应先前强度值,以及

[0729] (ii) 相应当前强度值。

[0730] 12. 根据条款11所述的系统,其中相应当前最大强度各自由通道特异性当前最大强度表征。

[0731] 13. 根据条款12所述的系统,其中相应当前状态值为从以下各者在当

[0732] 前测序循环处针对相应像素确定的相应当前最小强度:

[0733] (i) 相应先前强度值,以及

[0734] (ii) 相应当前强度值。

[0735] 14. 根据条款13所述的系统,其中相应当前最小强度各自由通道特异性当前最小强度表征。

[0736] 15. 根据条款14所述的系统,其中相应当前状态值为从以下各者在当前测序循环处针对相应像素确定的相应当前指数加权平均强度:

[0737] (i) 相应先前强度值,以及

[0738] (ii) 相应当前强度值。

[0739] 16. 根据条款15所述的系统,其中相应当前指数加权平均强度是基于对最近测序循环比对之前测序循环进行更多加权而确定的。

[0740] 17. 根据条款16所述的系统,其中相应当前指数加权平均强度各自由通道特异性当前指数加权平均强度表征。

[0741] 18. 根据条款17所述的系统,其中相应当前状态值为从以下各者在当前测序循环处针对相应像素确定的相应当前移动平均强度:

[0742] (i) 相应当前强度值,以及

[0743] (ii) 相应先前强度值的滚动子集。

[0744] 19. 根据条款18所述的系统,其中相应当前移动平均强度各自由通道特异性当前移动平均强度表征。

[0745] 20. 根据条款19所述的系统,其中相应当前强度值中的每个相应当前强度值基于相应当前强度值与相应当前启用状态强度和相应当前非启用状态强度的比较而归属于启用状态桶或非启用状态桶。

[0746] 21. 根据条款20所述的系统,其中相应当前启用状态强度为从以下各者在当前测序循环处针对相应像素确定的相应当前全局最大强度:

[0747] (i) 相应先前强度值。

[0748] 22. 根据条款21所述的系统,其中相应当前全局最大强度各自由通道特异性当前全局最大强度表征。

[0749] 23. 根据条款22所述的系统,其中相应当前非启用状态强度为从以下各者在当前测序循环处针对相应像素确定的相应当前全局最小强度:

[0750] (i) 相应先前强度值。

[0751] 24. 根据条款23所述的系统,其中相应当前全局最小强度各自由通道特异性当前

全局最小强度表征。

[0752] 25. 根据条款24所述的系统,其中相应当前状态值还包括在当前测序循环处针对相应像素生成的相应当前启用状态值和相应当前非启用状态值。

[0753] 26. 根据条款25所述的系统,其中相应当前启用状态值各自由通道特异性当前启用状态值表征。

[0754] 27. 根据条款26所述的系统,其中相应当前非启用状态值各自由通道特异性当前非启用状态值表征。

[0755] 28. 根据条款27所述的系统,其中目标通道中的像素的当前启用状态值是在当前测序循环处从以下各者确定的:

[0756] (i) 目标通道中的像素的在当前测序循环处被检测到并且归属于启用状态桶的当前强度值,以及

[0757] (ii) 目标通道中的像素的在先前测序循环处被检测到的先前强度值。

[0758] 29. 根据条款28所述的系统,其中当前强度值基于当前强度值与从目标通道中的先前强度值确定的全局最大值和全局最小值的比较而归属于启用状态桶。

[0759] 30. 根据条款29所述的系统,其中当前强度值基于当前强度值与在目标通道中的先前测序循环处确定的先前启用状态值和先前非启用状态值的比较而归属于启用状态桶。

[0760] 31. 根据条款30所述的系统,其中当前启用状态值为从当前强度值和先前强度值确定的指数加权平均值。

[0761] 32. 根据条款31所述的系统,其中当前启用状态值为从当前强度值和先前强度值确定的平均值。

[0762] 33. 根据条款32所述的系统,其中当前启用状态值为从当前强度值和先前强度值的滚动子集确定的移动平均值。

[0763] 34. 根据条款33所述的系统,其中当前启用状态值为从当前强度值和先前强度值确定的最小值。

[0764] 35. 根据条款34所述的系统,其中当前启用状态值为从当前强度值和先前强度值确定的最大值。

[0765] 36. 根据条款35所述的系统,其中当前启用状态值是从先前测序循环携带的,并且当当前强度值在当前测序循环处归属于非启用状态桶时不在当前测序循环处重新确定。

[0766] 37. 根据条款36所述的系统,其中目标通道中的像素的当前非启用状态值是在当前测序循环处从以下各者确定的:

[0767] (i) 目标通道中的像素的在当前测序循环处被检测到并且归属于非启用状态桶的当前强度值,以及

[0768] (ii) 目标通道中的像素的在先前测序循环处被检测到的先前强度值。

[0769] 38. 根据条款37所述的系统,其中当前强度值基于当前强度值与从目标通道中的先前强度值确定的全局最大值和全局最小值的比较而归属于非启用状态桶。

[0770] 39. 根据条款38所述的系统,其中当前强度值基于当前强度值与在目标通道中的先前测序循环处确定的先前非启用状态值和先前非启用状态值的比较而归属于非启用状态桶。

[0771] 40. 根据条款39所述的系统,其中当前非启用状态值为从当前强度值和先前强度

值确定的指数加权平均值。

[0772] 41. 根据条款40所述的系统,其中当前非启用状态值为从当前强度值和先前强度值确定的平均值。

[0773] 42. 根据条款41所述的系统,其中当前非启用状态值为从当前强度值和先前强度值的滚动子集确定的移动平均值。

[0774] 43. 根据条款42所述的系统,其中当前非启用状态值为从当前强度值和先前强度值确定的最小值。

[0775] 44. 根据条款43所述的系统,其中当前非启用状态值为从当前强度值和先前强度值确定的最大值。

[0776] 45. 根据条款44所述的系统,其中当前非启用状态值是从先前测序循环携带的,并且当当前强度值在当前测序循环处归属于启用状态桶时不在当前测序循环处重新确定。

[0777] 46. 根据条款1所述的系统,其中在测序运行的每个测序循环处迭代地生成相应当前状态值。

[0778] 47. 根据条款1所述的系统,其中存储器被进一步配置为针对测序运行的下一测序循环存储相应像素的相应下一强度值的序列。

[0779] 48. 根据条款47所述的系统,其中碱基调用器被进一步配置为响应于处理以下各者而针对当前测序循环生成碱基调用:

[0780] (i) 相应当前强度值,

[0781] (ii) 针对先前测序循环中一个或多个先前测序循环的相应先前强度值的序列中的相应先前强度值,

[0782] (iii) 针对下一测序循环中的一个或多个下一测序循环的相应下一强度值的序列中的相应下一强度值,以及

[0783] (iv) 相应状态值。

[0784] 49. 根据条款48所述的系统,其中相应状态值是利用相应先前强度值和相应下一强度值进行逐像素编码的。

[0785] 50. 根据条款3所述的系统,其中通道对应于利用特定激光器照明和通过特定光学滤波器成像的组合。

[0786] 51. 根据条款49所述的系统,其中通道对应于滤波器波长带。

[0787] 52. 根据条款51所述的系统,其中通道对应于测序循环处的成像事件。

[0788] 53. 根据条款1所述的系统,其中碱基调用器为神经网络。

[0789] 54. 根据条款53所述的系统,其中神经网络为卷积神经网络。

[0790] 55. 根据条款54所述的系统,其中卷积神经网络包括多个空间卷积层和多个时间卷积层。

[0791] 56. 根据条款3所述的系统,其中相应状态值被配置为表征相应像素的历史强度模式。

[0792] 57. 根据条款56所述的系统,其中历史强度模式被配置为补偿碱基调用器的碱基调用准确性的损失。

[0793] 58. 根据条款57所述的系统,其中历史强度模式被配置为当针对k-mer的碱基进行碱基调用时补偿碱基调用器的碱基调用准确性的损失。

[0794] 59. 根据条款1所述的系统,其中针对当前测序循环的碱基调用包括针对检测到相应当前信号值和相应先前信号值的一个或多个簇的碱基调用。

[0795] 60. 根据条款59所述的系统,其中相应状态值被配置为辨别簇的相应信号分布。

[0796] 61. 根据条款1所述的系统,其中使用针对已经执行的测序运行离线生成的测序图像来训练碱基调用器。

[0797] 62. 根据条款61所述的系统,其中在由碱基调用器进行碱基调用之前,针对已经执行的测序运行的每个测序循环离线计算相应像素的相应状态值。

[0798] 63. 根据条款62所述的系统,其中碱基调用器被训练为使用相应状态值来补偿碱基调用准确性的损失。

[0799] 64. 一种系统,包括:

[0800] 存储器,该存储器针对测序运行的当前测序循环存储多个信号单元中的相应信号单元的相应当前信号值;

[0801] 该存储器针对测序运行的在当前测序循环之前的一个或多个先前测序循环存储相应信号单元的相应先前信号值的序列;

[0802] 状态发生器,该状态发生器具有对存储器的访问权限,被配置为根据以下各者生成相应信号单元的相应当前状态值:

[0803] (i) 相应当前信号值,以及

[0804] (ii) 相应先前强度值的序列中的相应先前强度值,

[0805] 并且被配置为将相应当前状态值存储在存储器中;以及碱基调用器,该碱基调用器具有对存储器的访问权限,被配置为响应于处理以下各者而生成针对当前测序循环的碱基调用:

[0806] (i) 相应当前信号值,以及

[0807] (ii) 相应当前状态值。

[0808] 65. 根据条款64所述的系统,其中相应信号单元为图像的相应像素。

[0809] 66. 根据条款65所述的系统,其中相应当前信号值为在当前测序循环处检测到的相应像素的相应当前像素强度。

[0810] 67. 根据条款66所述的系统,其中相应先前信号值为在先前测序循环处检测到的相应像素的相应先前像素强度。

[0811] 68. 根据条款67所述的系统,其中针对当前测序循环的碱基调用包括针对检测到相应当前信号值和相应先前信号值的一个或多个簇的碱基调用。

[0812] 69. 根据条款68所述的系统,其中相应当前状态值被配置为辨别簇的相应信号分布。

[0813] 70. 根据条款69所述的系统,其中存储器被进一步配置为针对测序运行的下一测序循环存储相应信号单元的相应下一信号值的序列。

[0814] 71. 根据条款70所述的系统,其中碱基调用器被进一步配置为响应于处理而针对当前测序循环生成碱基调用。

[0815] (i) 相应当前信号值,

[0816] (ii) 针对先前测序循环中一个或多个先前测序循环的相应先前信号值的序列中的相应先前信号值,

[0817] (iii) 针对下一测序循环中的一个或多个下一测序循环的相应下一信号值的序列中的相应下一信号值,以及

[0818] (iv) 相应状态值。

[0819] 72. 根据条款64所述的系统,其中相应当前信号值为在当前测序循环处检测到的相应当前电压值。

[0820] 73. 根据条款72所述的系统,其中相应先前信号值为在先前测序循环处检测到的相应先前电压值。

[0821] 74. 根据条款64所述的系统,其中相应当前信号值为在当前测序循环处检测到的相应当前电流值。

[0822] 75. 根据条款74所述的系统,其中相应先前信号值为在先前测序循环处检测到的相应先前电流值。

[0823] 76. 一种系统,包括:

[0824] 基于神经网络的碱基调用器,该基于神经网络的碱基调用器被配置为处理测序循环的第一滑动窗口,该第一滑动窗口包括用于中心测序循环、一个或多个右侧翼测序循环和一个或多个左侧翼测序循环的测序图像;

[0825] 状态发生器,该状态发生器被配置为基于 (i) 在测序循环的第一滑动窗口中用于中心测序循环的测序图像中的像素强度,以及 (ii) 在测序循环的第一滑动窗口中用于中心测序循环之前的测序循环的测序图像中的像素强度而确定在测序循环的第一滑动窗口中的测序图像的每像素状态;并且

[0826] 该基于神经网络的碱基调用器被进一步配置为处理测序循环的第一滑动窗口中的测序图像连同每像素状态,以在测序循环的第一滑动窗口中对中心测序循环进行碱基调用。

[0827] 77. 根据条款76所述的系统,其中测序循环的第二滑动窗口包括用于中心测序循环、一个或多个右侧翼测序循环和一个或多个左侧翼测序循环的测序图像。

[0828] 其中状态发生器被配置为基于 (i) 在测序循环的第二滑动窗口中用于中心测序循环的测序图像中的像素强度,以及 (ii) 在测序循环的第二滑动窗口中用于中心测序循环之前的测序循环的测序图像中的像素强度而确定在测序循环的第二滑动窗口中的测序图像的每像素状态,并且

[0829] 其中基于神经网络的碱基调用器被进一步配置为处理测序循环的第二滑动窗口中的测序图像连同每像素状态,以在测序循环的第二滑动窗口中对中心测序循环进行碱基调用。

[0830] 78. 根据条款77所述的系统,其中测序循环的第三滑动窗口包括用于中心测序循环、一个或多个右侧翼测序循环和一个或多个左侧翼测序循环的测序图像。

[0831] 其中状态发生器被配置为基于 (i) 在测序循环的第三滑动窗口中用于中心测序循环的测序图像中的像素强度,以及 (ii) 在测序循环的第三滑动窗口中用于中心测序循环之前的测序循环的测序图像中的像素强度而确定在测序循环的第三滑动窗口中的测序图像的每像素状态,并且

[0832] 其中基于神经网络的碱基调用器被进一步配置为处理测序循环的第三滑动窗口中的测序图像连同每像素状态,以在测序循环的第三滑动窗口中对中心测序循环进行碱基

调用。

[0833] 79.一种系统,包括:

[0834] 基于神经网络的碱基调用器,该基于神经网络的碱基调用器具有多个强度模式检测器,该多个强度模式检测器中的强度模式检测器被配置为基于根据测序图像的当前滑动窗口中的像素的强度值的强度模式检测而执行目标碱基调用操作,其中由于强度模式检测器限制于测序图像的当前窗口而导致碱基调用准确性损失;

[0835] 强度上下文化逻辑,该强度上下文化逻辑被配置为基于(i)在测序图像的当前滑动窗口中的测序图像之前的测序图像中的像素的过去强度值以及(ii)在测序图像的当前滑动窗口中的目标测序图像中的像素的目标强度值而确定强度上下文数据;

[0836] 数据流逻辑,该数据流逻辑被配置为向强度上下文数据附加像素的强度值以生成强度上下文文化像素;并且

[0837] 该基于神经网络的碱基调用器被进一步配置为将强度模式检测器应用于强度上下文文化像素,并且生成针对目标碱基调用操作的碱基调用,强度上下文文化像素中的强度上下文数据补偿碱基调用准确性的损失。

[0838] 80.一种计算机实现的方法,所述方法包括:

[0839] 访问用于在测序循环的当前滑动窗口中的第一多个测序循环的当前窗口测序信号数据;

[0840] 访问用于在第一多个测序循环中的测序循环之前的第二多个测序循环的先前测序信号数据;

[0841] 基于当前窗口测序信号数据和先前测序信号数据而生成状态信号数据;以及

[0842] 响应于处理当前窗口测序信号数据和状态信号数据而对第一多个测序循环中的至少一个测序循环进行碱基调用。

[0843] 81.一种计算机实现的方法,所述方法包括:

[0844] 基于从测序运行的在当前测序循环之前的一个或多个先前测序循环确定的状态信息而对测序运行的当前测序循环进行碱基调用。

[0845] 82.一种系统,包括:

[0846] 存储器,该存储器存储用于测序运行的当前测序循环的簇的当前测序图像,以及用于测序运行的在当前测序循环之前的一个或多个先前测序循环的簇的先前测序图像;

[0847] 主机处理器,该主机处理器具有对存储器的访问权限并且被配置为针对当前测序循环基于当前测序图像和先前测序图像而生成测序图像中的像素的当前每像素状态;

[0848] 可配置处理器,所述可配置处理器具有对所述存储器的访问权限并且被配置为执行神经网络以产生碱基调用分类分数;以及

[0849] 数据流逻辑,所述数据流逻辑具有对所述存储器、所述主机处理器和所述可配置处理器的访问权限,并且被配置为

[0850] 响应于处理当前每像素状态和当前测序图像而将当前每像素状态和当前测序图像提供到神经网络并且使神经网络产生针对当前测序循环的簇的当前碱基调用分类分数,以及

[0851] 将当前碱基调用分类分数提供到主机。

[0852] 83.根据条款82所述的系统,其中存储器进一步存储由主机处理器在先前测序循

环中生成的像素的先前每像素状态。

[0853] 84. 根据条款83所述的系统,其中主机处理器被进一步配置为基于当前测序图像和先前每像素状态而生成当前每像素状态。

[0854] 条款集2-每孔状态通道

[0855] 1. 一种系统,包括:

[0856] 空间卷积神经网络,该空间卷积神经网络被配置为处理分析物的测序图像并且产生空间卷积特征;

[0857] 滤波逻辑,该滤波逻辑被配置为从空间卷积特征选择包含分析物的中心的空间卷积特征子集;

[0858] 状态生成逻辑,该状态生成逻辑被配置为生成针对空间卷积特征子集的每分析物状态;

[0859] 时间处理逻辑,该时间处理逻辑被配置为处理空间卷积特征子集和每分析物状态,并且产生时间紧凑输出;以及

[0860] 碱基调用逻辑,该碱基调用逻辑被配置为基于时间紧凑输出而产生碱基调用。

[0861] 2. 根据条款1所述的系统,其中每分析物状态为响应于独立地处理测序图像而通过线性模型在逐分析物的基础上生成的强度变化校正系数。

[0862] 3. 根据条款2所述的系统,其中强度变化校正系数包括补偿分析物的强度分布之间的尺度变化的放大系数。

[0863] 4. 根据条款3所述的系统,其中强度变化校正系数包括补偿沿着多个通道的分析物的强度分布之间的移位变化的通道特异性偏移系数。

[0864] 5. 根据条款4所述的系统,其中强度变化校正系数包括补偿移位变化的公共偏移系数。

[0865] 6. 根据条款5所述的系统,其中基于将在测序运行的当前测序循环处确定的当前强度统计值的分析与在测序运行的在当前测序循环之前的一个或多个先前测序循环处确定的历史强度统计值的分析相结合,在逐分析物的基础上生成强度变化校正系数。

[0866] 7. 根据条款1所述的系统,其中分析物的强度分布的特征在于多个通道中的通道的通道特异性强度值。

[0867] 8. 根据条款7所述的系统,其中每分析物状态的特征在于通道的通道特异性状态值。

[0868] 9. 根据条款8所述的系统,其中每分析物状态是在逐中心特征的基础上利用空间卷积特征子集进行编码的。

[0869] 10. 根据条款8所述的系统,其中用于通道子集的通道特异性状态值是在逐中心特征的基础上利用空间卷积特征子集进行编码的。

[0870] 11. 根据条款8所述的系统,其中通道特异性状态值跨通道被平均以生成泛通道状态值,其中泛通道状态值是在逐中心特征的基础上利用空间卷积特征子集进行编码的。

[0871] 12. 根据条款9所述的系统,其中每分析物状态是在逐中心特征的基础上利用空间卷积特征子集进行级联的。

[0872] 13. 根据条款9所述的系统,其中每分析物状态是在逐中心特征的基础上利用空间卷积特征子集进行求和的。

- [0873] 14. 根据条款1所述的系统,其中每分析物状态基于由空间处理逻辑在当前测序循环处和在先前测序循环处产生的空间卷积特征子集的特征值。
- [0874] 15. 根据条款14所述的系统,其中每分析物状态为特征值的平均值。
- [0875] 16. 根据条款14所述的系统,其中每分析物状态为特征值中的最大特征值。
- [0876] 17. 根据条款14所述的系统,其中每分析物状态为特征值中的最小特征值。
- [0877] 18. 根据条款14所述的系统,其中每分析物状态为特征指数权重值的指数加权平均值。
- [0878] 19. 根据条款18所述的系统,其中指数加权平均值是基于对最近测序循环比对之前测序循环进行更多加权而确定的。
- [0879] 20. 根据条款14所述的系统,其中每分析物状态为特征值的移动平均值。
- [0880] 21. 根据条款20所述的系统,其中移动平均值使用来自先前测序循环的特征值的滚动子集。
- [0881] 22. 根据条款14所述的系统,其中每分析物状态包括启用状态值和非启用状态值。
- [0882] 23. 根据条款22所述的系统,其中启用状态值和非启用状态值为通道特异性的。
- [0883] 24. 根据条款23所述的系统,其中启用状态值和非启用状态值是从先前碱基调用确定的。
- [0884] 25. 根据条款23所述的系统,其中启用状态值和非启用状态值是基于特征值的全局最大值和全局最小值而确定的。
- [0885] 26. 根据条款23所述的系统,其中启用状态值和非启用状态值是基于特征值的指数加权平均值而确定的。
- [0886] 27. 根据条款1所述的系统,其中输入包括针对当前测序循环、一个或多个先前测序循环以及一个或多个下一测序循环的测序图像的滑动窗口。
- [0887] 28. 根据条款27所述的系统,其中特征集合是从针对当前测序循环、先前测序循环和下一侧翼测序循环的测序图像的滑动窗口中的测序图像的像素提取的。
- [0888] 29. 根据条款28所述的系统,其中中心特征集合是从针对当前测序循环、先前测序循环和下一测序循环的特征集合挑选的。
- [0889] 30. 根据条款29所述的系统,其中每分析物状态集合是针对当前测序循环、先前测序循环和下一侧翼测序循环的中心特征集合生成的并利用其进行编码的。
- [0890] 31. 根据条款1所述的系统,其中空间处理逻辑和时间处理逻辑为神经网络。
- [0891] 32. 根据条款31所述的系统,其中神经网络为卷积神经网络。
- [0892] 33. 根据条款1所述的系统,其中每分析物状态被配置为表征像素的历史强度模式。
- [0893] 34. 根据条款33所述的系统,其中历史强度模式被配置为补偿碱基调用准确性的损失。
- [0894] 35. 根据条款34所述的系统,其中历史强度模式被配置为当针对k-mer的碱基进行碱基调用时补偿碱基调用准确性的损失。
- [0895] 36. 根据条款1所述的系统,其中每分析物状态被配置为辨别分析物的强度分布。
- [0896] 37. 根据条款1所述的系统,其中使用针对已经执行的测序运行离线生成的测序图像来训练空间处理逻辑和时间处理逻辑。

[0897] 38.根据条款37所述的系统,其中在碱基调用之前,针对已经执行的测序运行的每个测序循环离线计算每分析物状态。

[0898] 39.根据条款38所述的系统,其中空间处理逻辑和时间处理逻辑被训练以使用每分析物状态来补偿碱基调用准确性的损失。

[0899] 40.一种系统,包括:

[0900] 空间处理逻辑,该空间处理逻辑被配置为处理其中分析物的强度分布跨测序图像的像素分散的输入,并且产生其中分析物的强度分布被聚集到从像素提取的特征中的空间紧凑输出;

[0901] 分析物聚焦逻辑,该分析物聚焦逻辑被配置为从特征挑选表征在分析物的中心处检测到的峰强度的那些每分析物中心特征;

[0902] 状态生成逻辑,该状态生成逻辑被配置为生成针对每分析物中心特征的每分析物状态;

[0903] 时间处理逻辑,该时间处理逻辑被配置为处理每分析物中心特征和每分析物状态,并且产生时间紧凑输出;以及

[0904] 碱基调用逻辑,该碱基调用逻辑被配置为基于时间紧凑输出而产生碱基调用。

[0905] 条款集3-压缩特征状态通道

[0906] 1.一种系统,包括:

[0907] 空间卷积神经网络,该空间卷积神经网络被配置为处理簇的测序图像并且产生空间卷积特征;

[0908] 滤波逻辑,该滤波逻辑被配置为从空间卷积特征选择包含簇的中心的空间卷积特征子集;

[0909] 压缩逻辑,该压缩逻辑被配置为将空间卷积特征子集压缩为压缩特征集合,其中空间卷积特征子集具有M个通道,并且压缩特征集合具有N个通道,并且其中 $M > N$;

[0910] 上下文逻辑,该上下文逻辑被配置为访问压缩特征集合中的压缩特征的状态信息,并且将状态信息与压缩特征相加以生成有状态压缩特征集合;

[0911] 时间卷积神经网络,该时间卷积神经网络被配置为处理有状态压缩特征集合,并且产生时间卷积的有状态特征;以及

[0912] 碱基调用逻辑,该碱基调用逻辑被配置为基于时间卷积的有状态特征而生成针对簇的碱基调用。

[0913] 2.根据条款1所述的系统,其中每通道状态为响应于独立地处理测序图像而通过线性模型在逐分析物的基础上生成的强度变化校正系数。

[0914] 3.根据条款2所述的系统,其中强度变化校正系数包括补偿分析物的强度分布之间的尺度变化的放大系数。

[0915] 4.根据条款3所述的系统,其中强度变化校正系数包括补偿沿着多个通道的分析物的强度分布之间的移位变化的通道特异性偏移系数。

[0916] 5.根据条款4所述的系统,其中强度变化校正系数包括补偿移位变化的公共偏移系数。

[0917] 6.根据条款5所述的系统,其中基于将在测序运行的当前测序循环处确定的当前强度统计值的分析与在测序运行的在当前测序循环之前的一个或多个先前测序循环处确

定的历史强度统计值的分析相结合,在逐分析物的基础上生成强度变化校正系数。

[0918] 7.根据条款1所述的系统,其中分析物的强度分布的特征在于多个通道中的通道的通道特异性强度值。

[0919] 8.根据条款7所述的系统,其中每通道状态的特征在于通道的通道特异性状态值。

[0920] 9.根据条款8所述的系统,其中每通道状态是在逐位的基础上利用每分析物通道特异性位进行编码的。

[0921] 10.根据条款8所述的系统,其中通道的子集的通道特异性状态值是在逐位的基础上利用每分析物通道特异性位进行编码的。

[0922] 11.根据条款8所述的系统,其中通道特异性状态值跨通道被平均以生成泛通道状态值,其中泛通道状态值是在逐位的基础上利用每分析物通道特异性位进行编码的。

[0923] 12.根据条款9所述的系统,其中每通道状态是在逐位的基础上利用每分析物通道特异性位进行级联的。

[0924] 13.根据条款9所述的系统,其中每通道状态是在逐位的基础上利用每分析物通道特异性位进行求和的。

[0925] 14.根据条款1所述的系统,其中每通道状态是基于由空间处理逻辑在当前测序循环处和在先前测序循环处产生的每分析物通道特异性位的压缩值。

[0926] 15.根据条款14所述的系统,其中每通道状态为压缩值的平均值。

[0927] 16.根据条款14所述的系统,其中每通道状态为压缩值中的最大压缩值。

[0928] 17.根据条款14所述的系统,其中每通道状态为压缩值中的最小压缩值。

[0929] 18.根据条款14所述的系统,其中每通道状态为压缩值的指数加权平均值。

[0930] 19.根据条款18所述的系统,其中指数加权平均值是基于对最近测序循环比对之前测序循环进行更多加权而确定的。

[0931] 20.根据条款14所述的系统,其中每通道状态为压缩值的移动平均值。

[0932] 21.根据条款20所述的系统,其中移动平均值使用来自先前测序循环的压缩值的滚动子集。

[0933] 22.根据条款14所述的系统,其中每通道状态包括启用状态值和非启用状态值。

[0934] 23.根据条款22所述的系统,其中启用状态值和非启用状态值为通道特异性的。

[0935] 24.根据条款23所述的系统,其中启用状态值和非启用状态值是从先前碱基调用确定的。

[0936] 25.根据条款23所述的系统,其中启用状态值和非启用状态值是基于压缩值的全局最大值和全局最小值而确定的。

[0937] 26.根据条款23所述的系统,其中启用状态值和非启用状态值是基于压缩值的指数加权平均值而确定的。

[0938] 27.根据条款1所述的系统,其中输入包括针对当前测序循环、一个或多个先前测序循环以及一个或多个下一测序循环的测序图像的滑动窗口。

[0939] 28.根据条款27所述的系统,其中特征集合是从针对当前测序循环、先前测序循环和下一侧翼测序循环的测序图像的滑动窗口中的测序图像的通道特异性像素提取的,其中每分析物中心特征集合是从针对当前测序循环、先前测序循环和下一测序循环的特征集合挑选的。

