



(12) 发明专利申请

(10) 申请公布号 CN 113166804 A

(43) 申请公布日 2021. 07. 23

(21) 申请号 201980076121.5

(22) 申请日 2019.11.26

(30) 优先权数据

1819378.9 2018.11.28 GB

(85) PCT国际申请进入国家阶段日

2021.05.19

(86) PCT国际申请的申请数据

PCT/GB2019/053334 2019.11.26

(87) PCT国际申请的公布数据

W02020/109773 EN 2020.06.04

(71) 申请人 牛津纳米孔科技公司

地址 英国牛津市

(72) 发明人 蒂莫西·L·马辛厄姆

(74) 专利代理机构 上海君立衡知识产权代理事务
所(特殊普通合伙) 31389

代理人 黄庆

(51) Int.Cl.

G12Q 1/6869(2006.01)

G01N 33/487(2006.01)

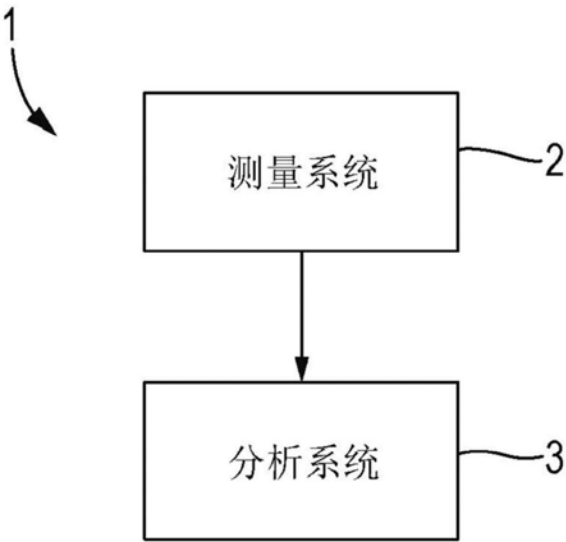
权利要求书4页 说明书36页 附图22页

(54) 发明名称

使用机器学习技术分析纳米孔信号

(57) 摘要

使用机器学习技术分析在包括一系列聚合物单元的聚合物相对于纳米孔易位期间从所述聚合物导出的信号。输出一系列权重分布,每个权重分布包括关于一组标签上的标签之间的转换的权重,所述一组标签表示所述聚合物中的可能类型的聚合物单元。来自所述权重分布的所述一系列聚合物单元的估计值。关于所述转换的权重的使用提高了所述技术的准确度。



1. 一种分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的方法, 所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,

所述方法包括:

使用输出一系列权重分布的机器学习技术来分析所述信号, 每个权重分布包括关于一组标签上的标签之间的转换的权重, 所述一组标签包含表示所述可能类型的聚合物单元的标签; 以及

从所述权重分布导出所述一系列聚合物单元的估计值。

2. 根据权利要求1所述的方法, 其中不允许标签之间的至少一个转换并且允许其它转换, 所述权重分布各自包括关于被允许的转换的权重。

3. 根据权利要求2所述的方法, 其中所述权重分布各自包括关于不被允许的转换的零权重。

4. 根据权利要求2或3所述的方法, 其中导出所述一系列聚合物单元的估计值的步骤将表示是否允许或不允许标签之间的转换的转换矩阵考虑在内。

5. 根据权利要求2到4中任一项所述的方法, 其中

所述一组标签包含关于每种类型的聚合物单元的第一标签和第二标签, 所述第一标签表示所述类型的聚合物单元的实例的开始, 并且所述第二标签表示所述类型的聚合物单元的所述实例中的停留, 其中

允许任何其它类型的聚合物单元从每个第一标签转换到所述第一标签,

允许相同类型的聚合物单元从每个第一标签转换到所述第一标签,

允许相同类型的聚合物单元从每个第一标签转换到所述第二标签,

不允许任何其它类型的聚合物单元从每个第一标签转换到所述第二标签,

允许相同类型的聚合物单元从每个第二标签转换到所述第一标签或允许任何其它类型的聚合物单元从每个第二标签转换到所述第一标签, 并且

允许相同类型的聚合物单元从每个第二标签转换到所述第二标签, 并且

不允许任何其它类型的聚合物单元从每个第二标签转换到所述第二标签。

6. 根据权利要求2到5中任一项所述的方法, 其中所述一组可能类型的聚合物单元包含总是以聚合物单元的已知序列出现的类型的聚合物单元, 允许与所述已知序列一致的转换并且不允许与所述已知序列相反的转换。

7. 根据前述权利要求中任一项所述的方法, 其中所述一系列聚合物单元中的相同类型的聚合物单元的连续实例以编码形式表示。

8. 根据权利要求7所述的方法, 其中所述标签包含关于每种类型的聚合物单元的多个标签, 其中关于每种类型的聚合物单元的所述多个标签表示所述一系列聚合物单元中所述类型的聚合物单元的连续实例。

9. 根据权利要求8所述的方法, 其中用于每种类型的聚合物单元的所述多个标签具有预先确定的循环顺序, 由此通过所述预先确定的循环顺序允许标签之间的一些转换并且通过所述预先确定的循环顺序不允许之间的其它转换, 所述权重分布各自包含关于通过所述预先确定的循环顺序允许的转换的权重。