[0940] 29.根据条款28所述的系统,其中通道特异性位集合是从针对当前测序循环、先前测序循环和下一测序循环的每分析物中心特征集合压缩的。

[0941] 30.根据条款29所述的系统,其中每通道状态集合是针对当前测序循环、先前测序循环和下一侧翼测序循环的通道特异性位集合生成的并利用其进行编码的。

[0942] 31.根据条款1所述的系统,其中空间处理逻辑和时间处理逻辑为神经网络。

[0943] 32.根据条款31所述的系统,其中神经网络为卷积神经网络。

[0944] 33.根据条款1所述的系统,其中每通道状态被配置为表征通道特异性像素的历史强度模式。

[0945] 34.根据条款33所述的系统,其中历史强度模式被配置为补偿碱基调用准确性的损失。

[0946] 35.根据条款34所述的系统,其中历史强度模式被配置为当针对k-mer的碱基进行碱基调用时补偿碱基调用准确性的损失。

[0947] 36.根据条款1所述的系统,其中每分析物状态被配置为辨别分析物的强度分布。

[0948] 37.根据条款1所述的系统,其中使用针对已经执行的测序运行离线生成的测序图像来训练空间处理逻辑和时间处理逻辑。

[0949] 38.根据条款37所述的系统,其中在碱基调用之前,针对已经执行的测序运行的每个测序循环离线计算每通道状态。

[0950] 39.根据条款38所述的系统,其中空间处理逻辑和时间处理逻辑被训练以使用每通道状态来补偿碱基调用准确性的损失。

[0951] 40.一种系统,包括:

[0952] 空间处理逻辑,该空间处理逻辑被配置为处理其中分析物的强度分布跨测序图像的通道特异性像素分散的输入,并且产生其中分析物的强度分布被聚集到从通道特异性像素提取的特征中的空间紧凑输出;

[0953] 分析物聚焦逻辑,该分析物聚焦逻辑被配置为从特征挑选表征在分析物的中心处检测到的峰强度的那些每分析物中心特征;

[0954] 压缩逻辑,该压缩逻辑被配置为将每分析物中心特征提取为每分析物通道特异性位,其中每分析物中心特征具有M个特征通道,其中每分析物通道特异性位具有N个位通道,并且其中 $M > N$;

[0955] 状态生成逻辑,该状态生成逻辑被配置为生成针对每分析物通道特异性位的每通道状态;

[0956] 时间处理逻辑,该时间处理逻辑被配置为处理每分析物通道特异性位和每通道状态,并且产生时间紧凑输出;以及

[0957] 碱基调用逻辑,该碱基调用逻辑被配置为基于时间紧凑输出而产生碱基调用。

[0958] 条款集4-每孔状态到每像素状态的变换

[0959] 1.一种系统,包括:

[0960] 每分析物状态生成逻辑,该每分析物状态生成逻辑被配置为响应于处理分析物的测序图像而生成每分析物状态,并且以稀疏表示对每分析物状态进行编码;

[0961] 每像素状态生成逻辑,该每像素状态生成逻辑被配置为响应于处理稀疏表示而将每分析物状态变换为测序图像中的像素的每像素状态,并且以密集表示对每像素状态进行

编码;以及

[0962] 碱基调用器,该碱基调用器被配置为响应于处理利用以密集表示编码的每像素状态补充的测序图像中的像素而生成碱基调用。

[0963] 2.根据条款1所述的系统,其中每像素状态生成逻辑使用转置卷积。

[0964] 3.根据条款2所述的系统,其中转置卷积被训练以将以稀疏表示编码的每分析物状态变换为以密集表示编码的每像素状态。

[0965] 4.根据条款2所述的系统,其中碱基调用器包括空间卷积层和时间卷积层。

[0966] 5.根据条款4所述的系统,其中以密集表示编码的每像素状态是利用测序图像中的像素进行逐像素编码的,以供碱基调用器的第一空间卷积层处理。

[0967] 6.根据条款5所述的系统,其中以密集表示编码的每像素状态被供应作为到碱基调用器的后续空间卷积层的输入。

[0968] 7.根据条款1所述的系统,其中每分析物状态为响应于处理测序图像而通过线性模型在逐分析物的基础上生成的强度变化校正系数。

[0969] 8.根据条款7所述的系统,其中强度变化校正系数包括补偿分析物的强度分布之间的尺度变化的放大系数。

[0970] 9.根据条款8所述的系统,其中强度变化校正系数包括补偿沿着多个通道的分析物的强度分布之间的移位变化的通道特异性偏移系数。

[0971] 10.根据条款9所述的系统,其中强度变化校正系数包括补偿移位变化的公共偏移系数。

[0972] 11.根据条款10所述的系统,其中基于将在测序运行的当前测序循环处确定的当前强度统计值的分析与在测序运行的在当前测序循环之前的一个或多个先前测序循环处确定的历史强度统计值的分析相结合,在逐分析物的基础上生成强度变化校正系数。

[0973] 12.根据条款1所述的系统,其中分析物的强度分布的特征在于多个通道中的通道的通道特异性强度值。

[0974] 13.根据条款12所述的系统,其中每分析物状态的特征在于通道的通道特异性状态值。

[0975] 14.根据条款1所述的系统,其中每像素状态生成逻辑使用插值。

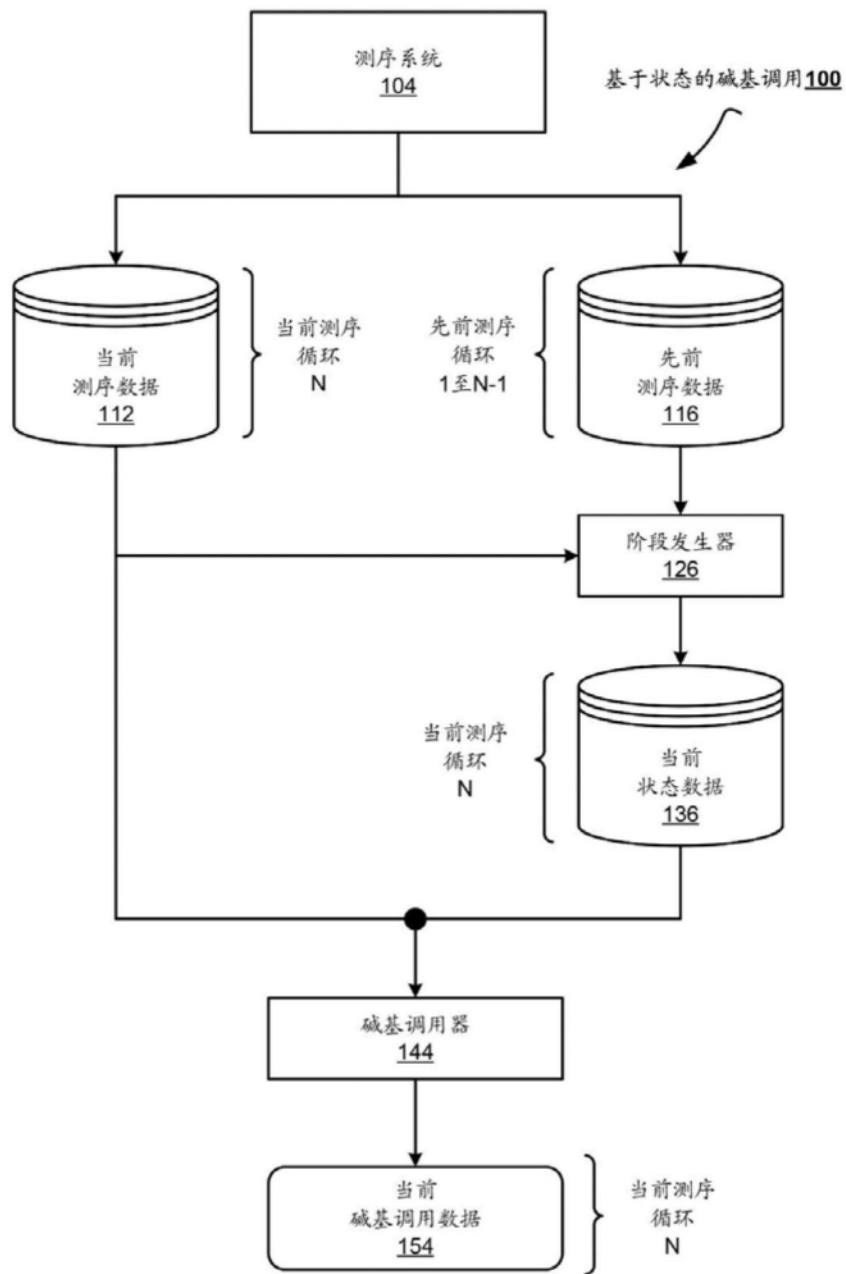


图1

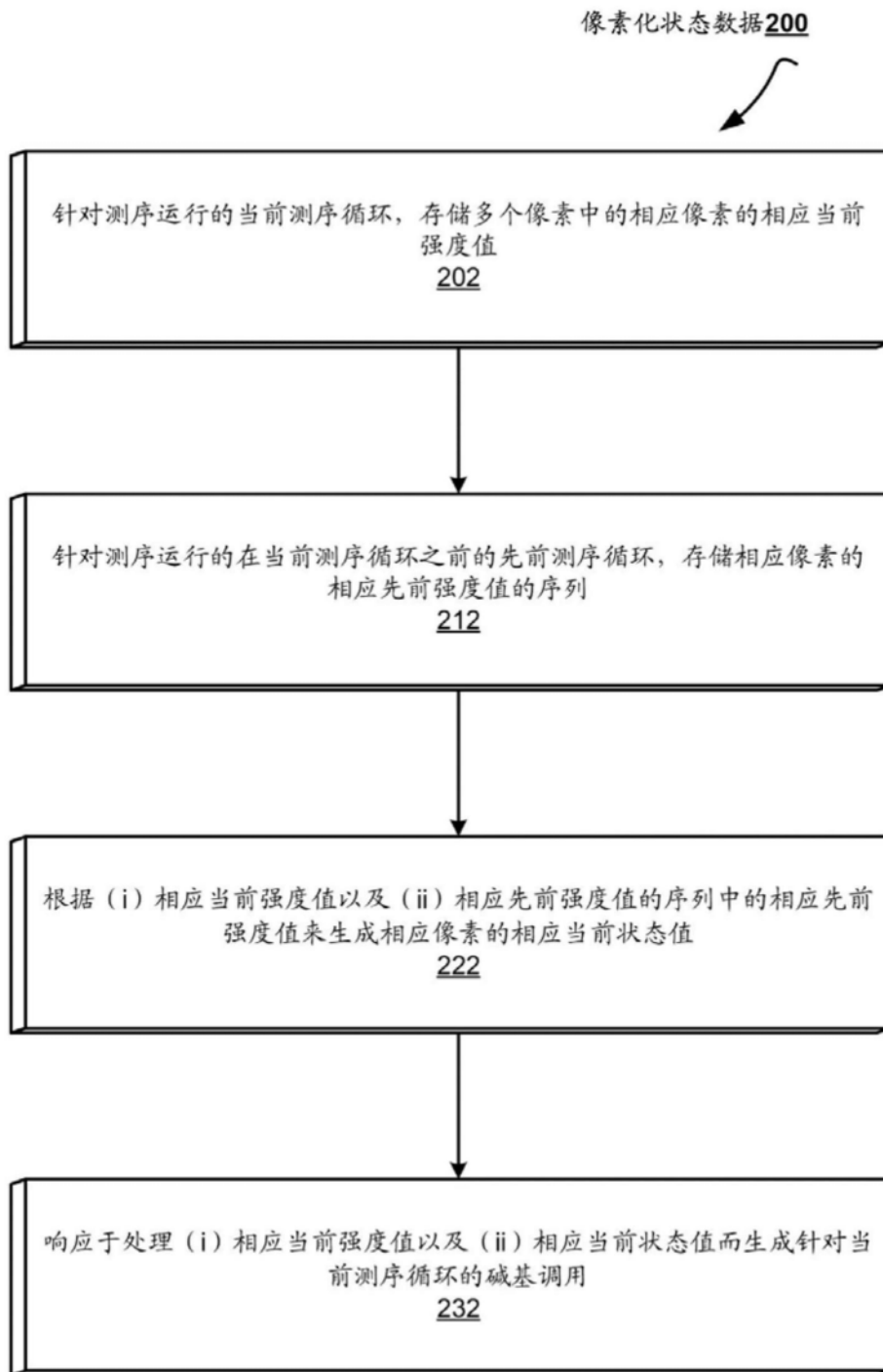


图2

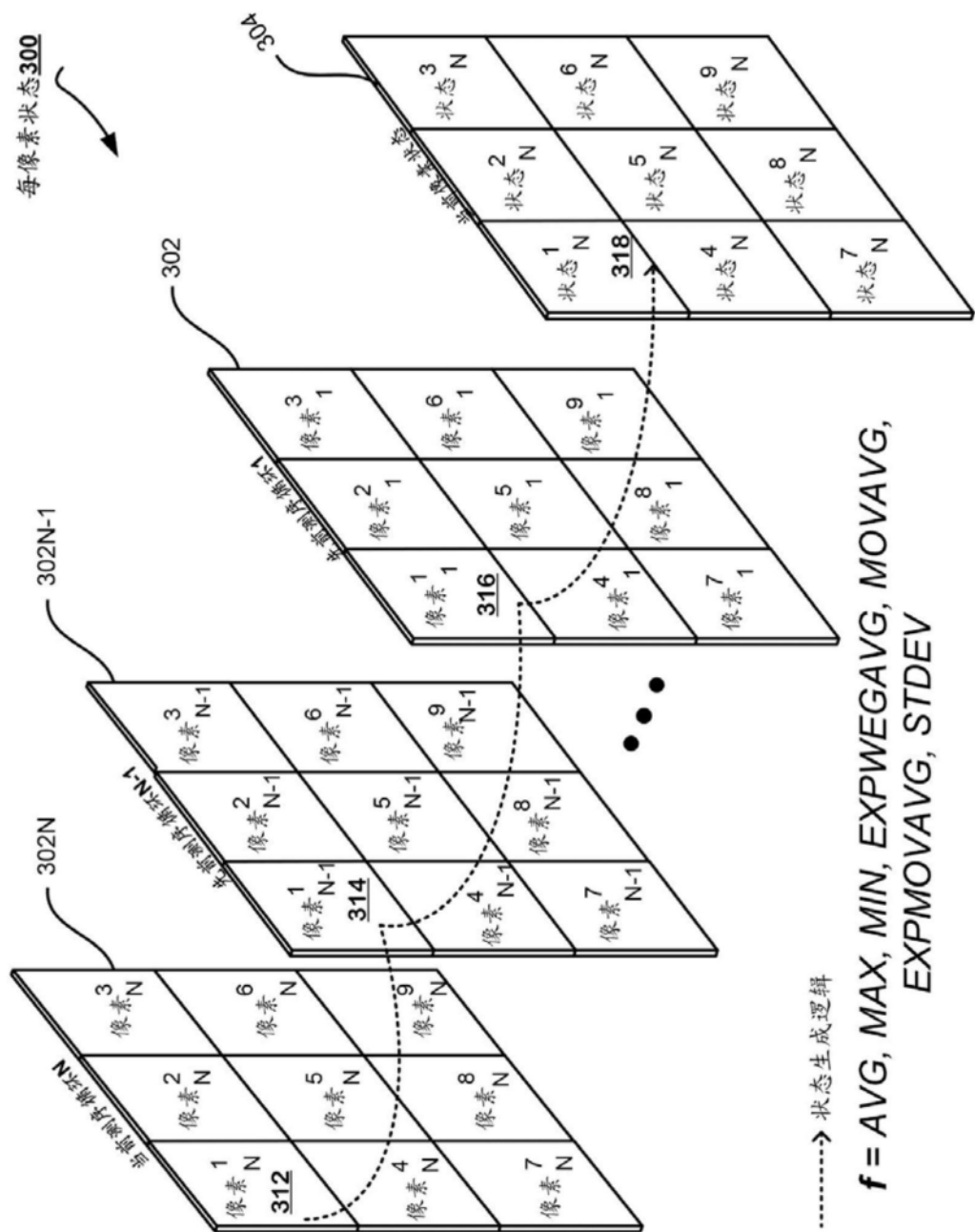


图3

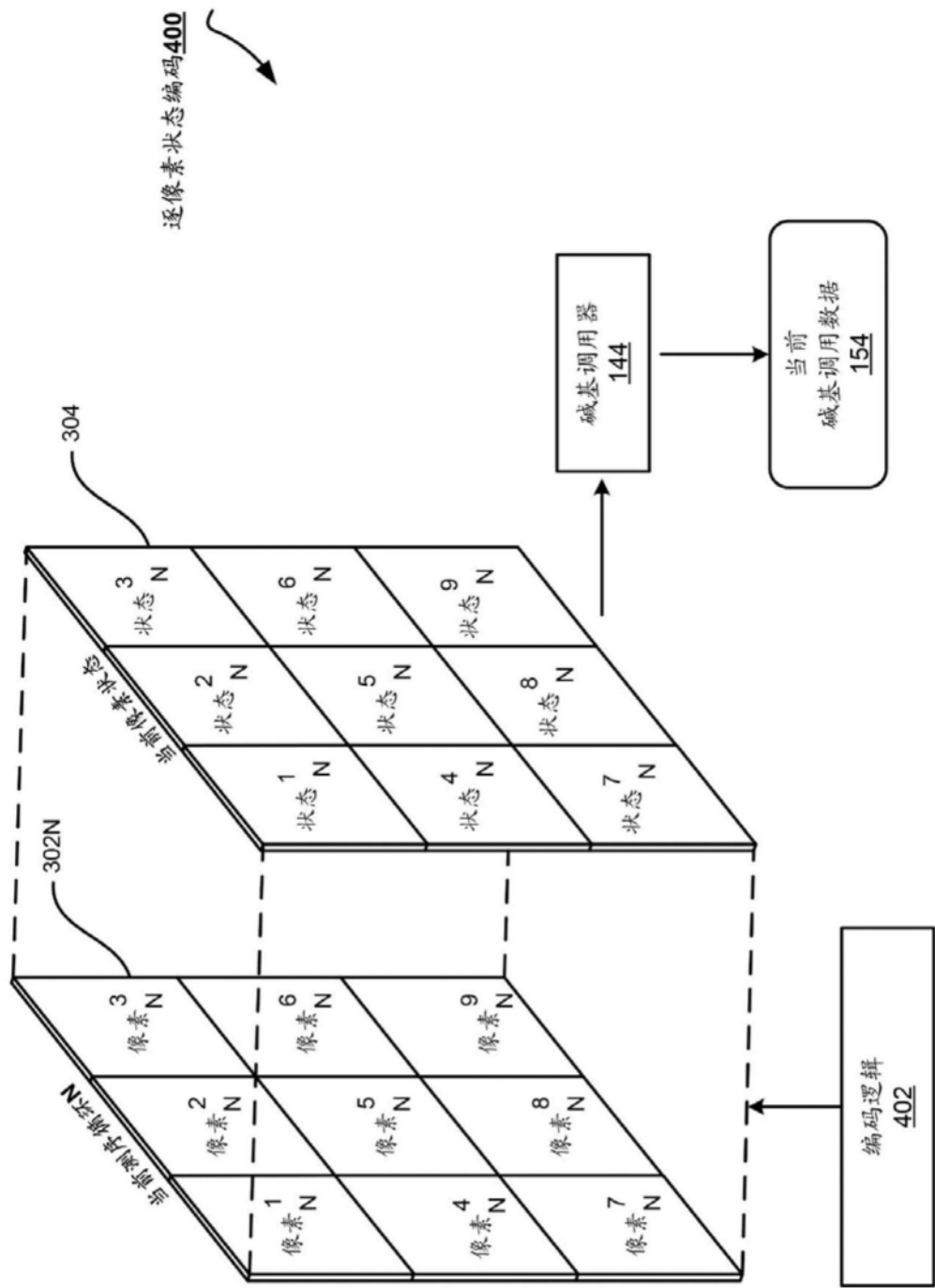


图4

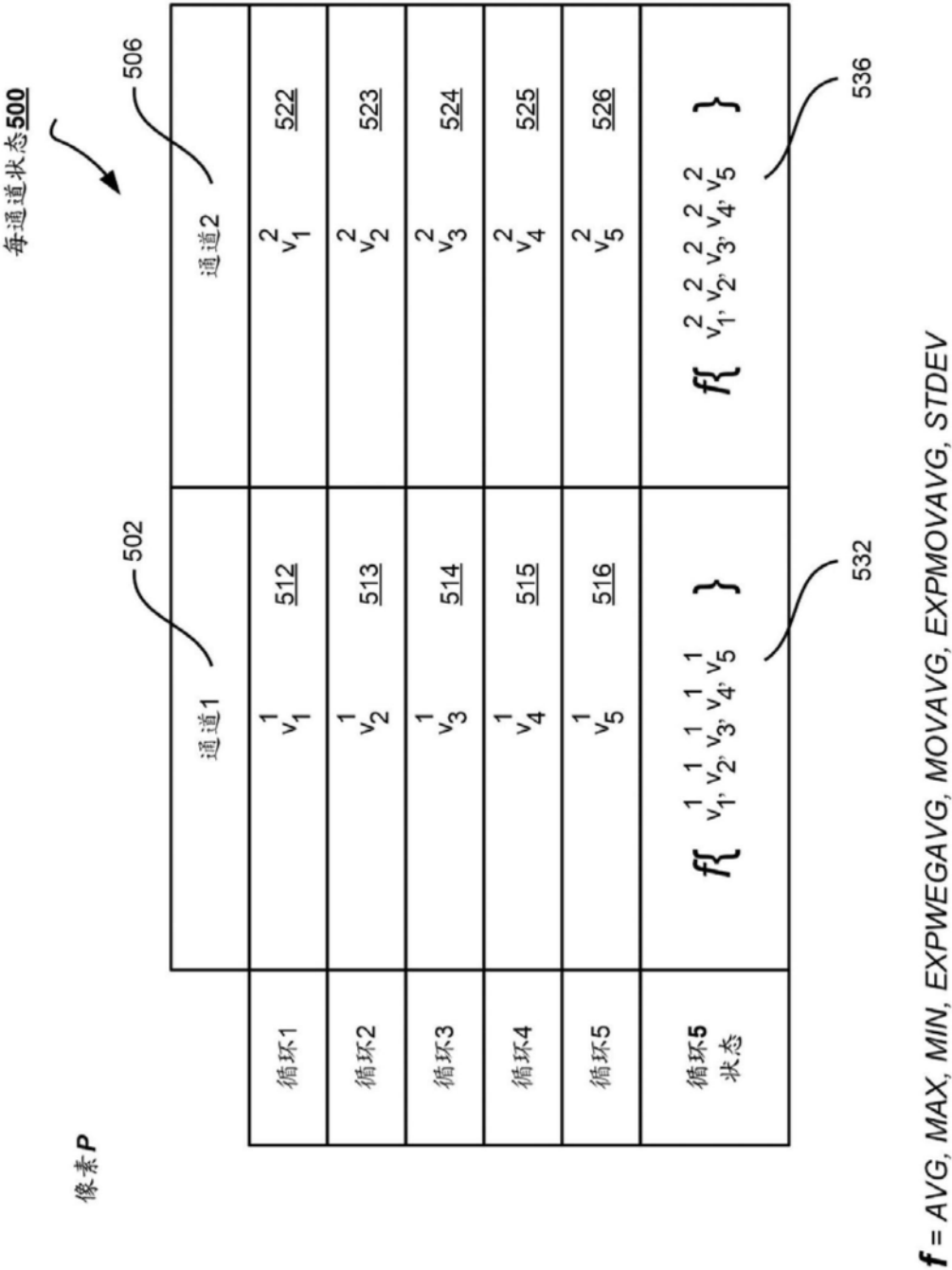


图5

每通道MIN状态600

当前测序循环N

像素601	通道1 强度值602	通道2 强度值603	通道1MIN 状态值604	通道2MIN 状态值605
(1, 1)	x_N^1	y_N^1	$\text{MIN}\{x_N^1 \dots x_1^1\}$	$\text{MIN}\{y_N^1 \dots y_1^1\}$
(1, 2)	x_N^2	y_N^2	$\text{MIN}\{x_N^2 \dots x_1^2\}$	$\text{MIN}\{y_N^2 \dots y_1^2\}$
(1, 3)	x_N^3	y_N^3	$\text{MIN}\{x_N^3 \dots x_1^3\}$	$\text{MIN}\{y_N^3 \dots y_1^3\}$
(1, 4)	x_N^4	y_N^4	$\text{MIN}\{x_N^4 \dots x_1^4\}$	$\text{MIN}\{y_N^4 \dots y_1^4\}$
(1, 5)	x_N^5	y_N^5	$\text{MIN}\{x_N^5 \dots x_1^5\}$	$\text{MIN}\{y_N^5 \dots y_1^5\}$
...	...	...	...	...
(i, j)	x_N^k	y_N^k	$\text{MIN}\{x_N^k \dots x_1^k\}$	$\text{MIN}\{y_N^k \dots y_1^k\}$

图6

当前测序循环N		泛通道MIN状态700	
像素601	通道1 强度值602	通道2 强度值603	泛通道AVG MIN状态值 708
(1, 1)	$1 \times N$	$1 y N$	
(1, 2)	$2 \times N$	$2 y N$	
(1, 3)	$3 \times N$	$3 y N$	图6中的列604和605 的逐像素平均值
(1, 4)	$4 \times N$	$4 y N$	
(1, 5)	$5 \times N$	$5 y N$	
...	...	...	
(i, j)	$k \times N$	$k y N$	

图7

每通道MAX状态800

当前测序循环N

像素601	通道1 强度值602	通道2 强度值603	通道1MAX 状态值804	通道2MAX 状态值805
(1, 1)	x_N^1	y_N^1	$\text{MAX}\{x_N^1 \dots x_1^1\}$	$\text{MAX}\{y_N^1 \dots y_1^1\}$
(1, 2)	x_N^2	y_N^2	$\text{MAX}\{x_N^2 \dots x_1^2\}$	$\text{MAX}\{y_N^2 \dots y_1^2\}$
(1, 3)	x_N^3	y_N^3	$\text{MAX}\{x_N^3 \dots x_1^3\}$	$\text{MAX}\{y_N^3 \dots y_1^3\}$
(1, 4)	x_N^4	y_N^4	$\text{MAX}\{x_N^4 \dots x_1^4\}$	$\text{MAX}\{y_N^4 \dots y_1^4\}$
(1, 5)	x_N^5	y_N^5	$\text{MAX}\{x_N^5 \dots x_1^5\}$	$\text{MAX}\{y_N^5 \dots y_1^5\}$
...	...	...	...	...
(i, j)	x_N^k	y_N^k	$\text{MAX}\{x_N^k \dots x_1^k\}$	$\text{MAX}\{y_N^k \dots y_1^k\}$

图8

当前测序循环N

像素601

像素601	通道1 强度值602	通道2 强度值603	泛通道AVG MAX状态值 908
(1, 1)	1 x N	1 y N	
(1, 2)	2 x N	2 y N	
(1, 3)	3 x N	3 y N	
(1, 4)	4 x N	4 y N	图8中的列804和805 的逐像素平均值
(1, 5)	5 x N	5 y N	
...	...	...	
(i, j)	k x N	k y N	

图9

每通道AVG状态1000

当前测序循环N

像素601	通道1 强度值602	通道2 强度值603	通道1AVG 状态值1004	通道2AVG 状态值1005
(1, 1)	$\begin{matrix} 1 \\ x_N \end{matrix}$	$\begin{matrix} 1 \\ y_N \end{matrix}$	$\text{AVG} \begin{Bmatrix} x_N & \dots & x_N^1 \end{Bmatrix}$	$\text{AVG} \begin{Bmatrix} y_N & \dots & y_N^1 \end{Bmatrix}$
(1, 2)	$\begin{matrix} 2 \\ x_N \end{matrix}$	$\begin{matrix} 2 \\ y_N \end{matrix}$	$\text{AVG} \begin{Bmatrix} x_N & \dots & x_N^2 \end{Bmatrix}$	$\text{AVG} \begin{Bmatrix} y_N & \dots & y_N^2 \end{Bmatrix}$
(1, 3)	$\begin{matrix} 3 \\ x_N \end{matrix}$	$\begin{matrix} 3 \\ y_N \end{matrix}$	$\text{AVG} \begin{Bmatrix} x_N & \dots & x_N^3 \end{Bmatrix}$	$\text{AVG} \begin{Bmatrix} y_N & \dots & y_N^3 \end{Bmatrix}$
(1, 4)	$\begin{matrix} 4 \\ x_N \end{matrix}$	$\begin{matrix} 4 \\ y_N \end{matrix}$	$\text{AVG} \begin{Bmatrix} x_N & \dots & x_N^4 \end{Bmatrix}$	$\text{AVG} \begin{Bmatrix} y_N & \dots & y_N^4 \end{Bmatrix}$
(1, 5)	$\begin{matrix} 5 \\ x_N \end{matrix}$	$\begin{matrix} 5 \\ y_N \end{matrix}$	$\text{AVG} \begin{Bmatrix} x_N & \dots & x_N^5 \end{Bmatrix}$	$\text{AVG} \begin{Bmatrix} y_N & \dots & y_N^5 \end{Bmatrix}$
...	...	...	...	...
(i, j)	$\begin{matrix} k \\ x_N \end{matrix}$	$\begin{matrix} k \\ y_N \end{matrix}$	$\text{AVG} \begin{Bmatrix} x_N & \dots & x_N^k \end{Bmatrix}$	$\text{AVG} \begin{Bmatrix} y_N & \dots & y_N^k \end{Bmatrix}$

图10

泛通道AVG AVG状态1100



当前测序循环N

像素601	通道1 强度值602	通道2 强度值603	泛通道AVG AVG状态值 1108
(1, 1)	1 x N	1 y N	
(1, 2)	2 x N	2 y N	
(1, 3)	3 x N	3 y N	图10中的列1004和1005 的逐像素平均值
(1, 4)	4 x N	4 y N	
(1, 5)	5 x N	5 y N	
...	...	...	
(i, j)	k x N	k y N	

图11

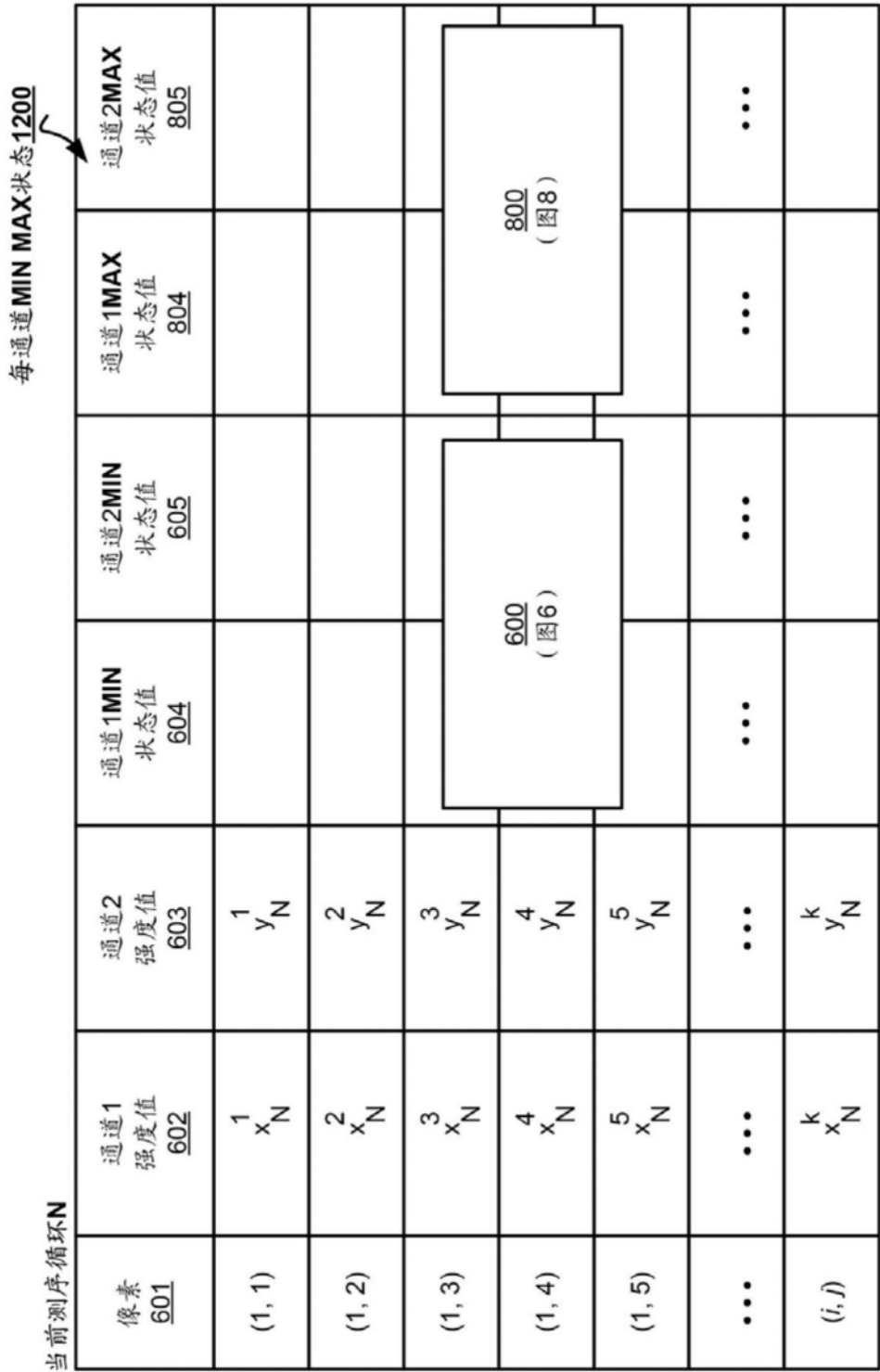


图12

每通道MIN AVG状态1300

当前测序循环N	像素 601	通道1 强度值 602	通道2 强度值 603	通道1MIN 状态值 604	通道2MIN 状态值 605	通道1AVG 状态值 1004	通道2AVG 状态值 1005
	(1, 1)	1_{x_N}	1_{y_N}				
	(1, 2)	2_{x_N}	2_{y_N}				
	(1, 3)	3_{x_N}	3_{y_N}				
	(1, 4)	4_{x_N}	4_{y_N}				
	(1, 5)	5_{x_N}	5_{y_N}				
	...	...	...	...	...	...	...
	(i, j)	k_{x_N}	k_{y_N}				

600图6

1000图10

图13

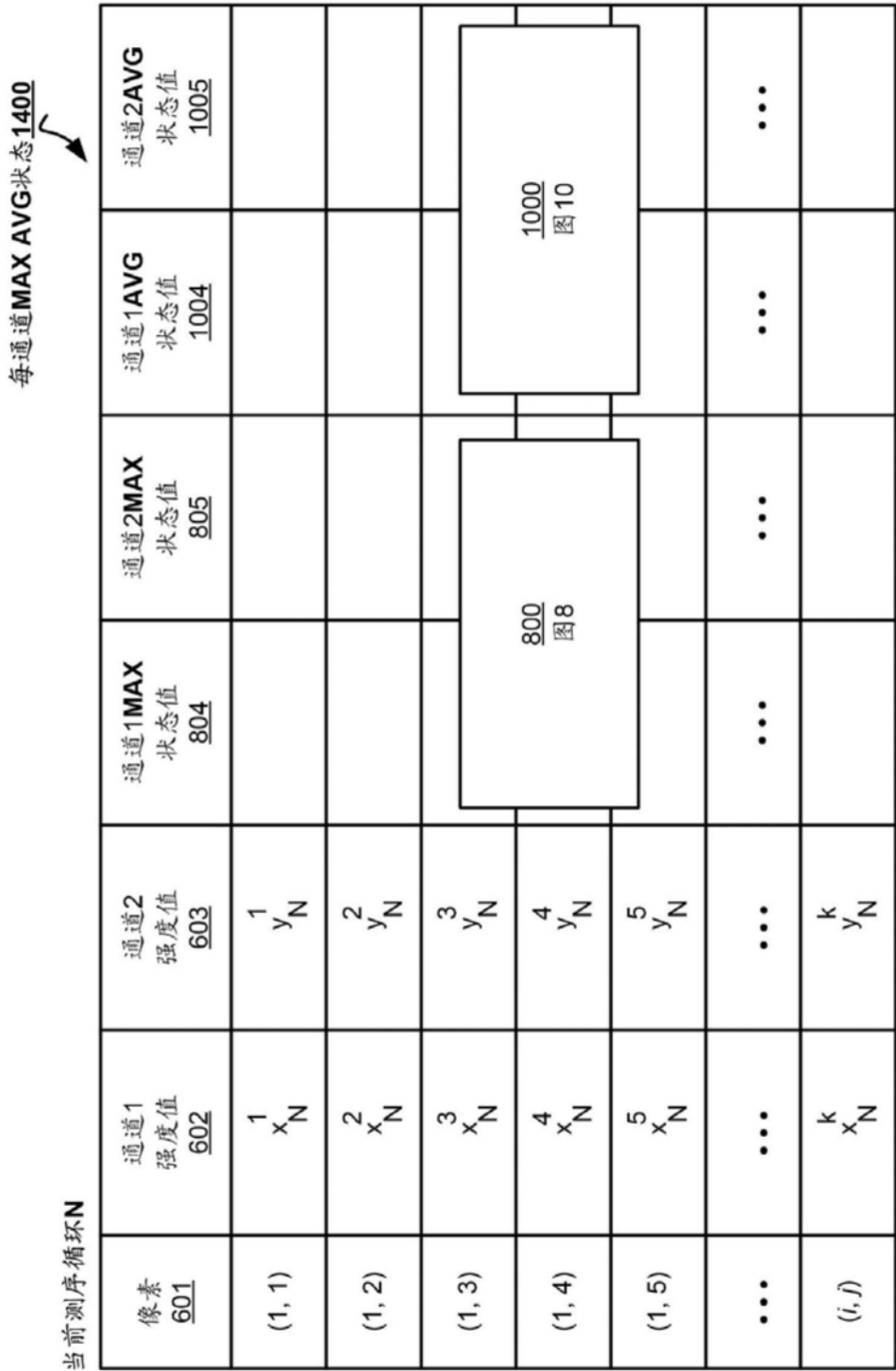


图14

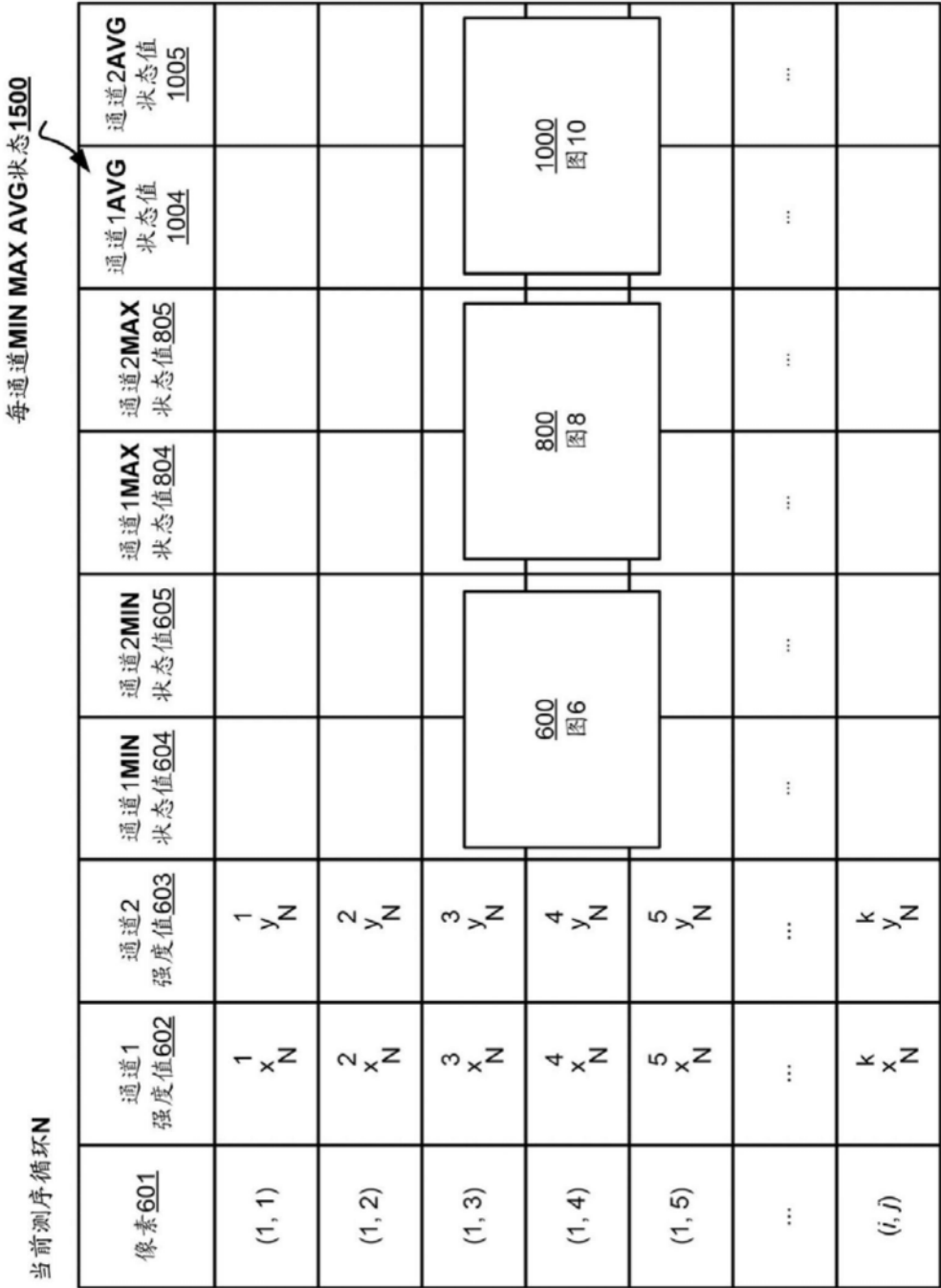


图15

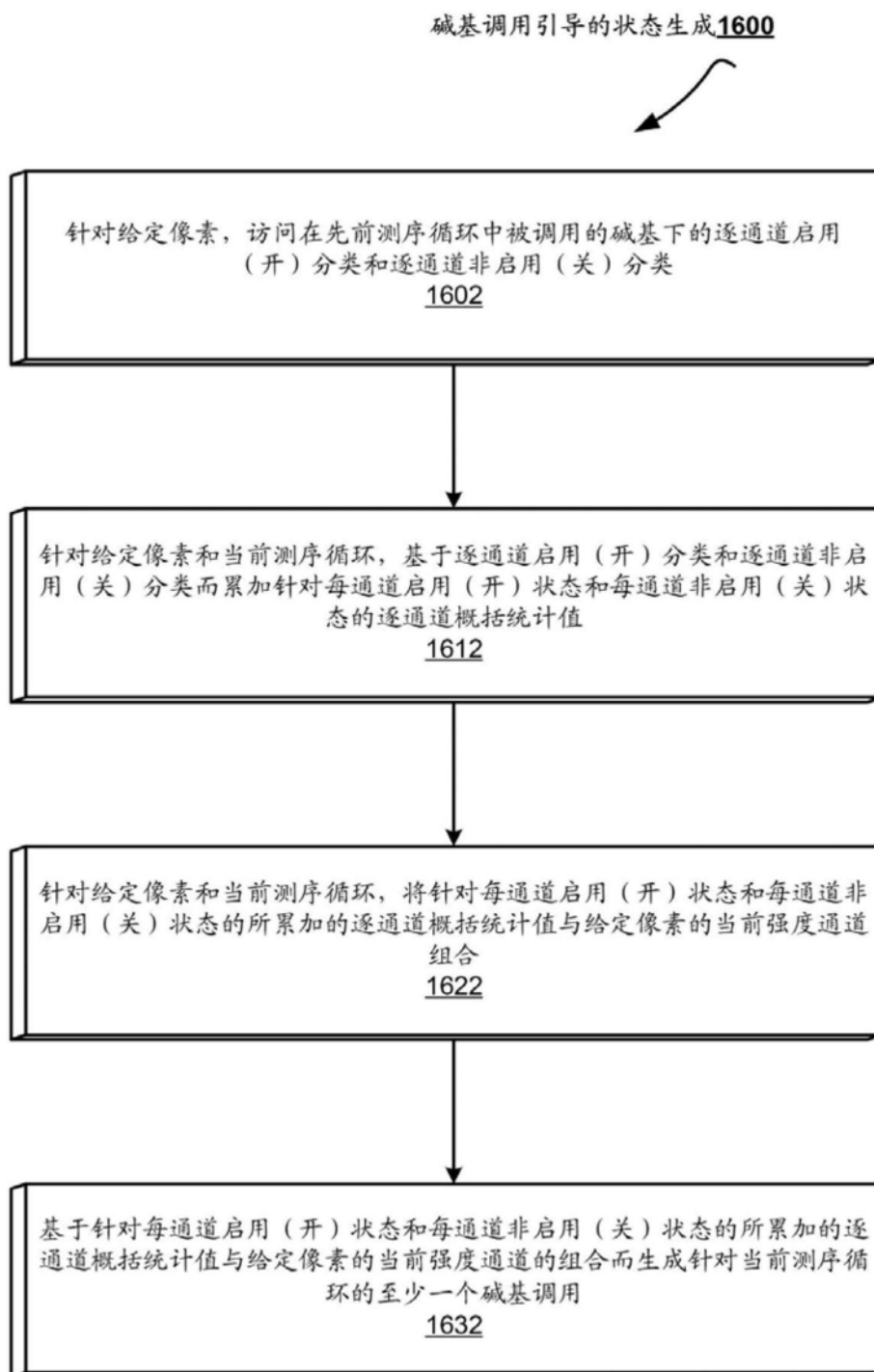


图16

碱基调用引导的开和关状态生成1700

像素P		通道1	通道2
循环1	碱基调用G	关 x_1^P	关 y_1^P
循环2	碱基调用C	关 x_2^P	开 y_2^P
循环3	碱基调用T	开 x_3^P	开 y_3^P
循环4	碱基调用A	开 x_4^P	关 y_4^P
循环5	碱基调用C	关 x_5^P	开 y_5^P
循环6	碱基调用G	关 x_6^P	关 y_6^P
循环7	碱基调用T	开 x_7^P	开 y_7^P
循环8	碱基调用A	开 x_8^P	关 y_8^P
循环9	碱基调用C	关 x_9^P	开 y_9^P
循环10	碱基调用T	开 x_{10}^P	开 y_{10}^P

图17A

针对循环11的每通道状态计算		$f = \text{AVG, MAX, MIN, EXPWEGAVG, MOVAVG, EXPMOVAVG, STDEV}$
通道1非启用 (关) 状态	1702	$f\{x_1^P, x_2^P, x_5^P, x_6^P, x_9^P\}$
通道1启用 (开) 状态	1712	$f\{x_3^P, x_4^P, x_7^P, x_8^P, x_{10}^P\}$
通道2非启用 (关) 状态	1722	$f\{y_1^P, y_4^P, y_6^P, y_8^P\}$
通道2启用 (开) 状态	1732	$f\{y_2^P, y_3^P, y_5^P, y_7^P, y_9^P, y_{10}^P\}$

图17B

循环11	通道1	通道2	1702	1712	1722	1732
	x_{11}^P	y_{11}^P				

图17C

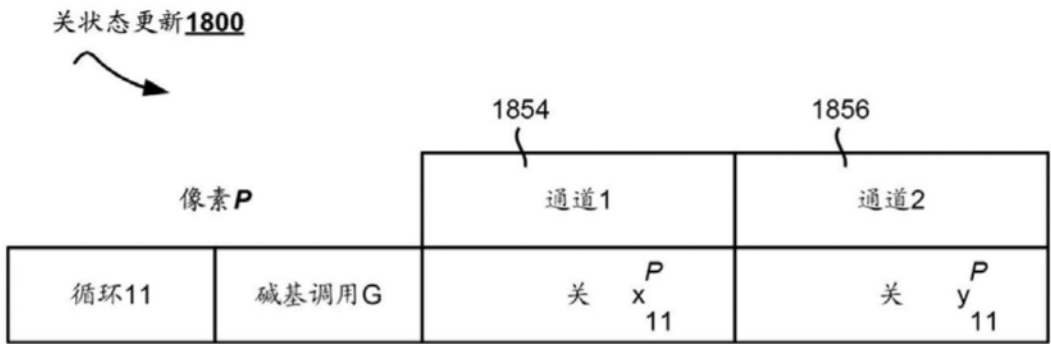


图18A

针对循环12的每通道状态计算		$f = \text{AVG, MAX, MIN, EXPWEGAVG, MOVAVG, EXPMOVAVG, STDEV}$
通道1 非启用（关）状态	1802	$f\{ \underline{1702}, x_{11}^P \}$
通道1 启用（开）状态	1812	<u>1712</u> （没有变化）
通道2 非启用（关）状态	1822	$f\{ \underline{1722}, y_{11}^P \}$
通道2 启用（开）状态	1832	<u>1732</u> （没有变化）

图18B

循环12	1864 通道1	1866 通道2	<u>1802</u>	<u>1812</u>	<u>1822</u>	<u>1832</u>
	x_{12}^P	y_{12}^P				

图18C

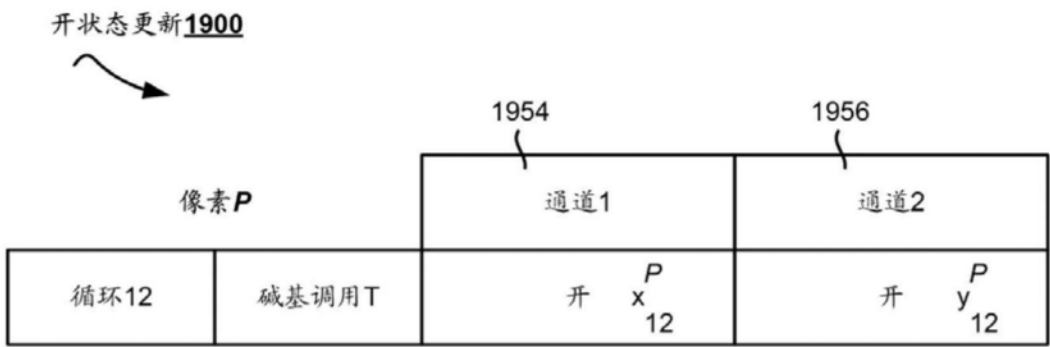


图19A

针对循环13的每通道状态计算		$f = \text{AVG}, \text{MAX}, \text{MIN}, \text{EXPWEGAVG}, \text{MOVAVG}, \text{EXPMOVAVG}, \text{STDEV}$	
通道1 非启用（关）状态	1902	1802	（没有变化）
通道1 启用（开）状态	1912	$f\{1812, x_{12}^P\}$	
通道2 非启用（关）状态	1922	1822	（没有变化）
通道2 启用（开）状态	1932	$f\{1832, y_{12}^P\}$	

图19B

循环13	通道1	通道2	1902	1913	1922	1932
	x_{13}^P	y_{13}^P				

图19C

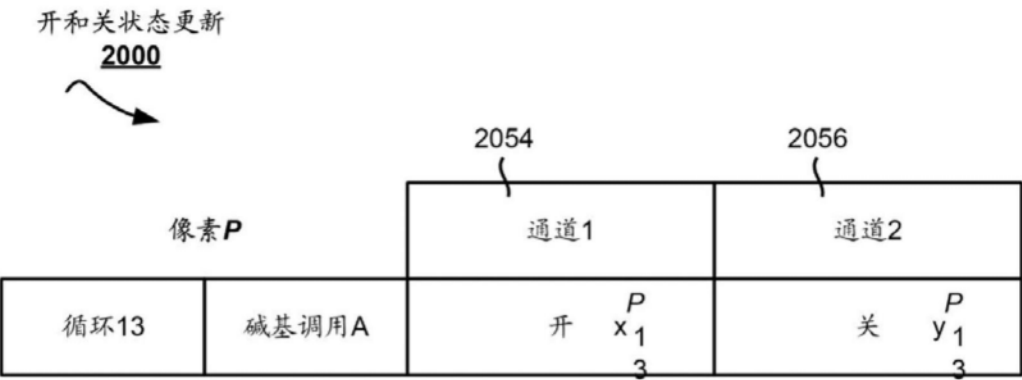


图20A

针对循环14的每通道状态计算		$f = \text{AVG}, \text{MAX}, \text{MIN}, \text{EXPWEGAVG}, \text{MOVAVG}, \text{EXPMOVAVG}, \text{STDEV}$
通道1 非启用（关）状态	2002	<u>1902</u> （没有变化）
通道1 启用（开）状态	2012	$f\{ \underline{1912}, x_{13}^P \}$
通道2 非启用（关）状态	2022	$f\{ \underline{1922}, y_{13}^P \}$
通道2 启用（开）状态	2032	<u>1932</u> （没有变化）

图20B

循环14	通道1	通道2	<u>2002</u>	<u>2014</u>	<u>2022</u>	<u>2032</u>
	x_{14}^P	y_{14}^P				

图20C

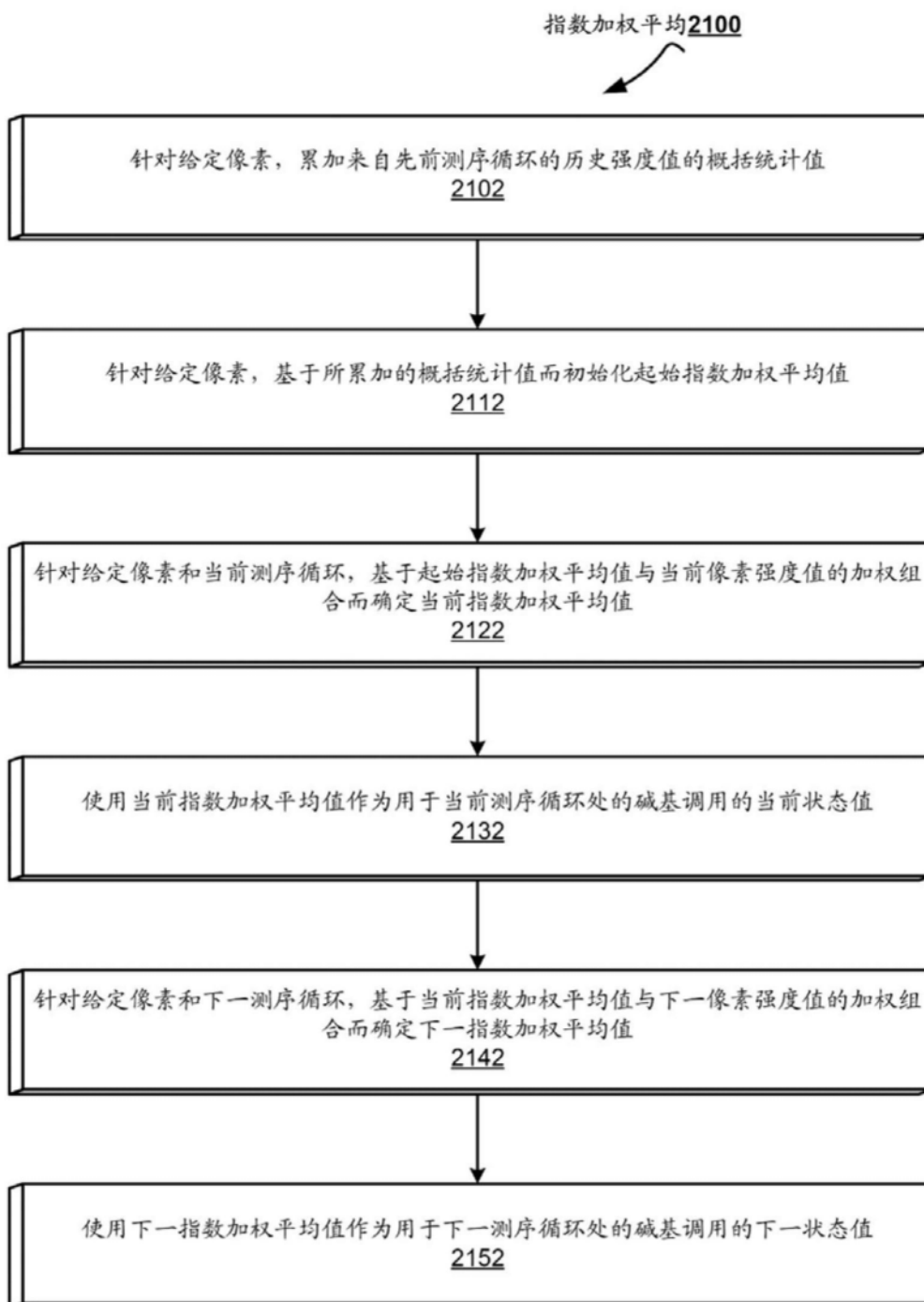


图21

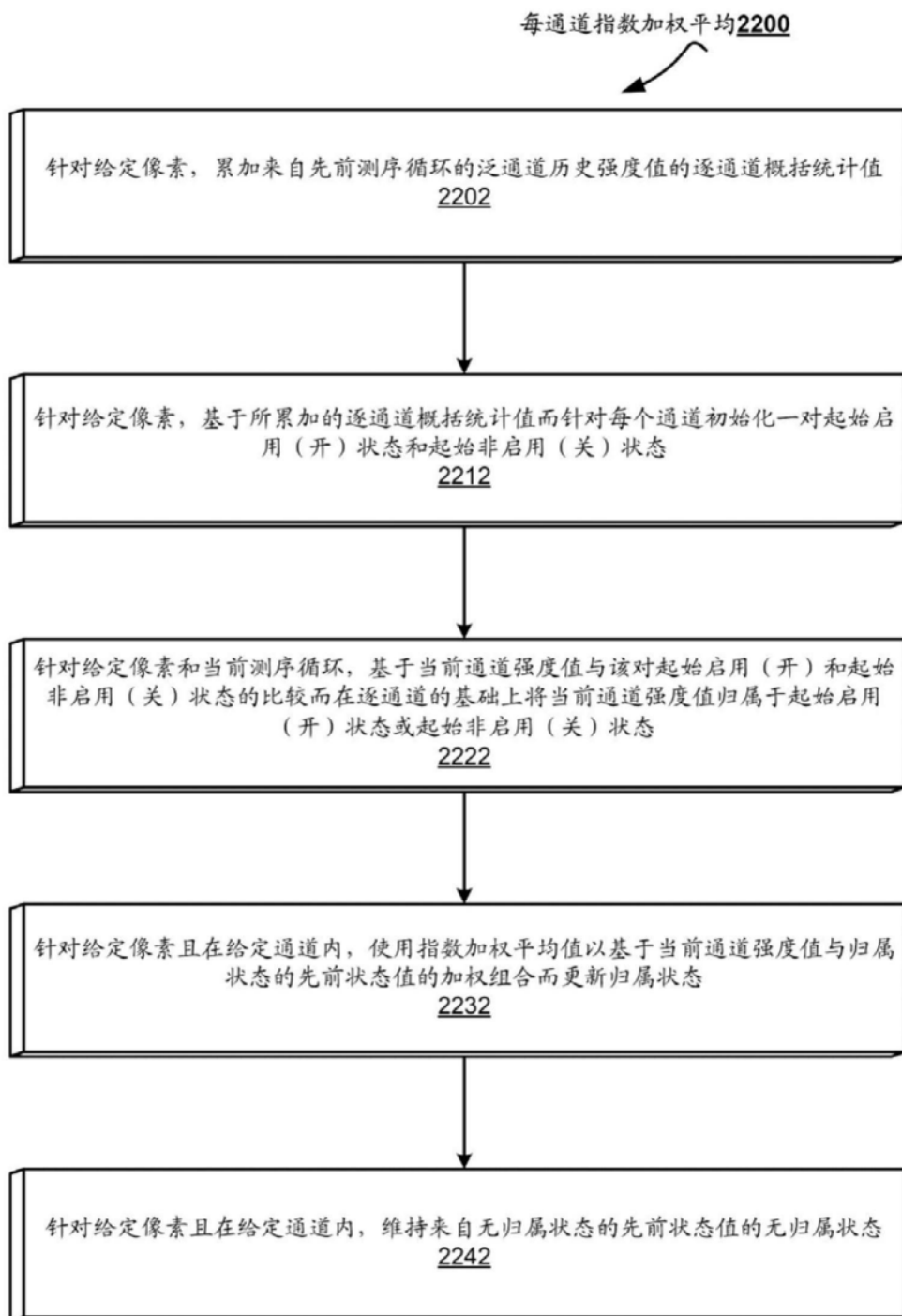


图22

像素P 使用全局最大值和全局最小值以及指数加权平均值的每通道开和关状态确定2300

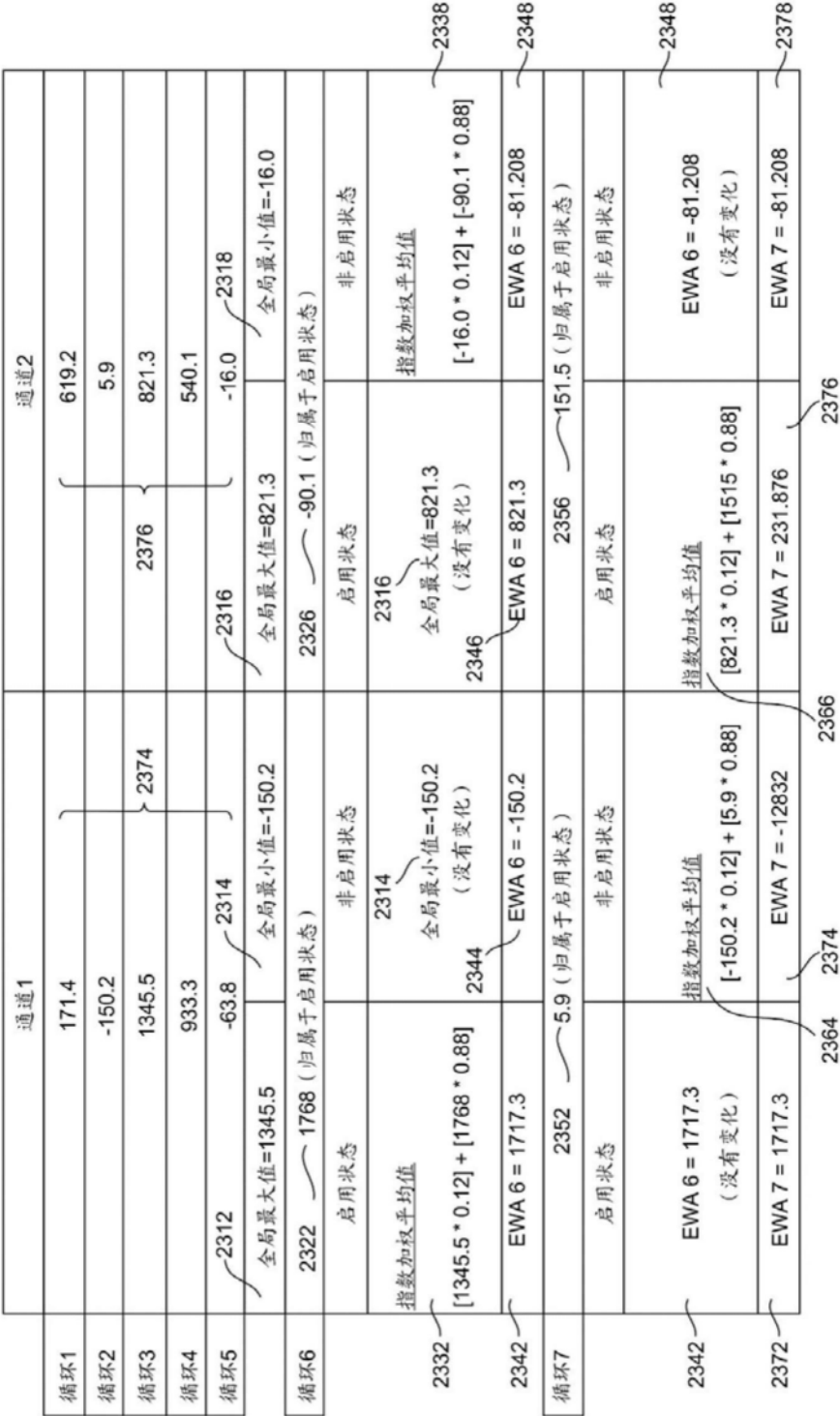


图23

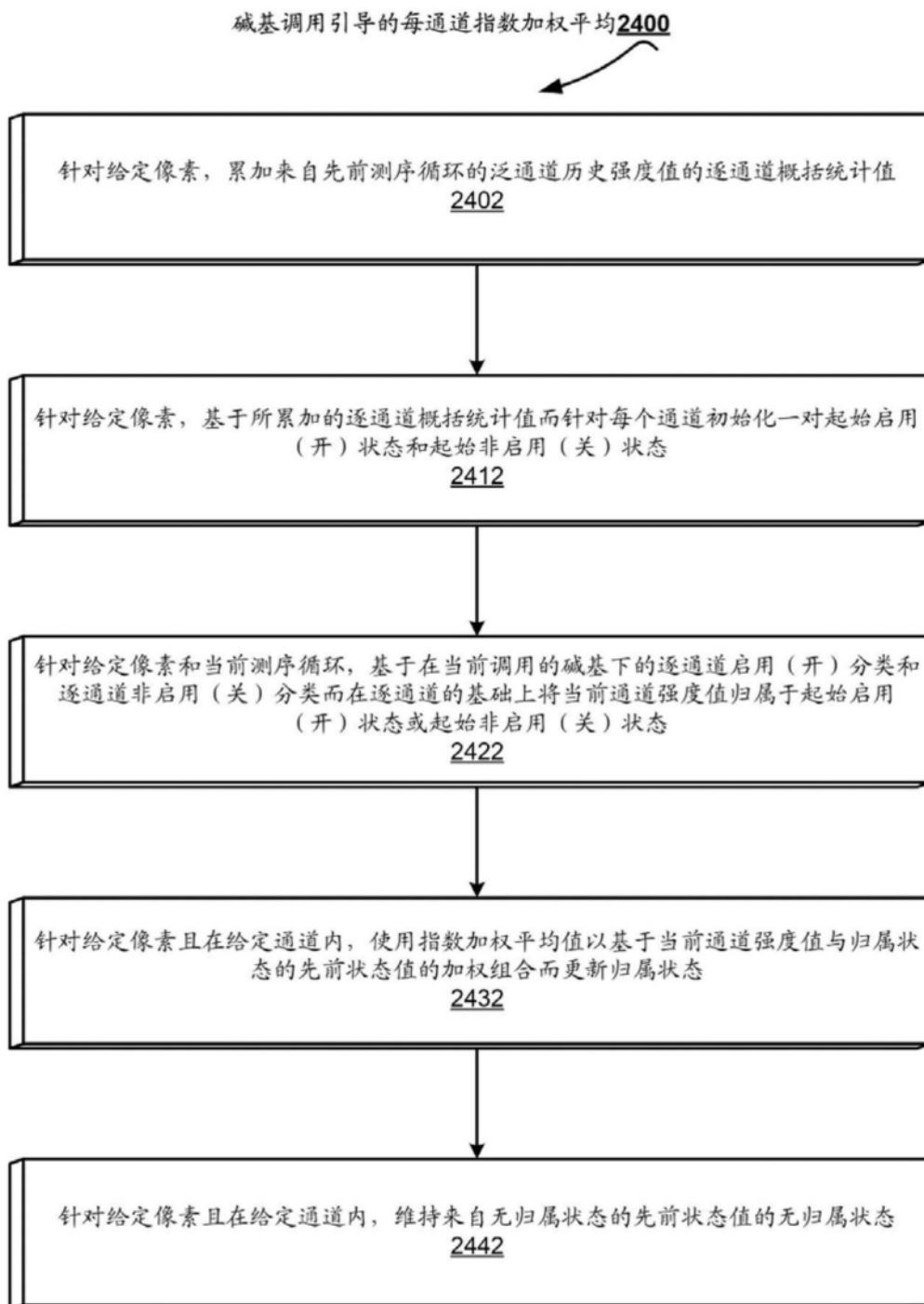


图24

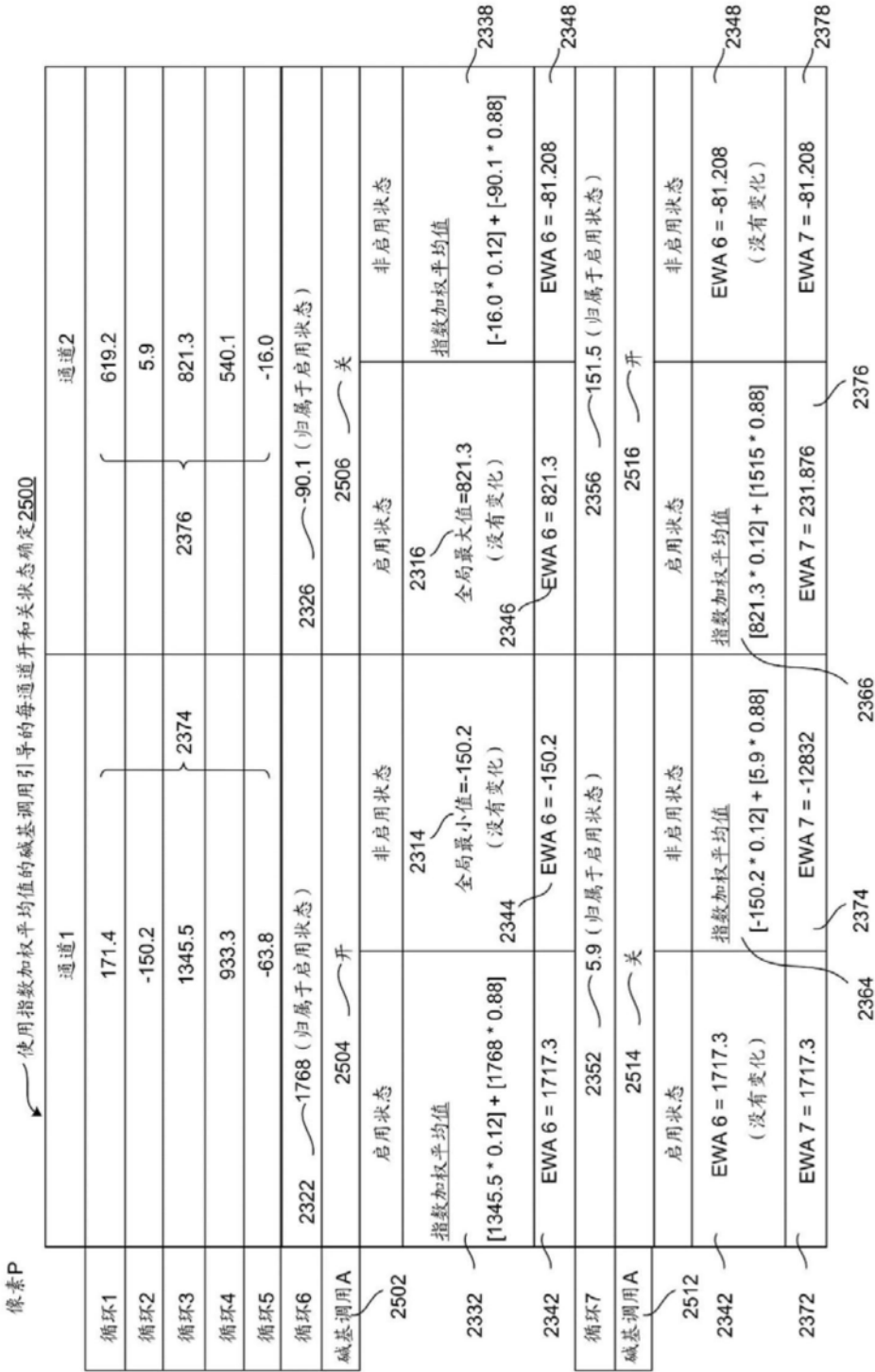


图25

针对非簇像素的碱基调用引导的状态生成2600

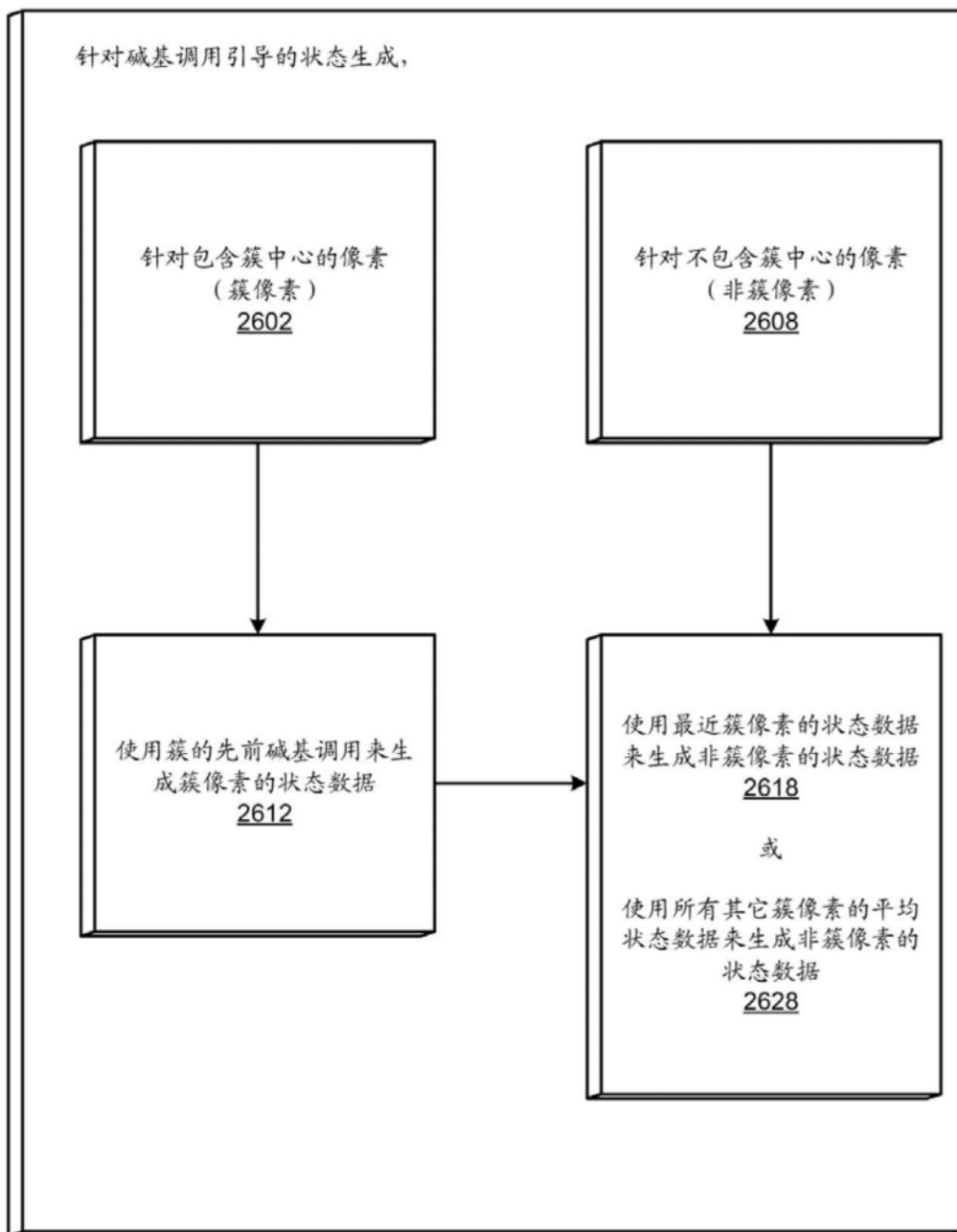


图26

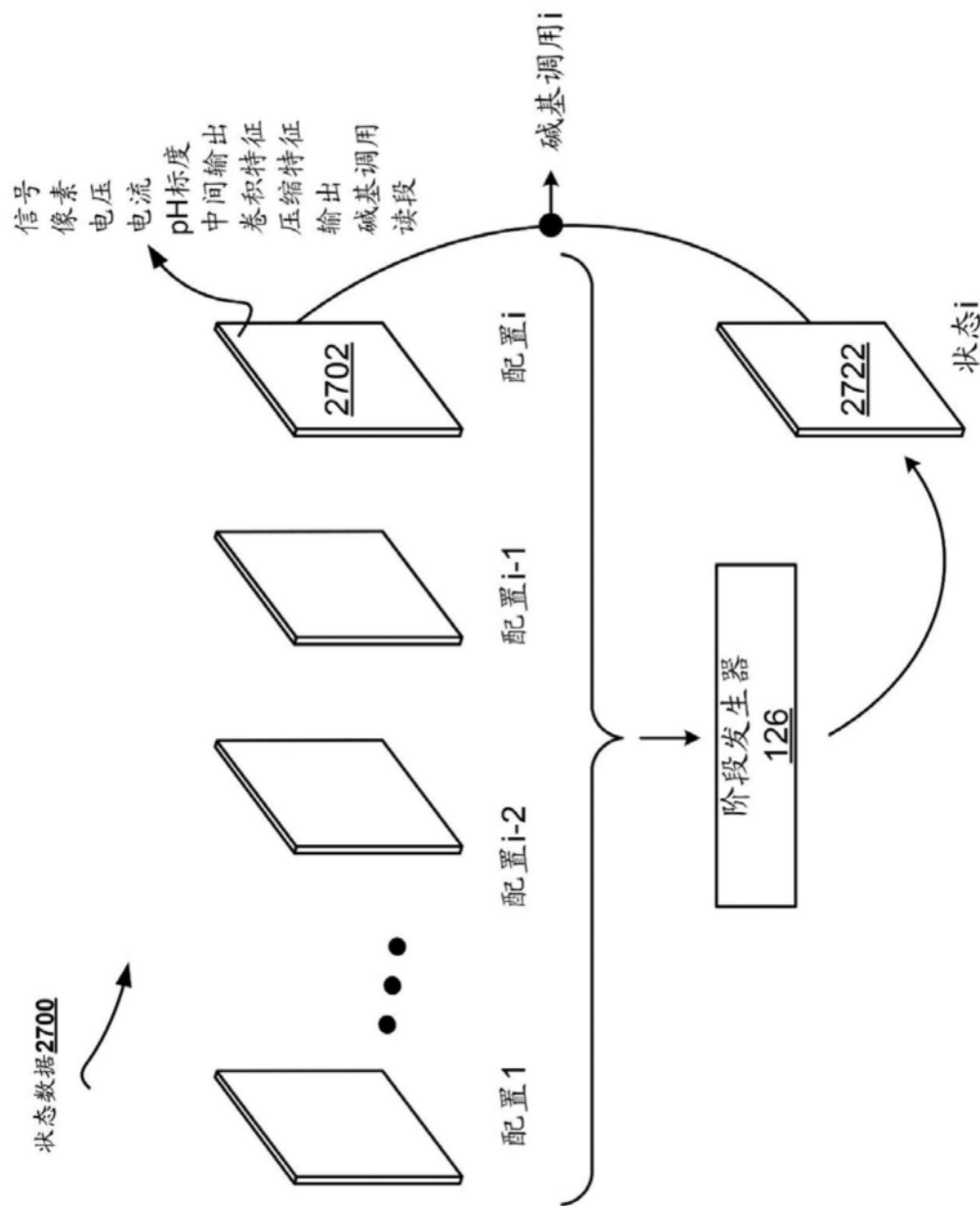


图27

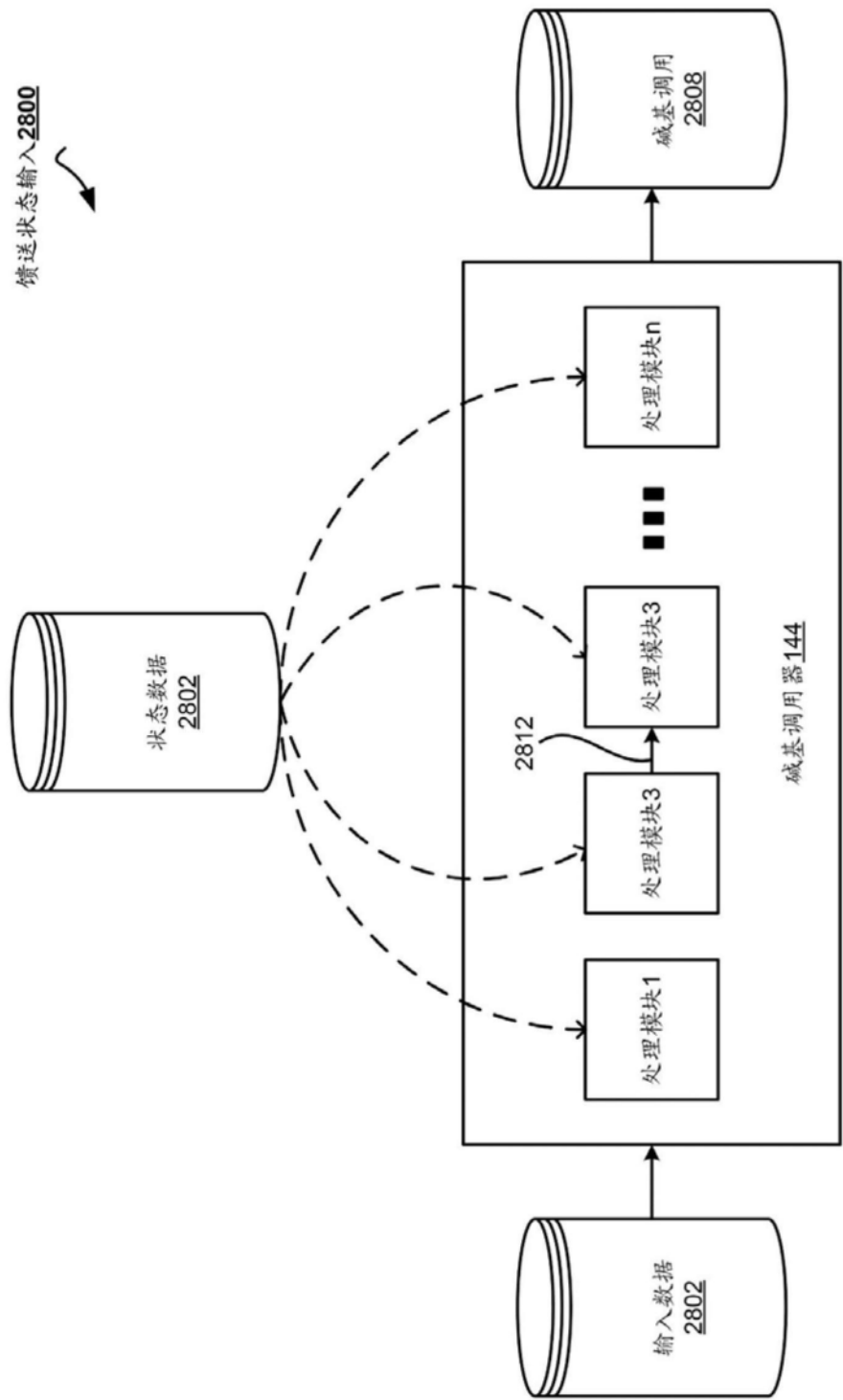


图28

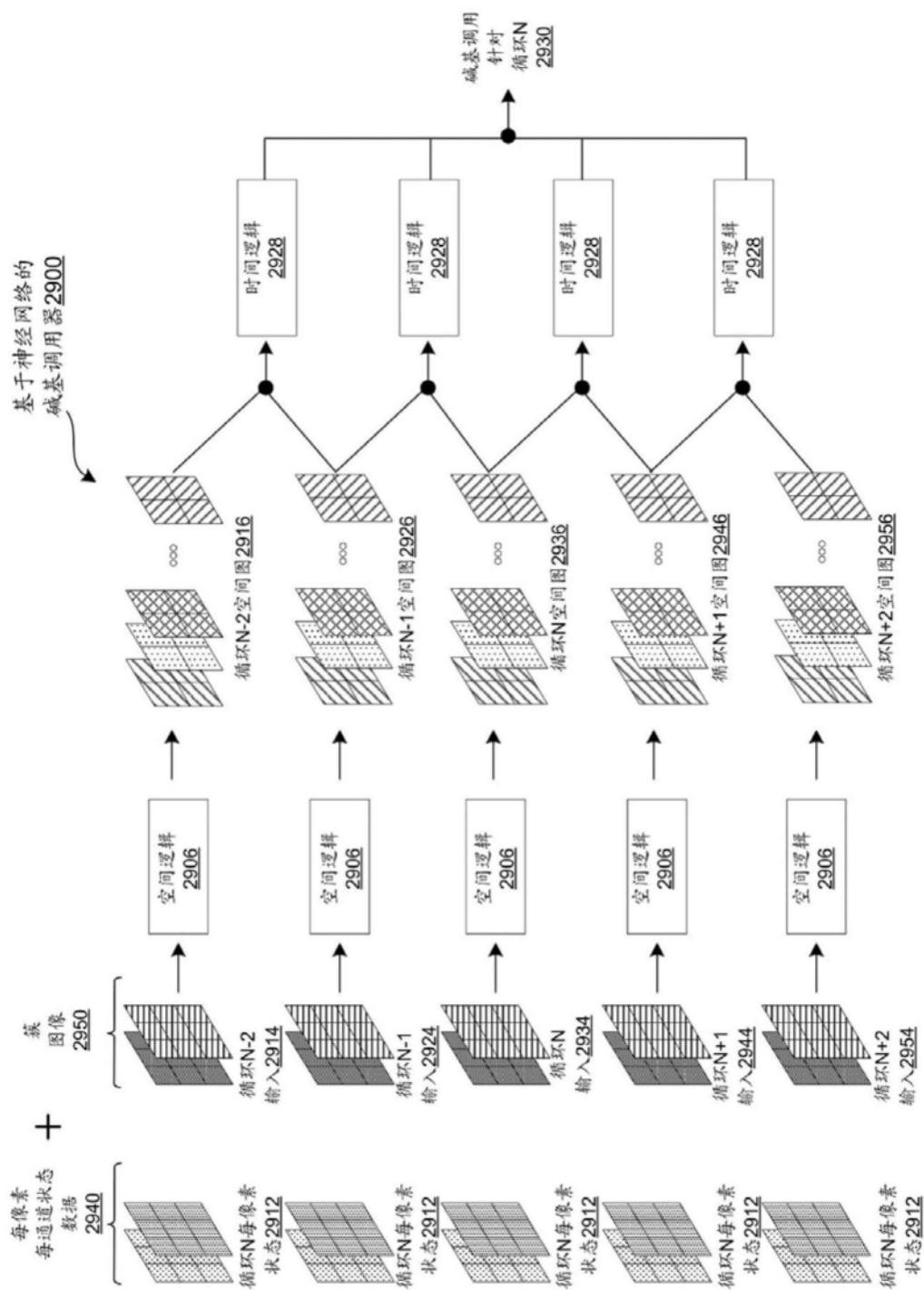


图29

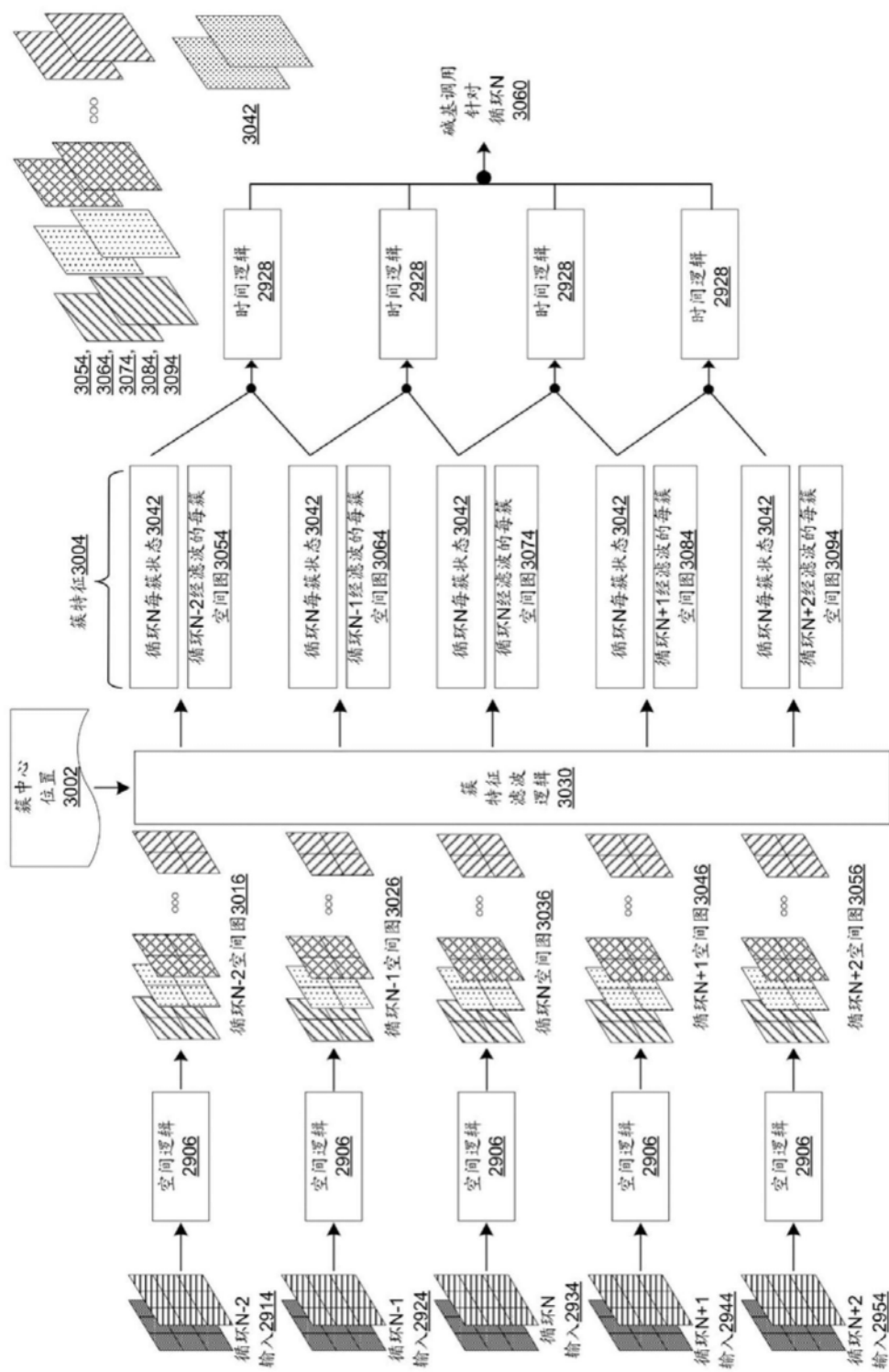


图30

基于每簇像素的状态生成3100



图31

基于每簇空间特征的状态生成3200

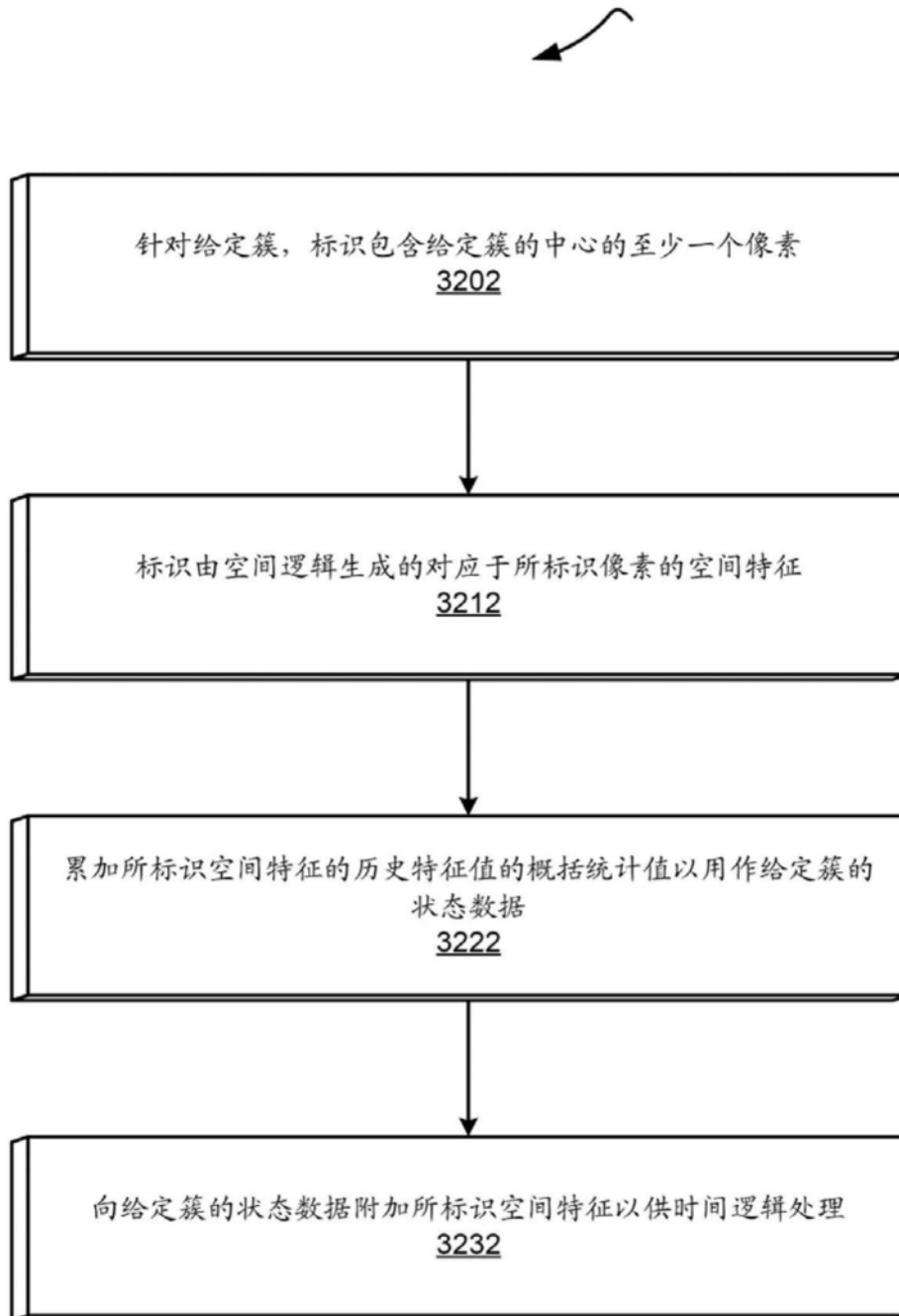


图32

基于每簇插值特征的状态生成3300

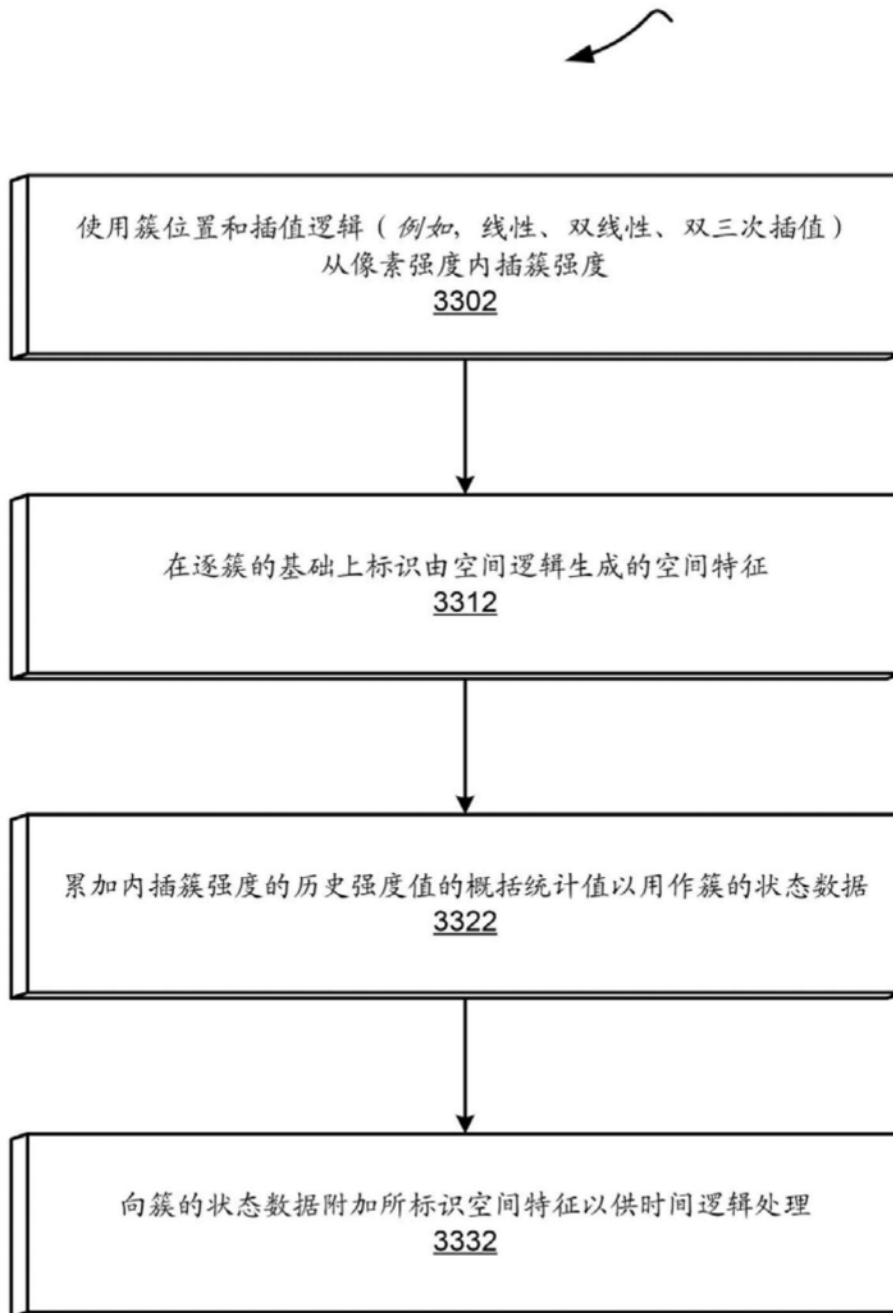


图33

基于每簇RTA系数的状态生成3400

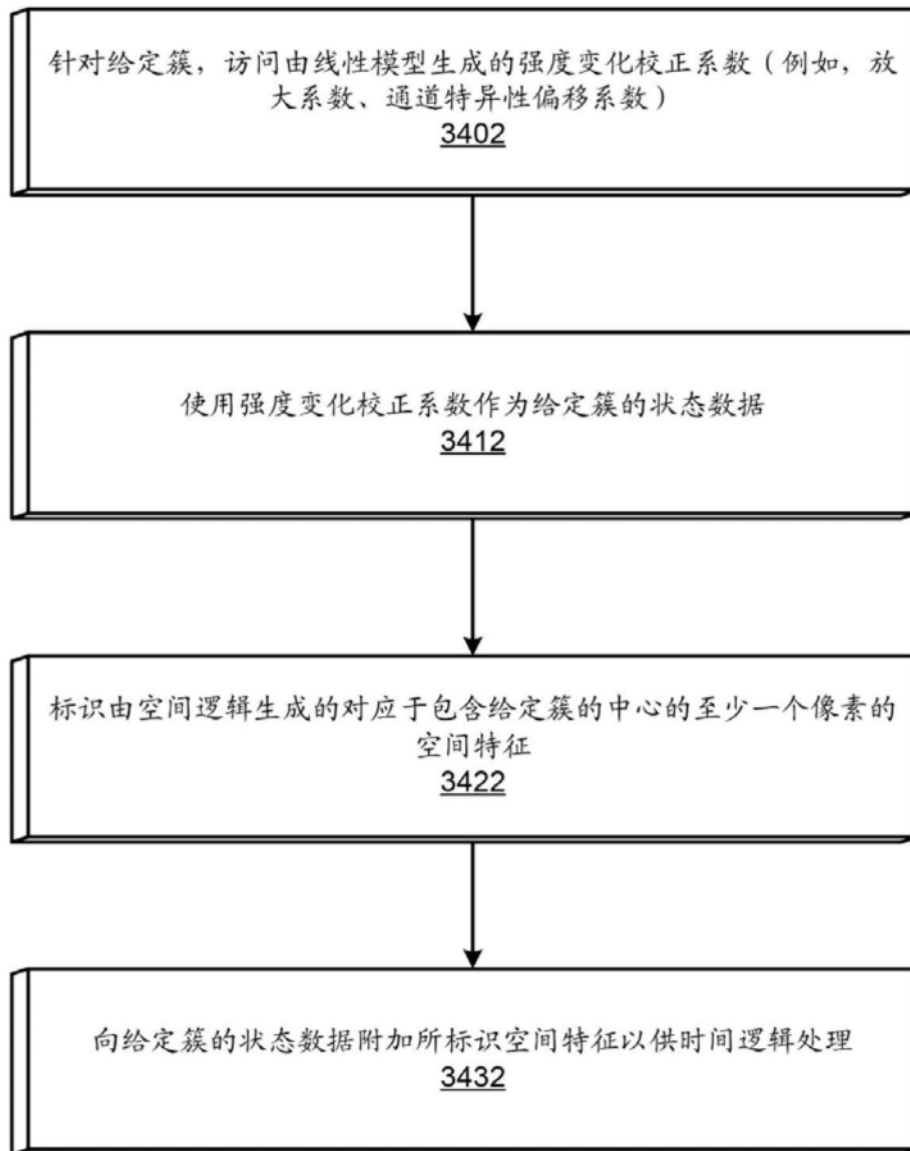


图34

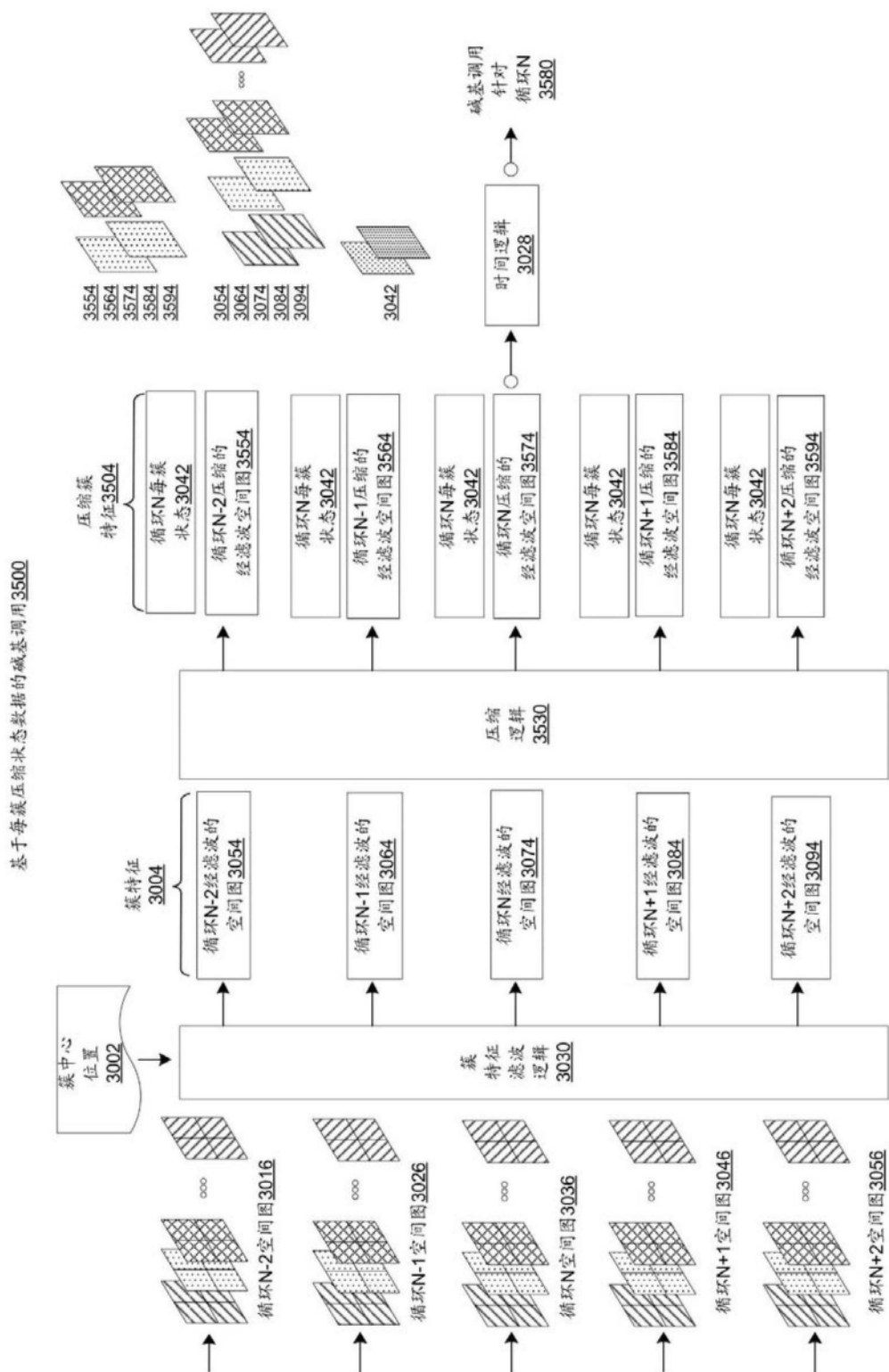


图35

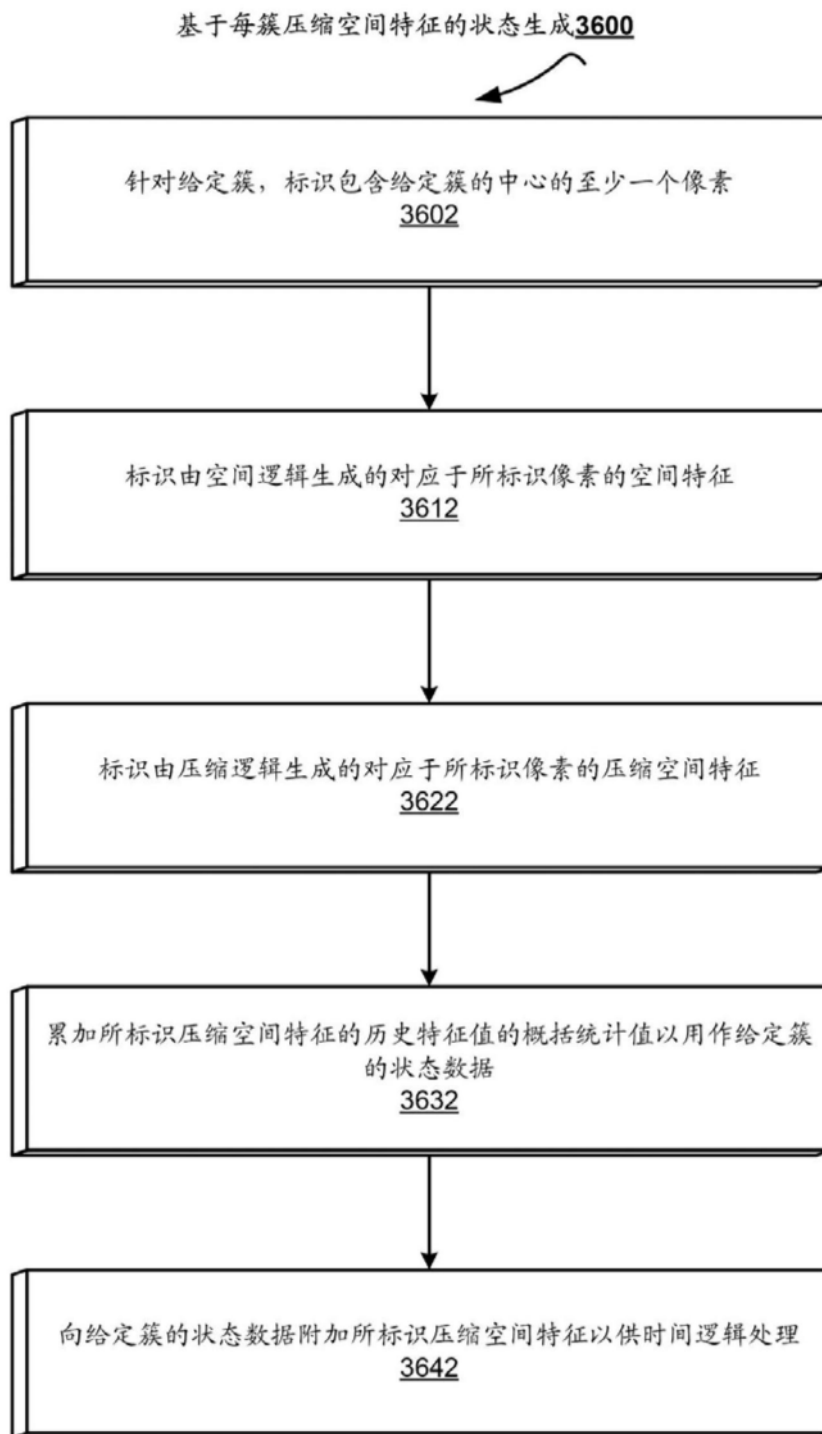


图36

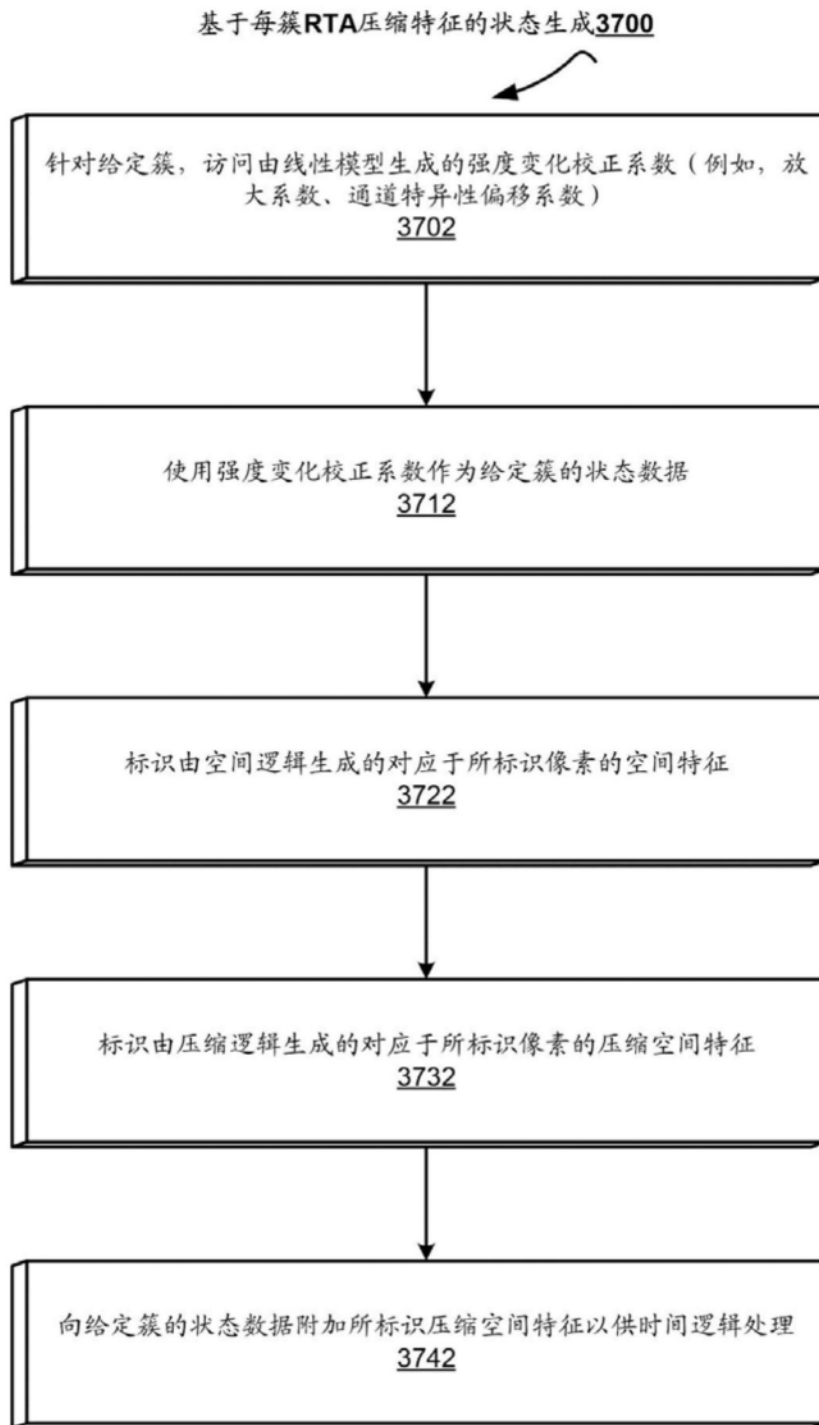


图37

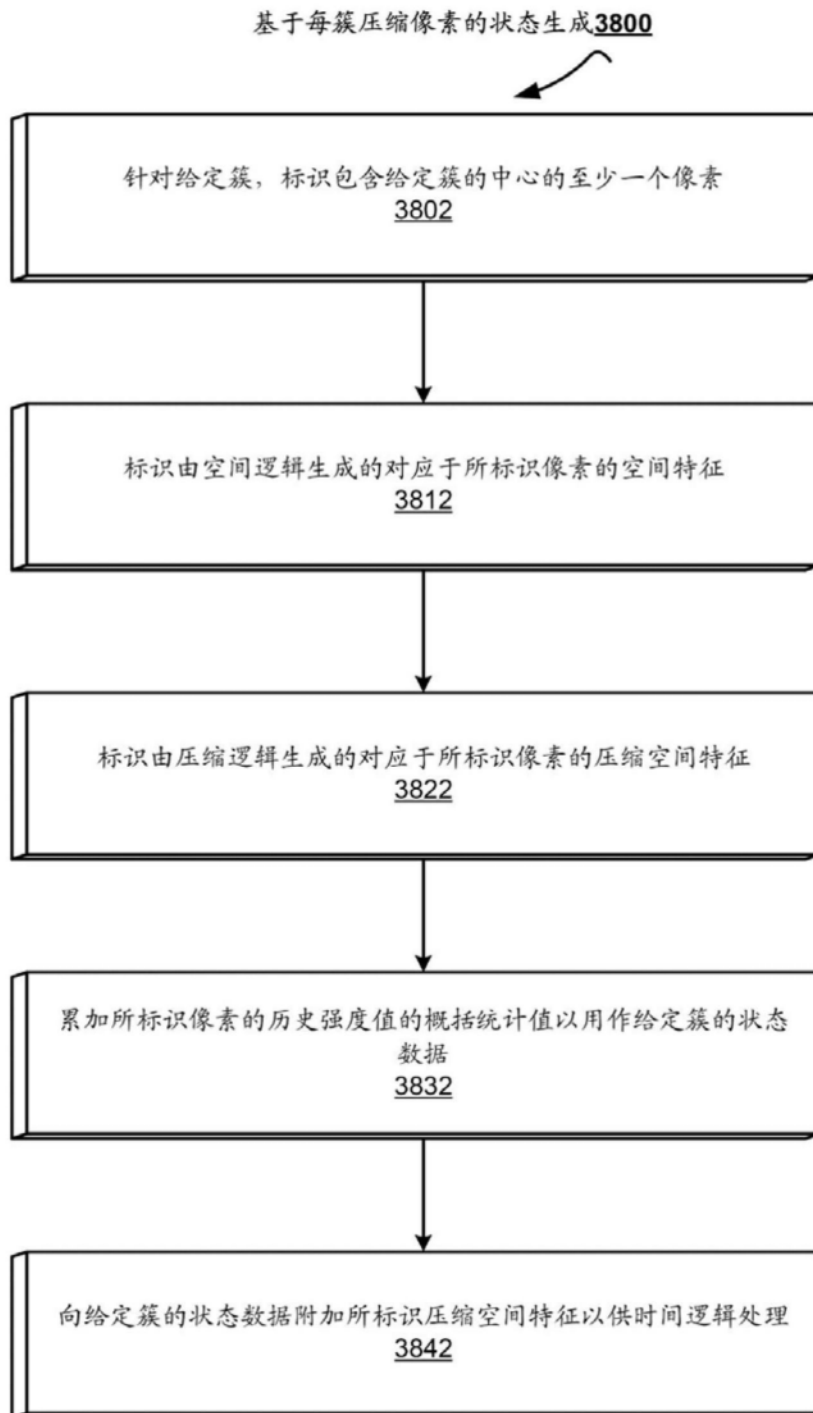


图38

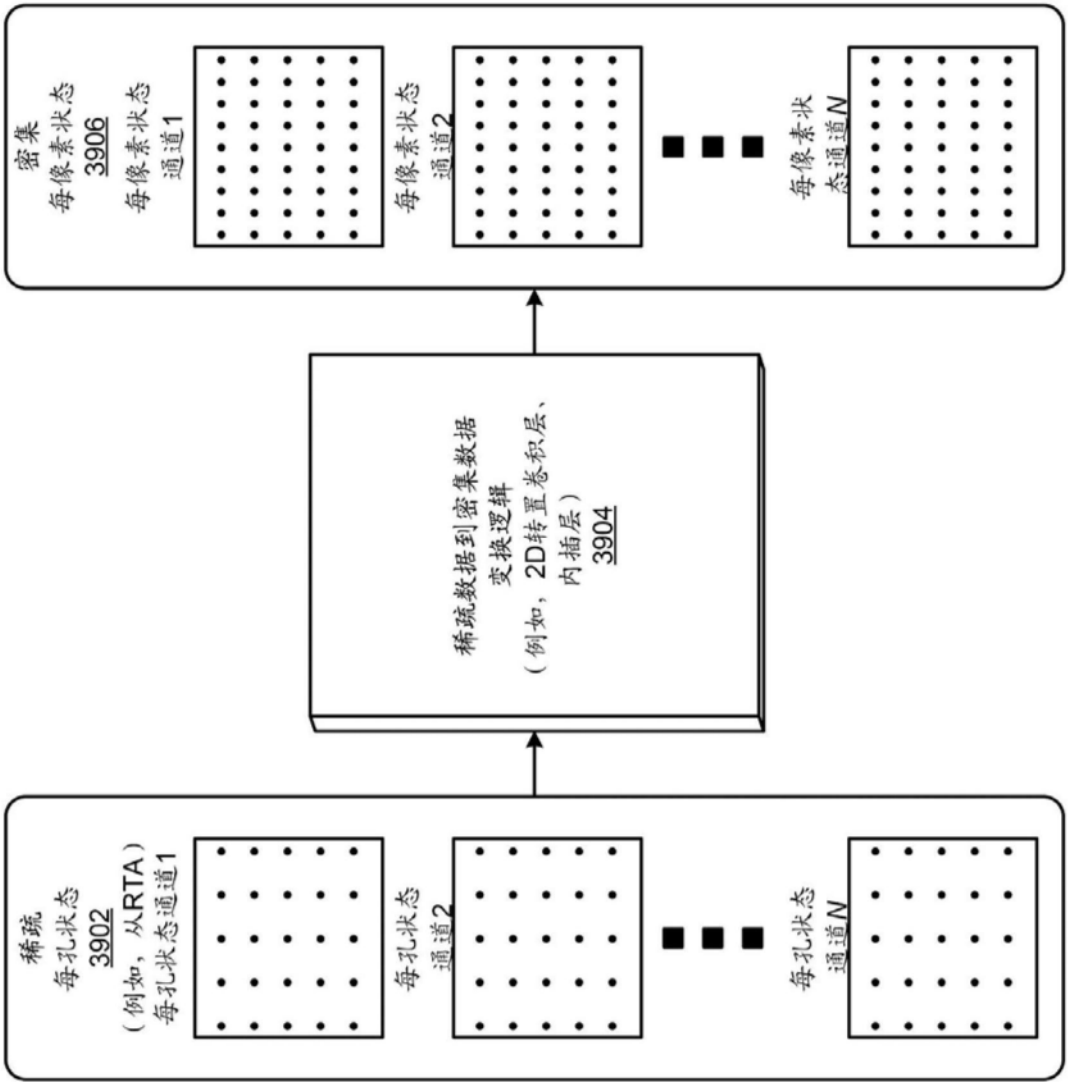


图39

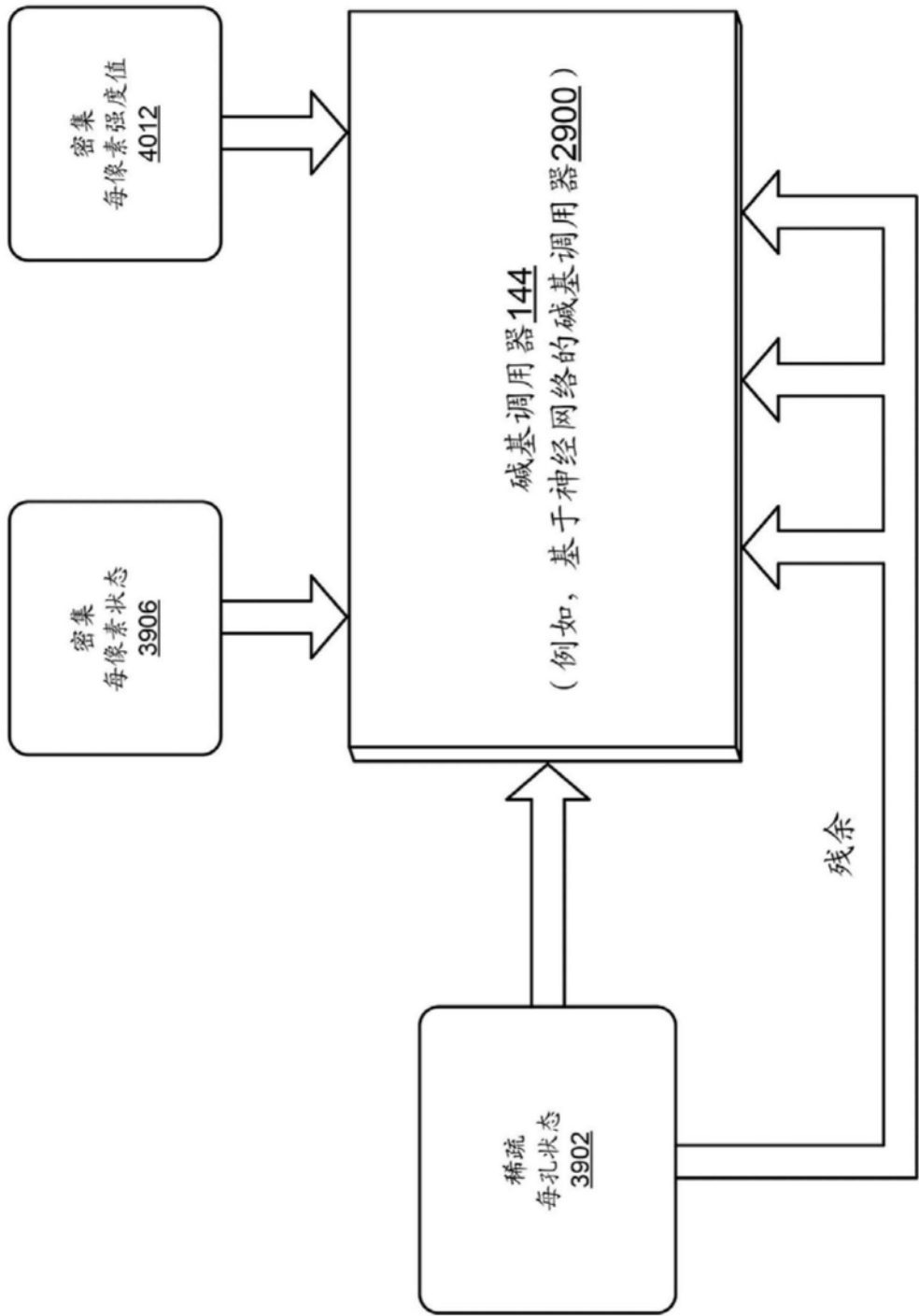


图40

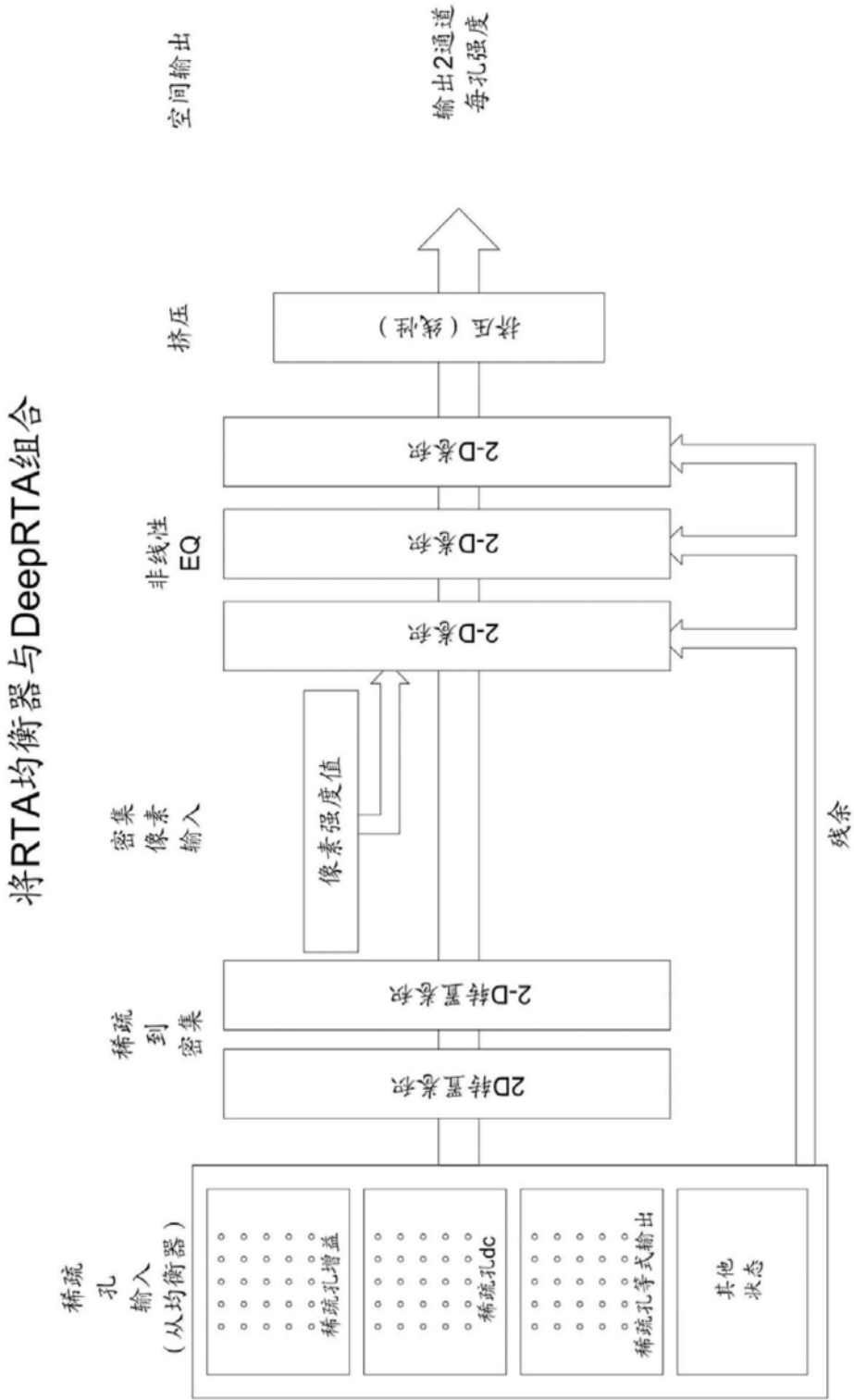


图41

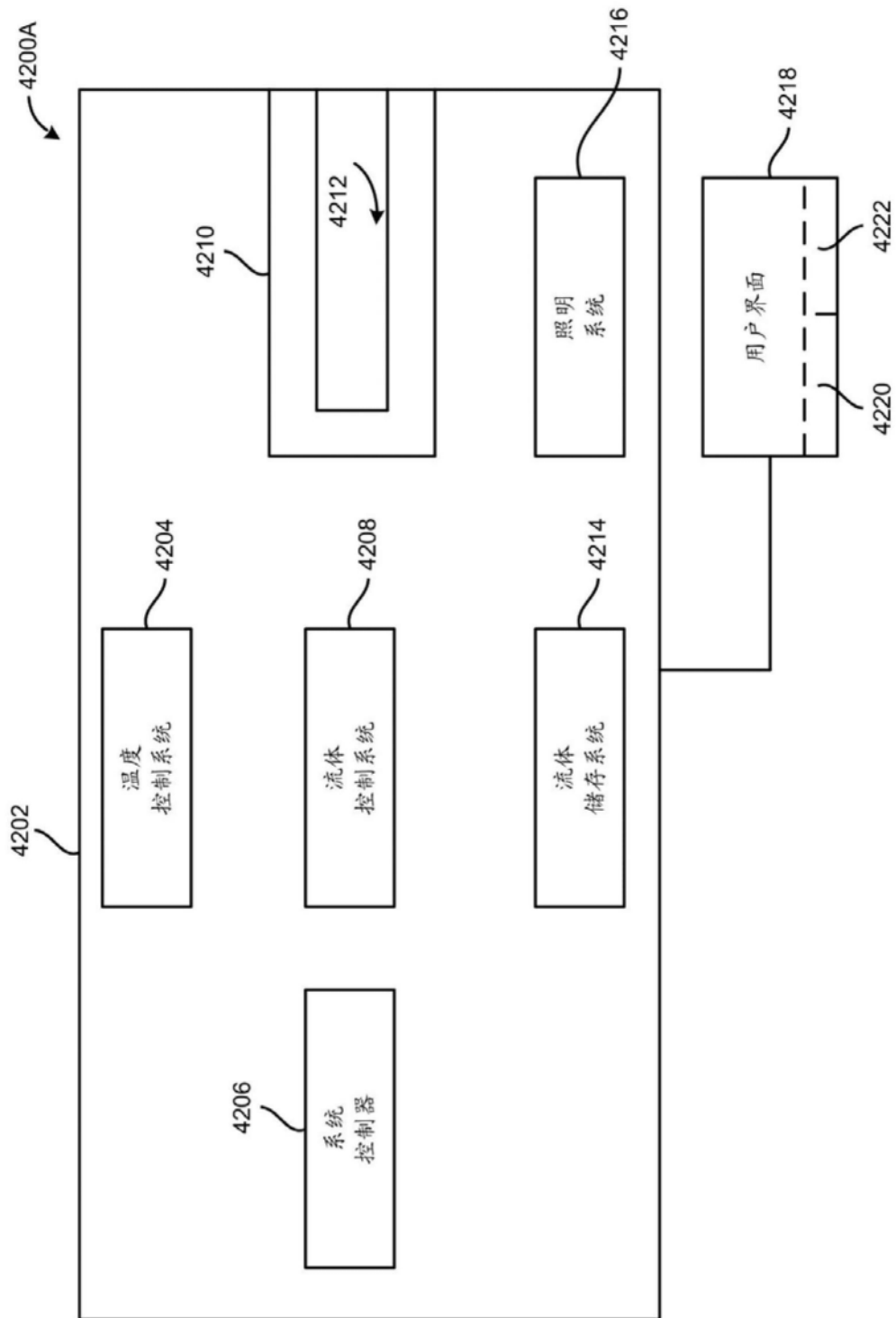


图42A

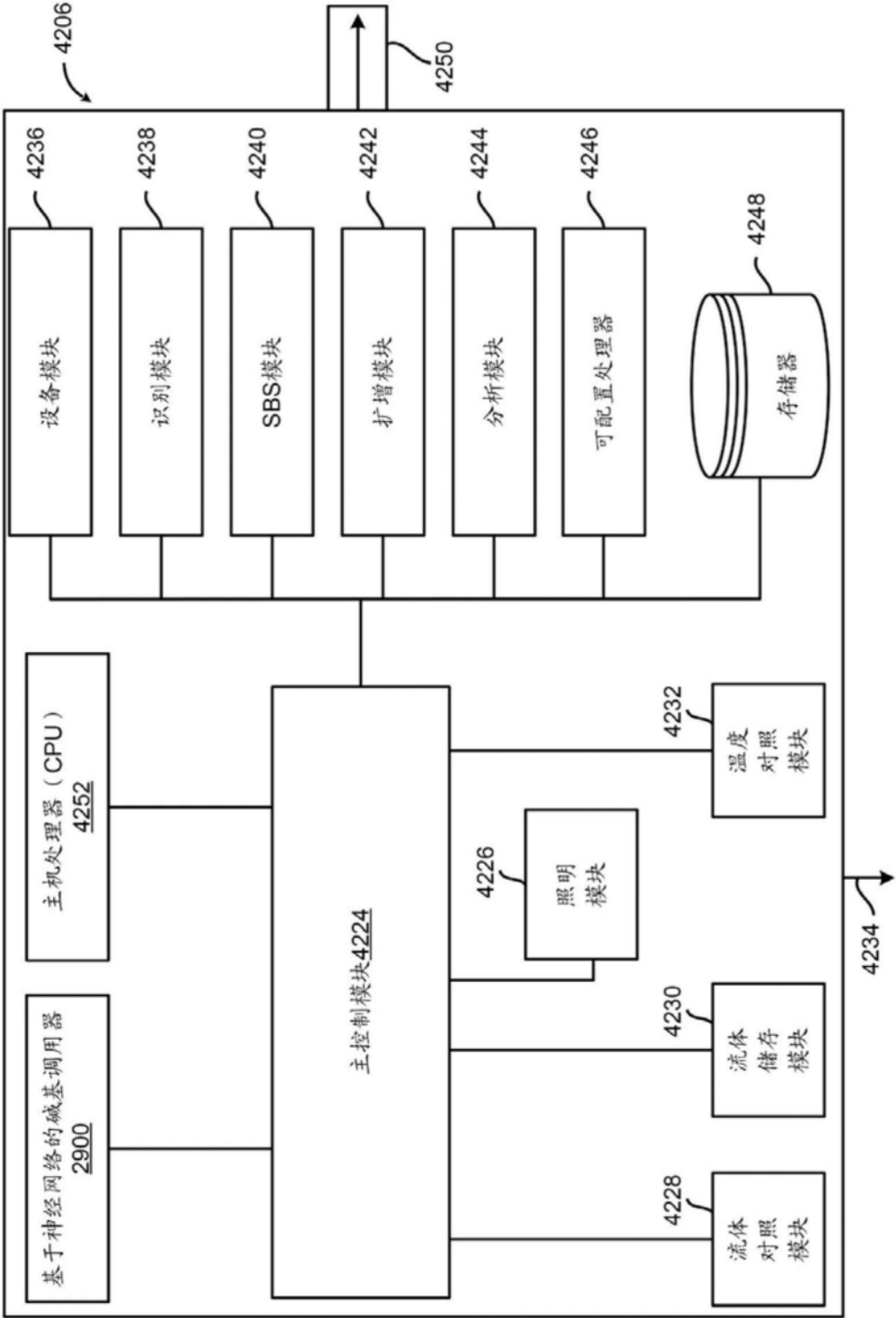


图42B

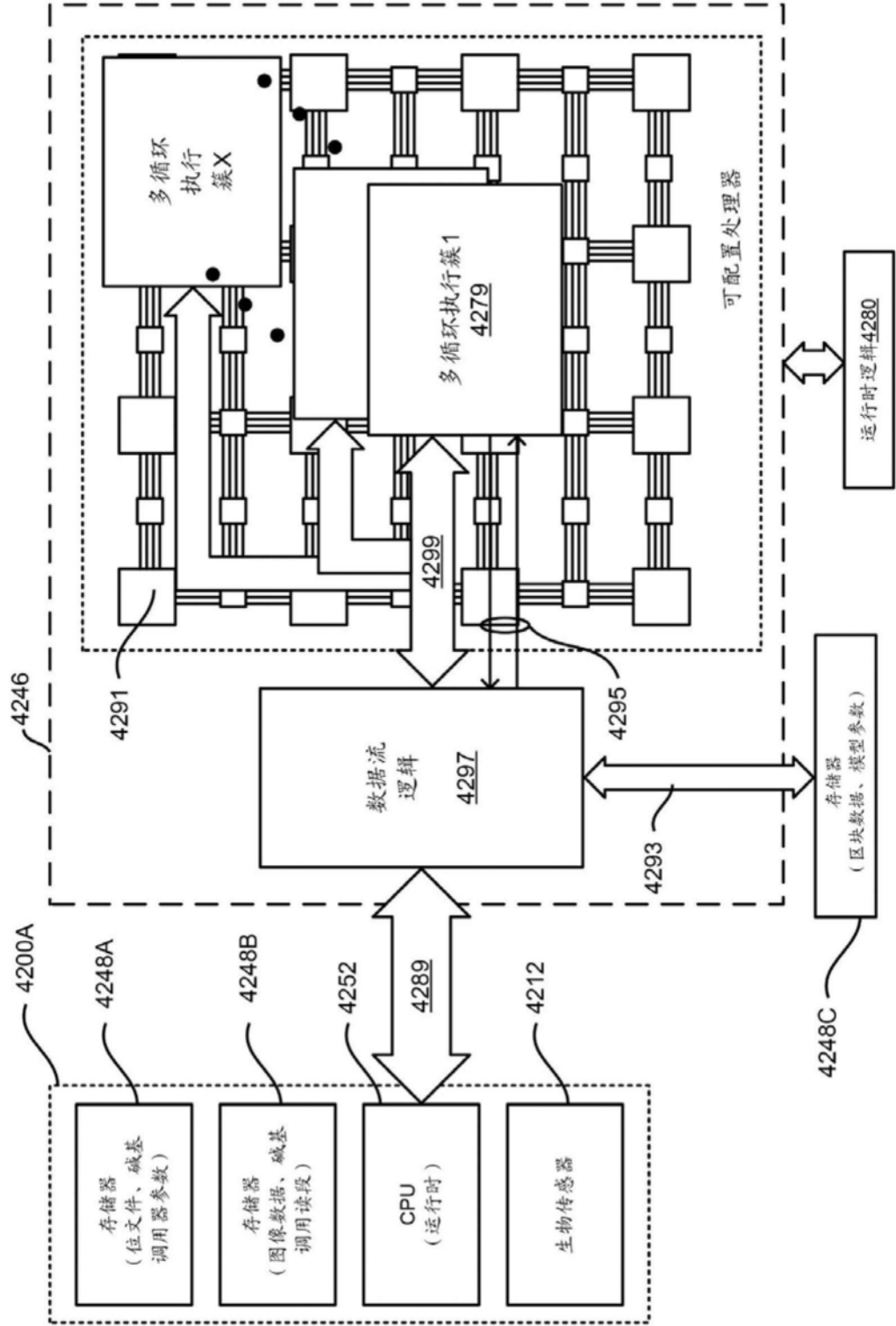


图42C

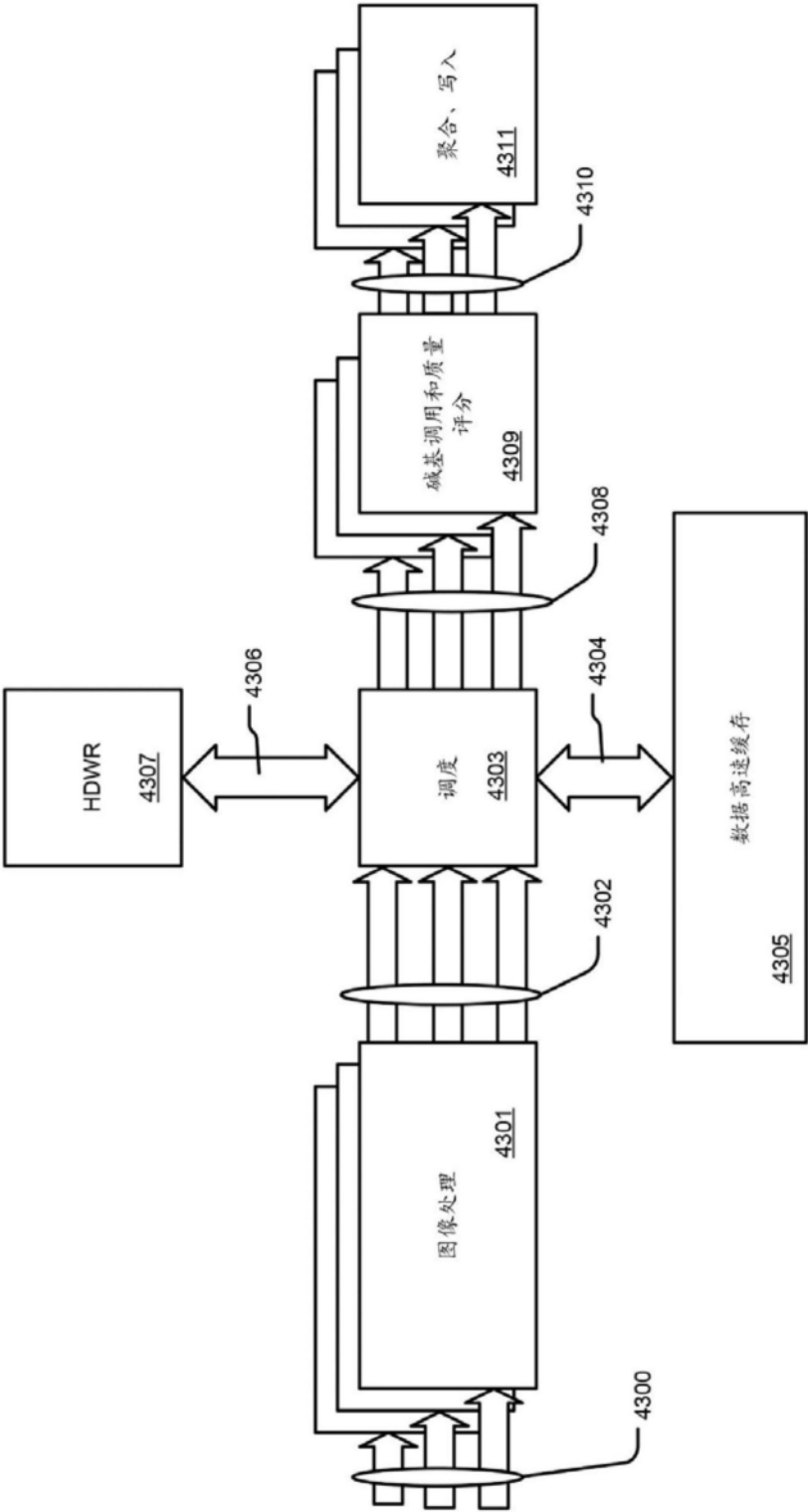


图43A

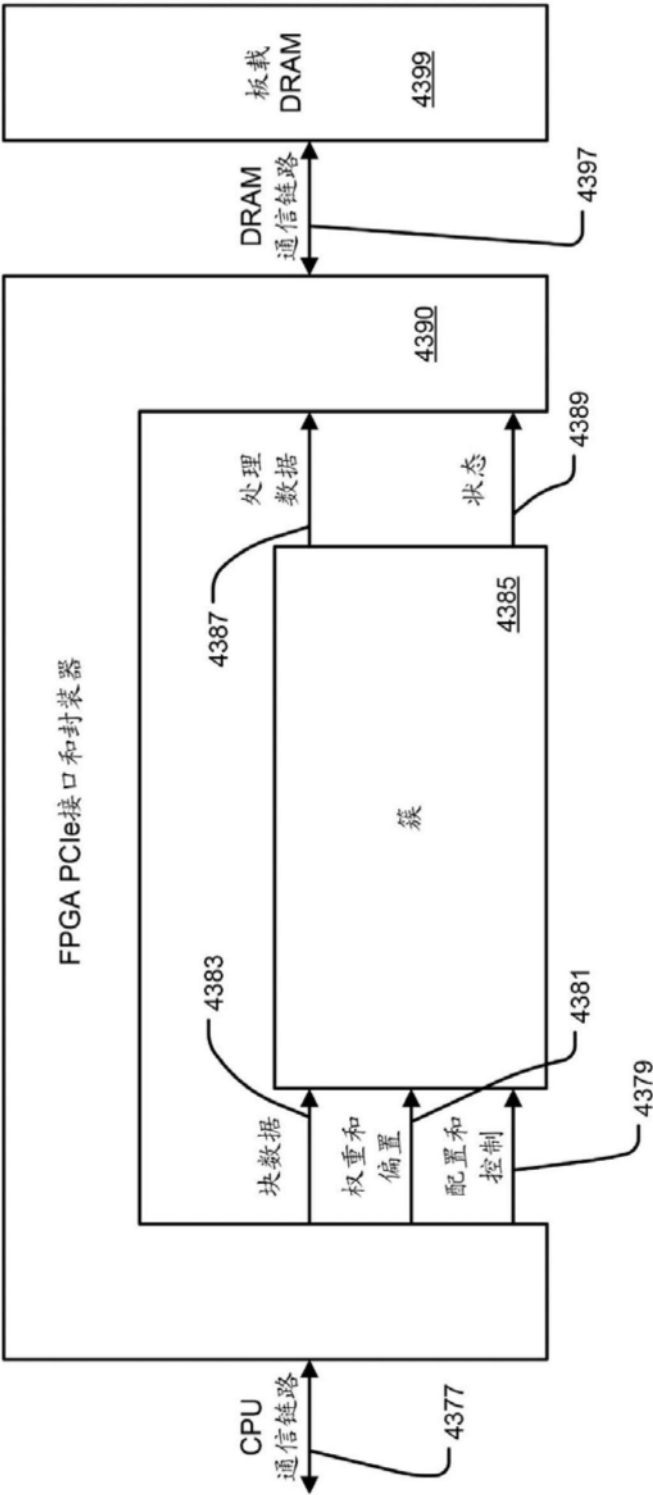


图43B

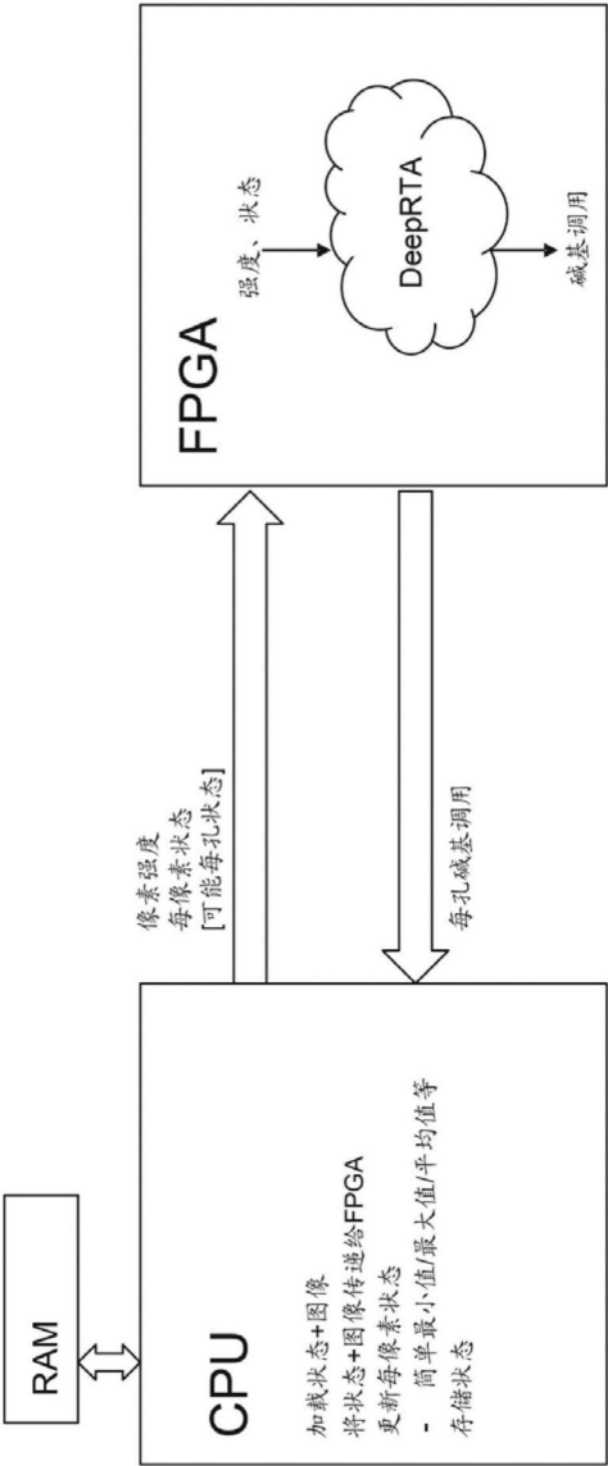


图44

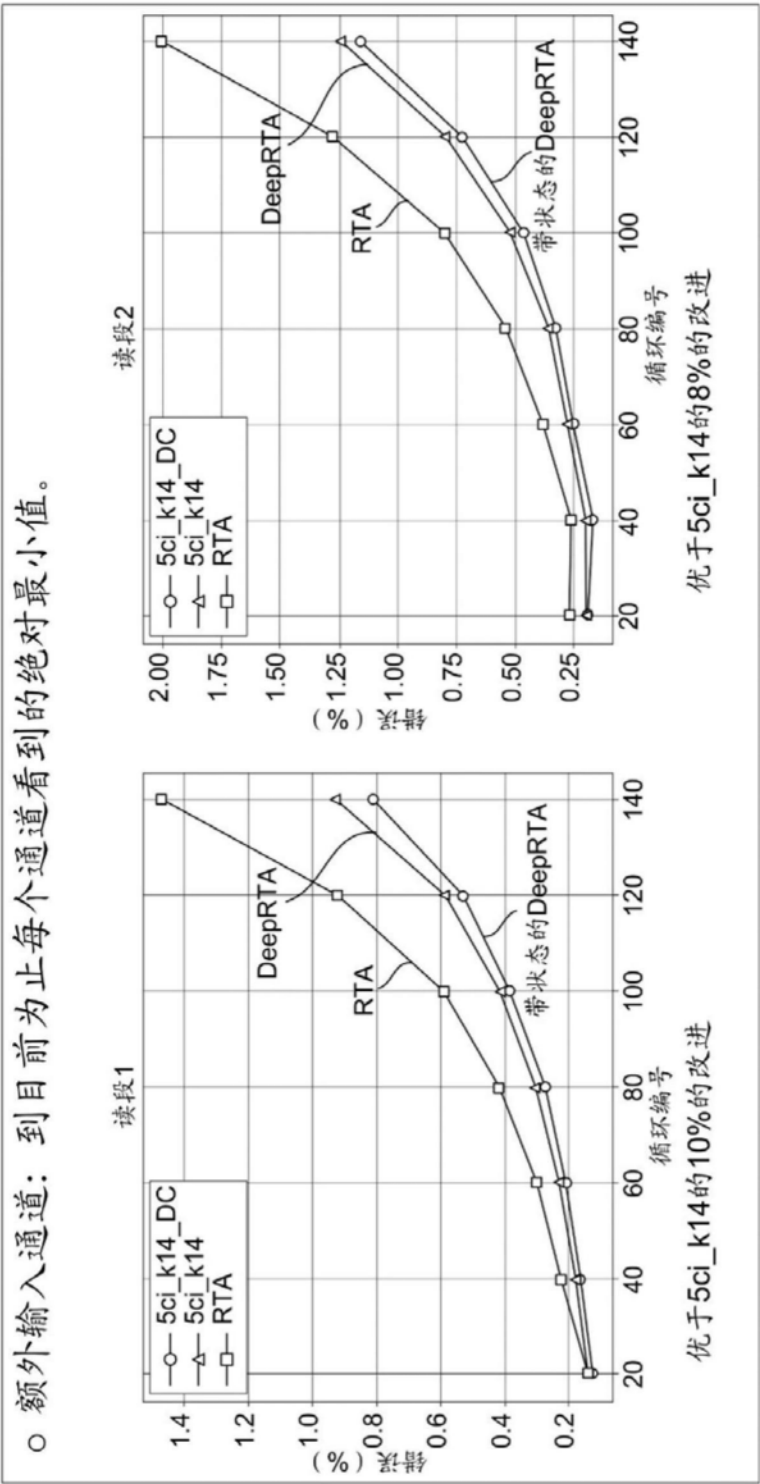


图45

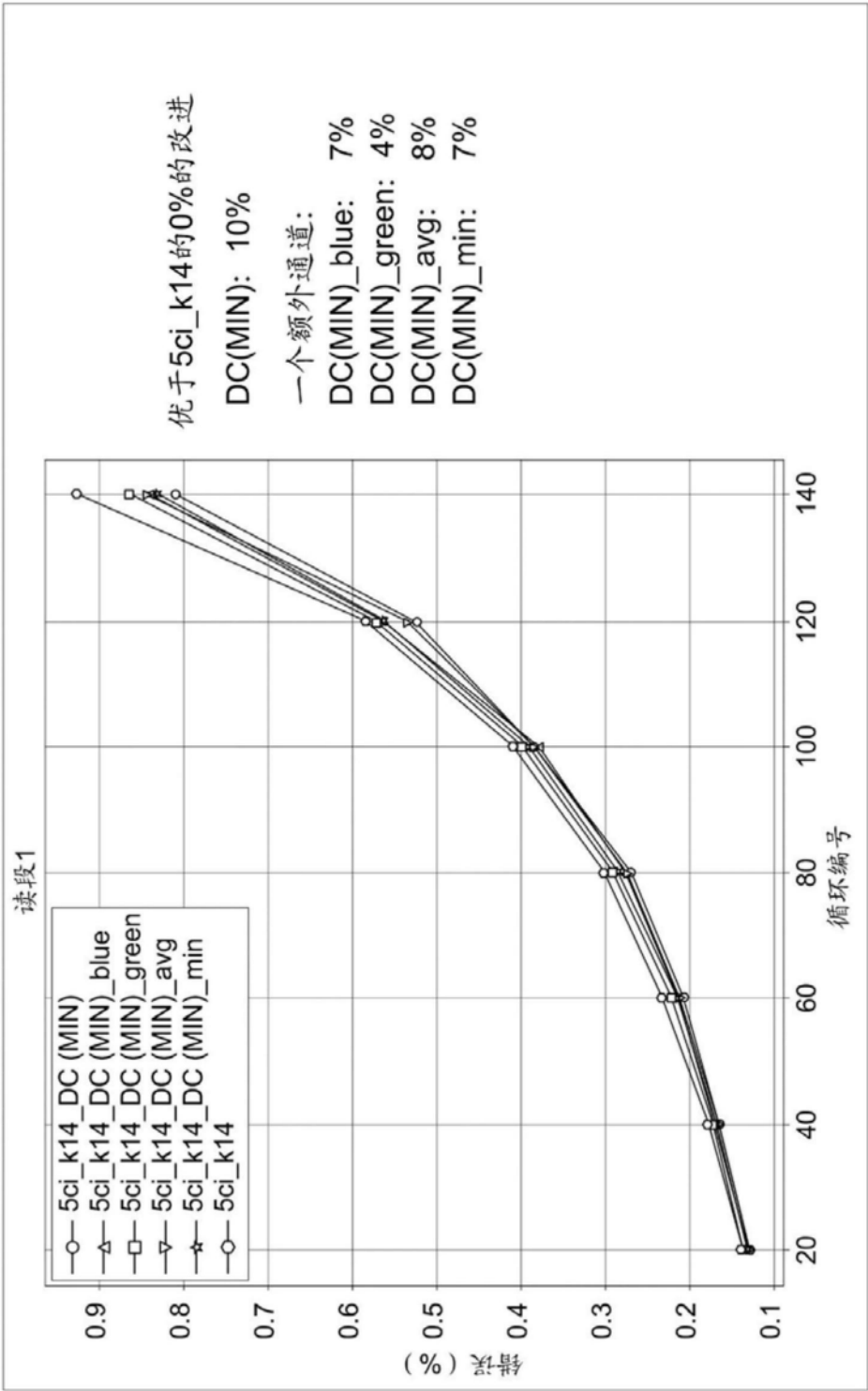


图46

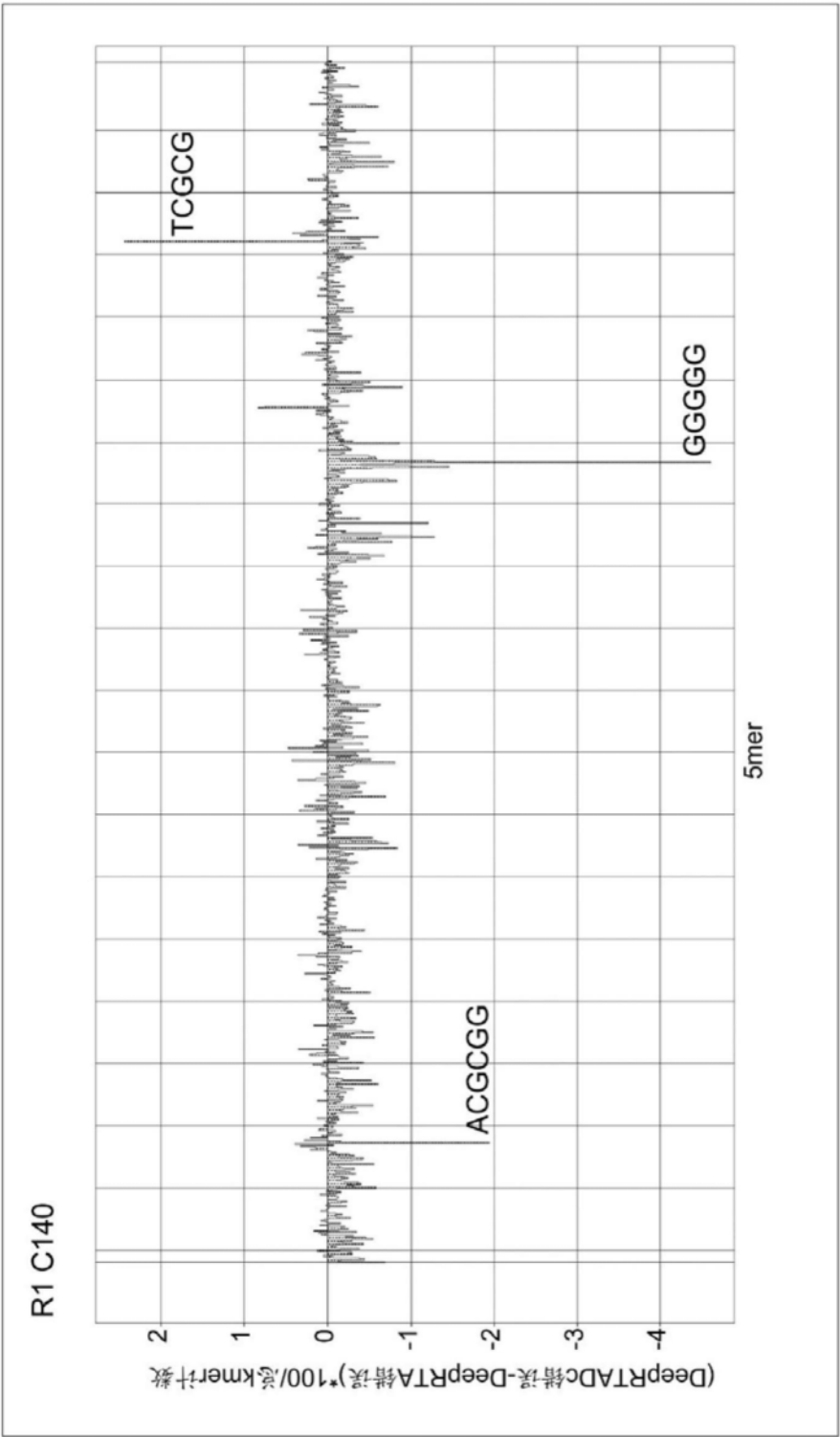
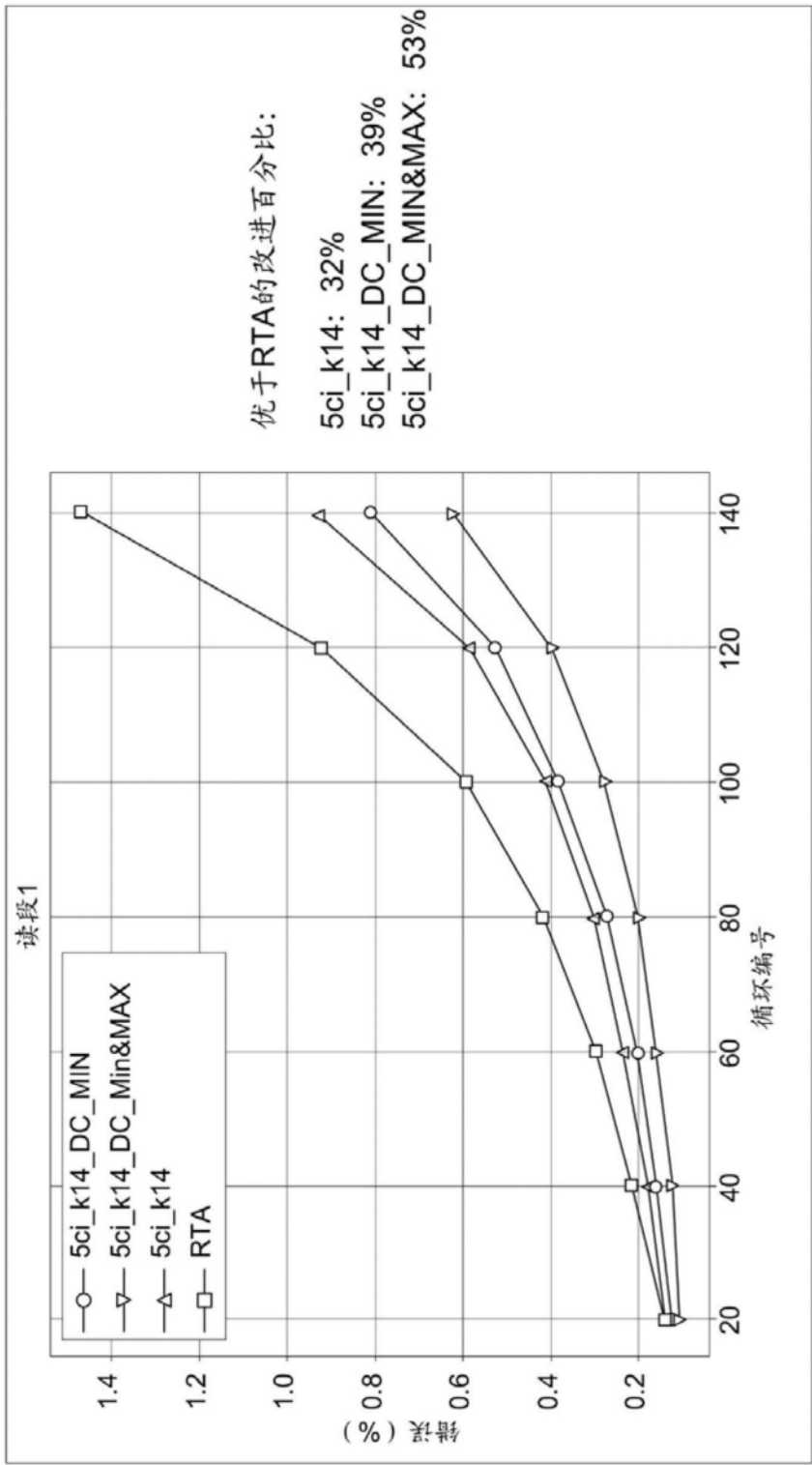


图47



与无状态的DeepRTA系统相比，已展示错误率降低了30%
- “计算”成本可忽略不计
- 硬件兼容

图48

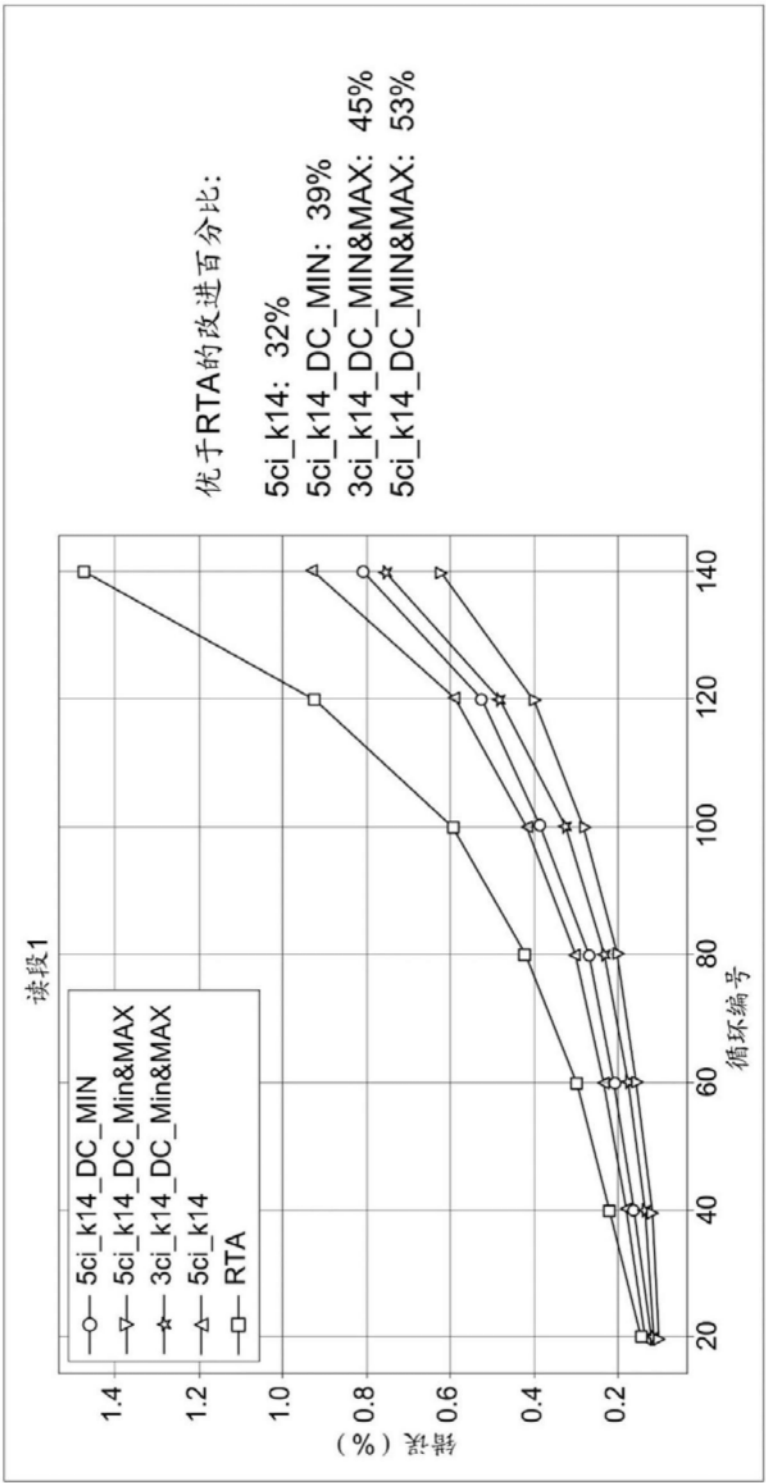


图49

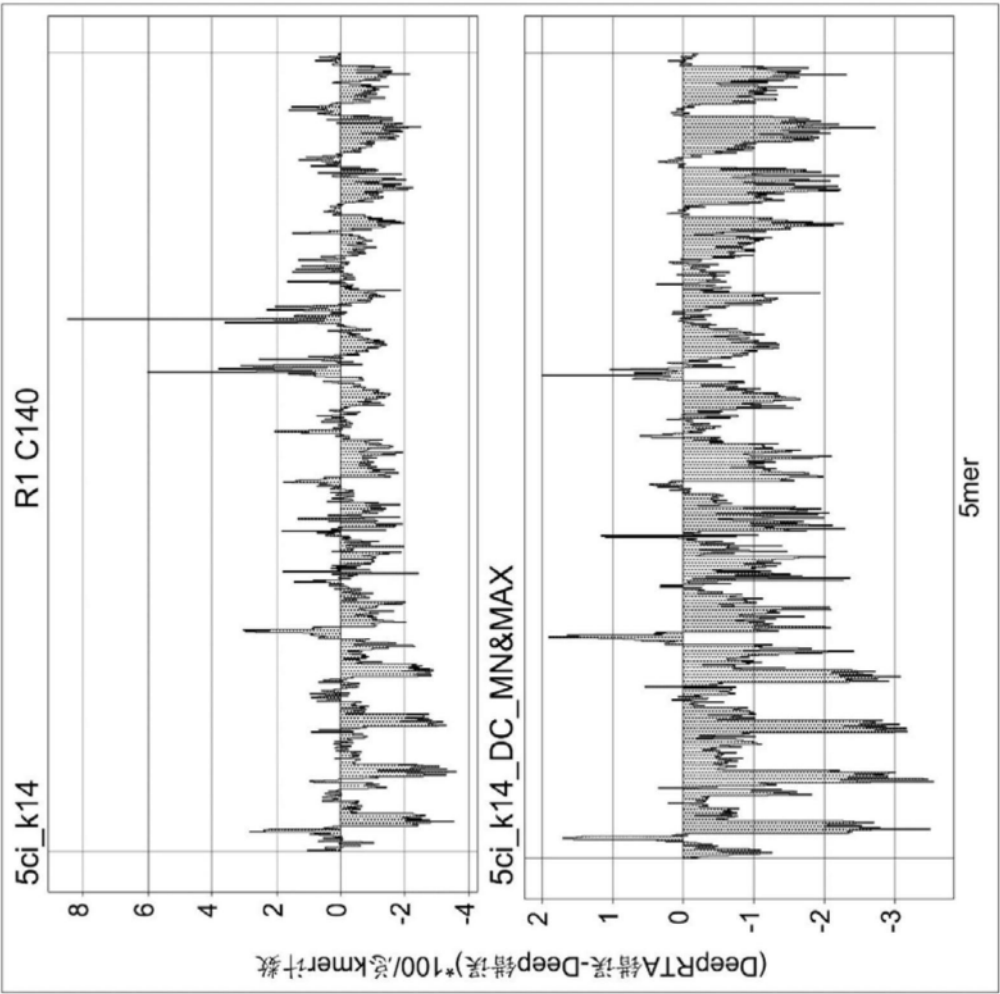


图50

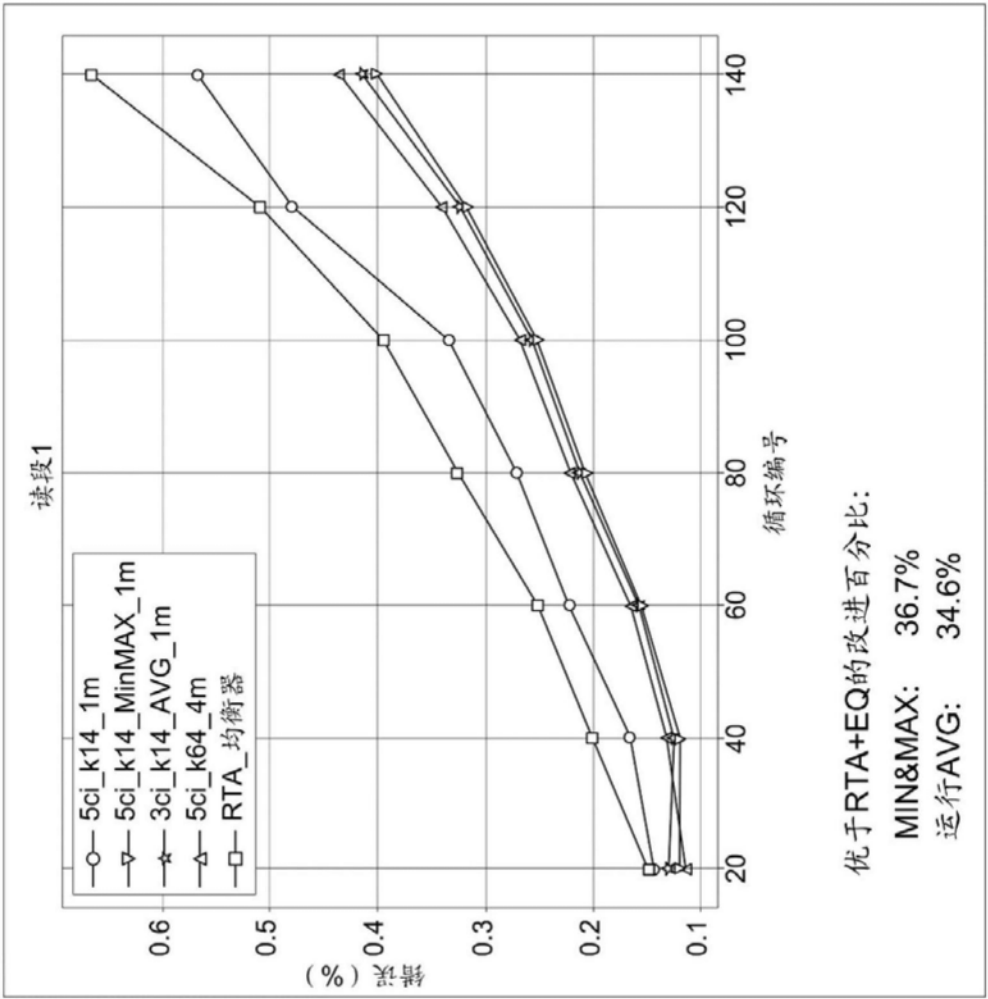


图51

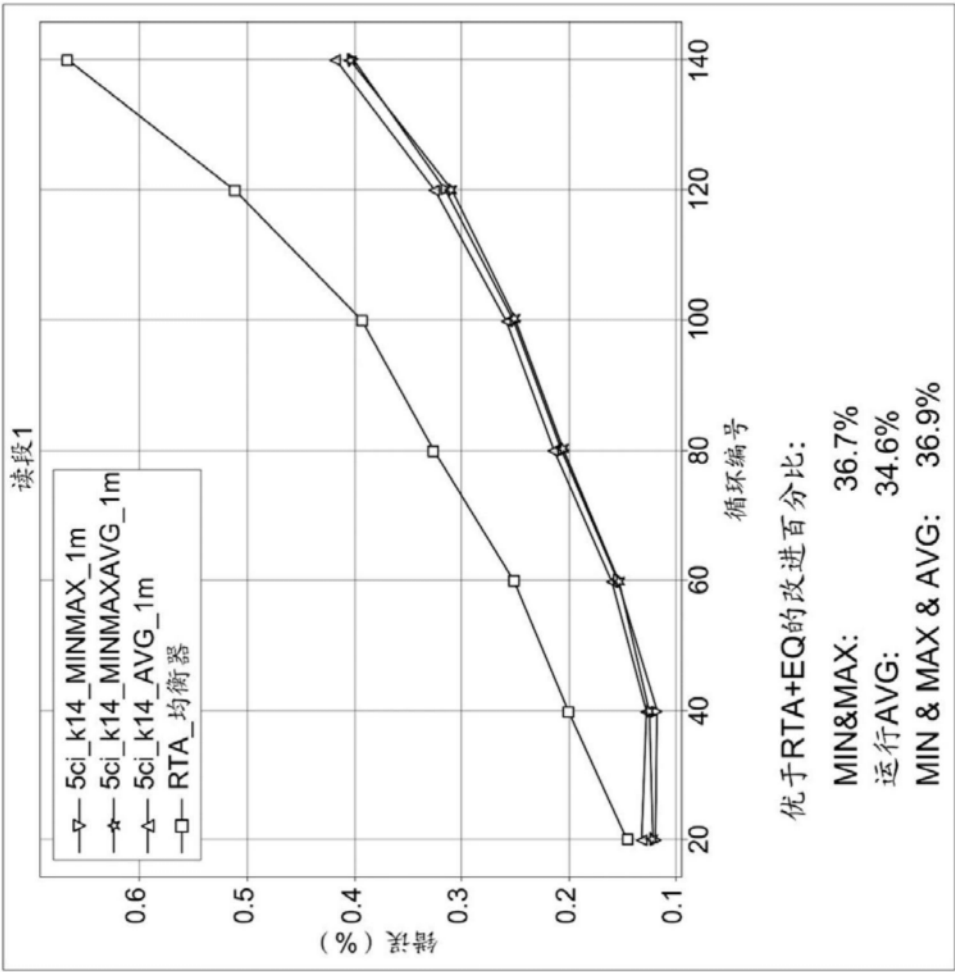


图52

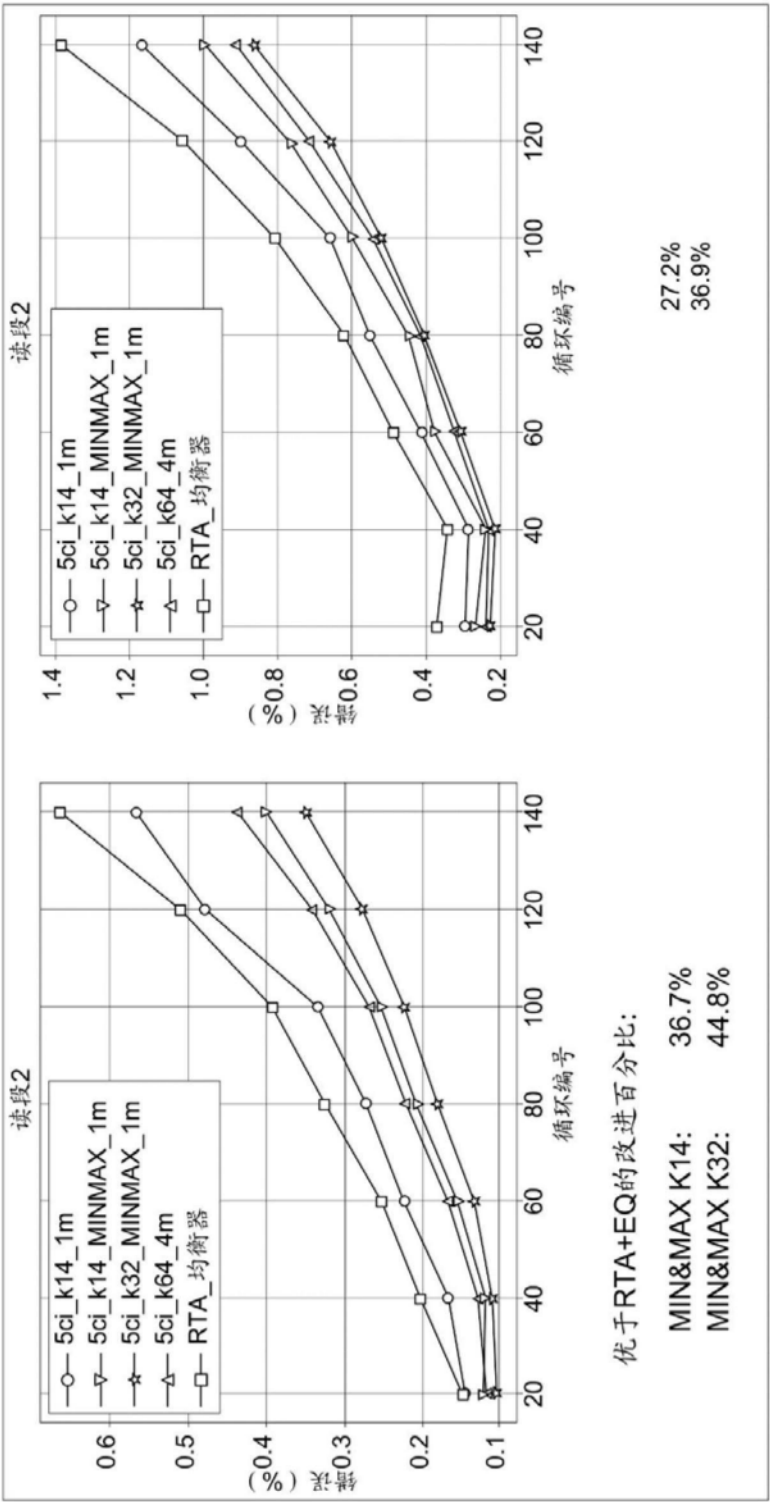


图53

使用每簇信息

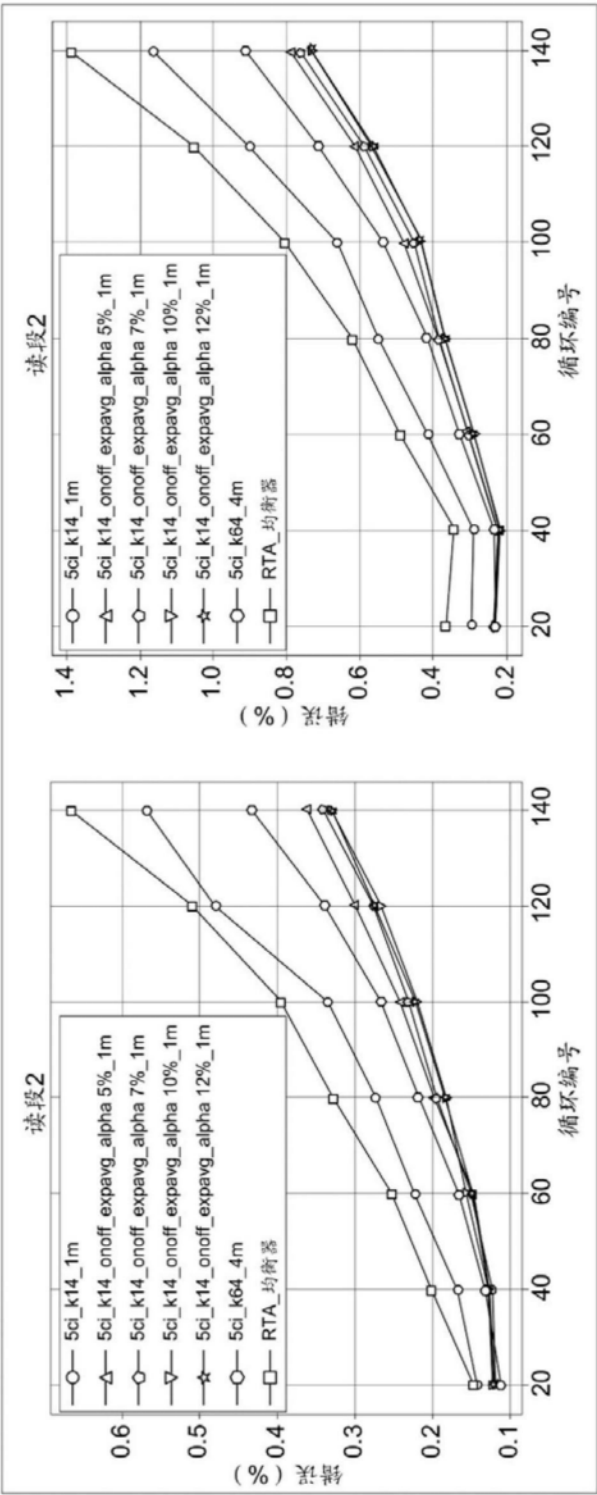


图54

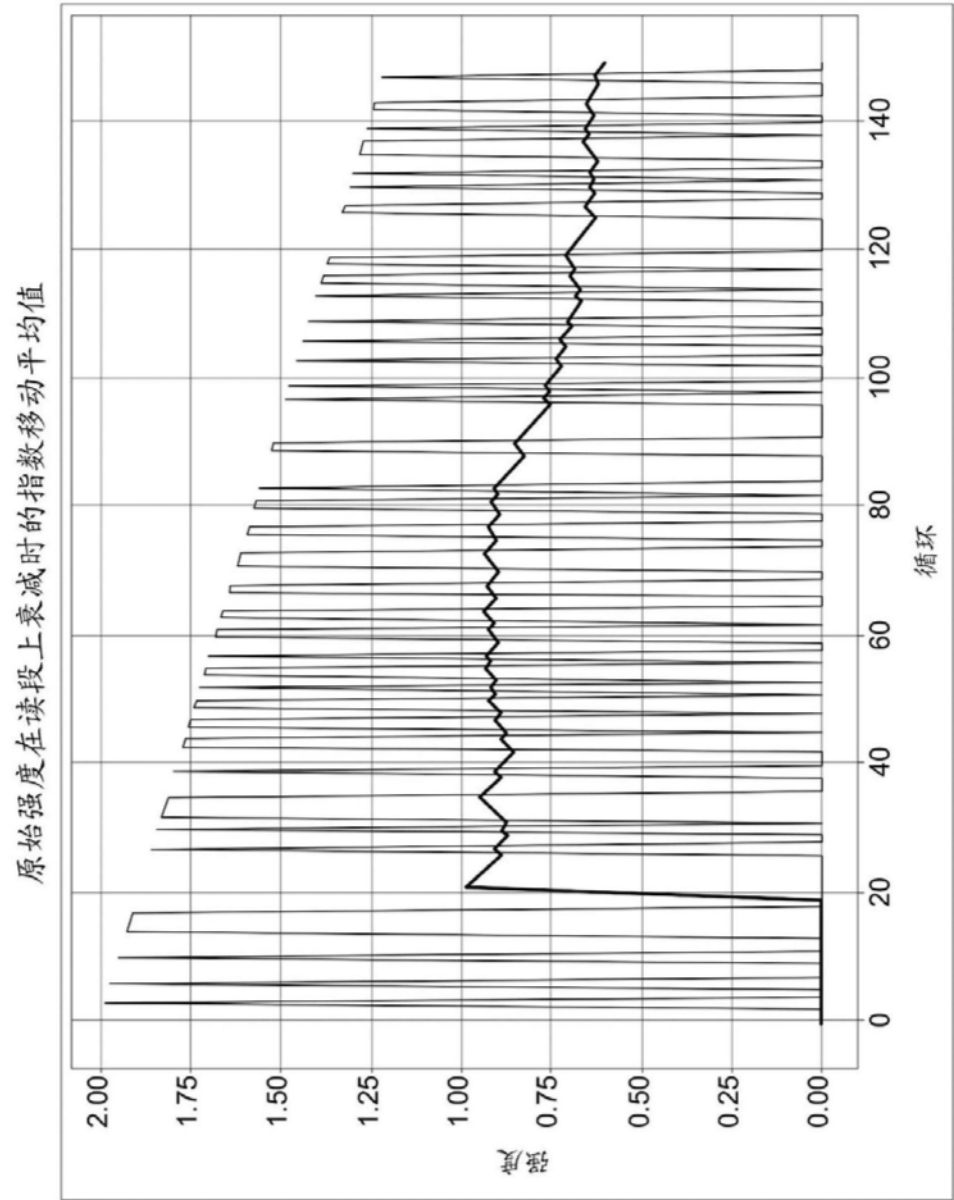


图55

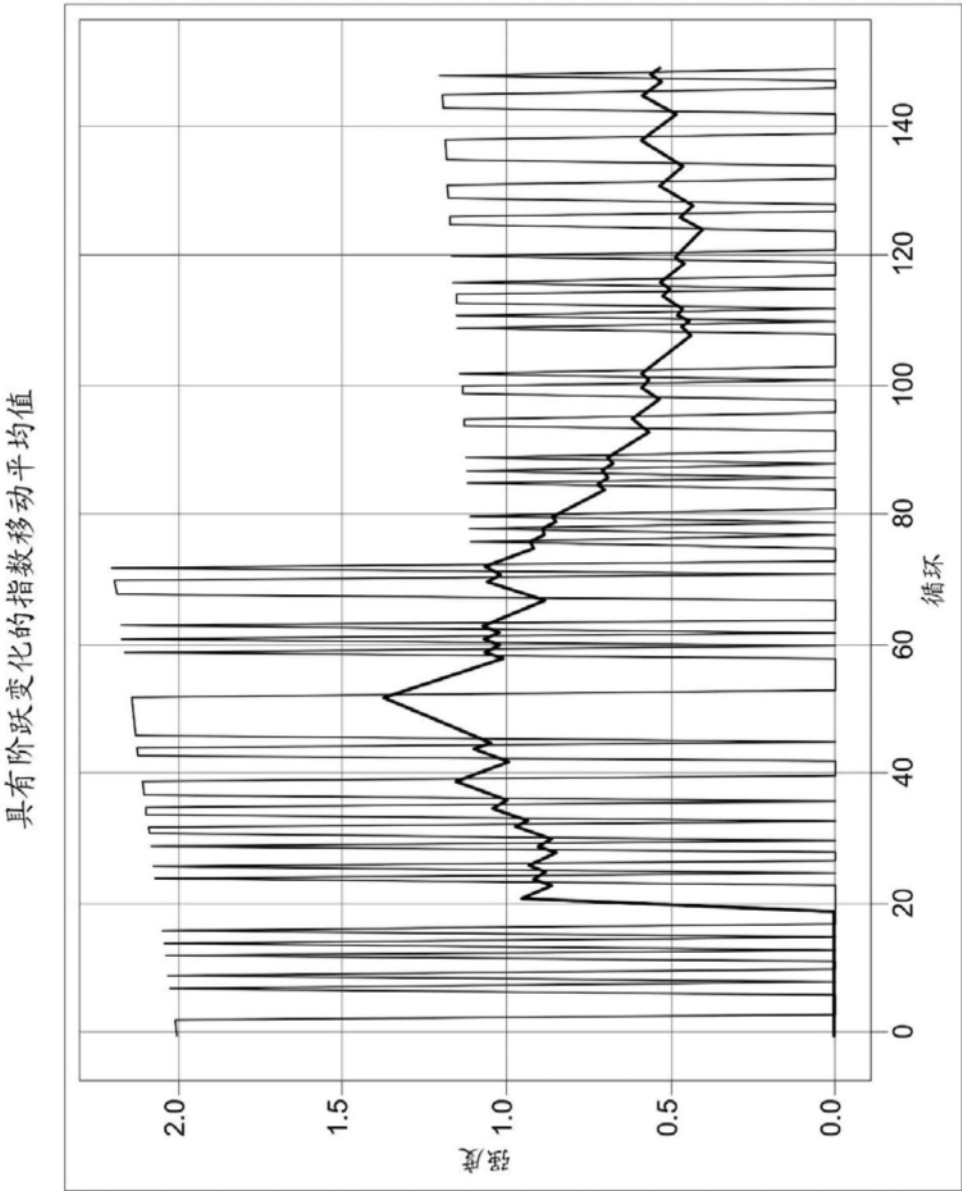


图56

关于25X人类构建的推断结果

	读段 的数量	错误率 (R1、R2、avg)	SNP 重新调用	SNP 精度	SNP F1	Indel 重新调用	Indel 精度	Indel F1
RTA + Eq	605,717,554	0.41,0.72,0.56	0.990142	0.998558	0.994332	0.971573	0.978162	0.974856
DeepRTA (k14_1m	602,022,776	0.36,0.63,0.50	0.990382	0.998508	0.994680	0.976982	0.977608	0.977295
DeepRTA_extrachannels (k14_1m_onoff_expavg	609,209,246	0.31,0.54,0.42	0.990385	0.99849	0.994724	0.977727	0.977748	0.977737

图57