10. 根据权利要求8或9所述的方法, 其中用于每种类型的聚合物单元的所述多个标签是用于每种类型的聚合物单元的两个标签。

11. 根据权利要求7所述的方法, 其中所述一系列聚合物单元中的相同类型的聚合物单元的连续实例以游程长度编码形式表示。

12. 根据权利要求11所述的方法, 其中所述标签包含关于每种类型的聚合物单元的不同游程长度的多个标签。

13. 根据权利要求11所述的方法, 其中

所述标签包含关于每种类型的聚合物单元的标签, 并且

对于每种类型的聚合物单元, 所述权重分布包括在相同类型的聚合物单元的连续实例的可能长度上的另外的权重。

14. 根据权利要求13所述的方法, 其中对于每种类型的聚合物单元, 所述另外的权重包括在相同类型的聚合物单元的连续实例的一组可能长度上的权重的分类分布。

15. 根据权利要求13所述的方法, 其中对于每种类型的聚合物单元, 所述另外的权重包括在相同类型的聚合物单元的连续实例的可能长度上的参数化分布的参数。

16. 根据权利要求13到15中任一项所述的方法, 其中 (a) 对于给定聚合物单元的类型和先前聚合物单元的类型的可能对、(b) 对于给定聚合物单元的类型和后续聚合物单元的类型的可能对, 或 (c) 对于给定聚合物单元的类型、先前聚合物单元的类型和后续聚合物单元的类型的可能三联体, 所述权重分布包括在相同类型的聚合物单元的连续实例的可能长度上的另外的权重。

17. 根据前述权利要求中任一项所述的方法, 其中所述可能类型的聚合物单元包含具有未修饰形式和经修饰形式的类型的聚合物单元。

18. 根据权利要求17所述的方法, 其中所述一组标签包含关于具有未修饰形式和经修饰形式的所述类型的聚合物单元的标签。

19. 根据权利要求18所述的方法, 其中对于具有所述未修饰形式和所述经修饰形式的所述类型的聚合物单元中的每个聚合物单元的所述未修饰形式和所述经修饰形式, 每个权重分布包括另外的权重。

20. 根据前述权利要求中任一项所述的方法, 其中所述一组标签包含至少一个表示每种类型的聚合物单元的标签。

21. 根据前述权利要求中任一项所述的方法, 其中所述一组标签进一步包含至少一个表示所述一系列聚合物单元中的空白和/或停留的标签。

22. 根据前述权利要求中任一项所述的方法, 其中所述机器学习技术是包括至少一个递归层的神经网络。

23. 根据权利要求22所述的方法, 其中所述至少一个递归层是双向递归层。

24. 根据权利要求22或23所述的方法, 其中所述神经网络相对于通过所述一系列权重分布的所有标签路径应用所述权重分布的全局归一化。

25. 根据权利要求22到24中任一项所述的方法, 其中所述神经网络包含至少一个卷积层, 所述至少一个卷积层布置在所述至少一个递归层之前并且执行所述信号的加窗部分的卷积。

26. 根据前述权利要求中任一项所述的方法, 其中所述权重表示后验概率。

27. 根据前述权利要求中任一项所述的方法, 其中从所述权重分布导出所述一系列聚合物单元的估计值的步骤使用链结式时间分类执行。

28. 根据前述权利要求中任一项所述的方法, 其中从所述权重分布导出聚合物单元的估计值的步骤包括导出关于相应的权重分布的标签和对导出的标签进行游程长度压缩。

29. 根据前述权利要求中任一项所述的方法, 其中从所述权重分布导出所述一系列聚合物单元的估计值的步骤包括估计按所述权重分布计通过所述一系列权重分布的最可能标签路径并且从被估计为最可能的所述标签路径导出所述一系列聚合物单元的所述估计值。

30. 根据权利要求1到28中任一项所述的方法, 其中从所述权重分布导出所述一系列聚合物单元的估计值的步骤包括估计关于每个权重分布最可能的所述标签、将通过所述一系列权重分布的前向和后向标签路径考虑在内并且从被估计为最可能的所述标签导出所述一系列聚合物单元的所述估计值。

31. 根据前述权利要求中任一项所述的方法, 其中所述纳米孔是蛋白孔。

32. 根据前述权利要求中任一项所述的方法, 其中所述聚合物是多核苷酸, 并且所述聚合物单元是核苷酸。

33. 根据前述权利要求中任一项所述的方法, 其中所述信号从以下性质中的一个或多个的测量结果导出: 离子电流、阻抗、隧穿性质、场效应晶体管电压和光学性质。

34. 根据前述权利要求中任一项所述的方法, 所述方法在计算机设备中执行。

35. 根据前述权利要求中任一项所述的方法, 其进一步包括在所述聚合物相对于纳米孔易位期间从所述聚合物导出所述信号。

36. 一种用于分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的分析设备, 所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,

所述分析设备包括:

机器学习单元, 所述机器学习单元被布置成对所述信号执行机器学习技术并且输出一系列权重分布, 每个权重分布包括关于一组标签上的标签之间的转换的权重, 所述一组标签包含表示所述可能类型的聚合物单元的标签; 以及

估计单元, 所述估计单元被布置成从所述权重分布导出所述一系列聚合物单元的估计值。

37. 一种纳米孔测量和分析系统, 其包括:

测量装置, 所述测量装置被布置成在聚合物相对于纳米孔易位期间从所述聚合物导出信号; 以及

根据权利要求36所述的分析设备。

38. 一种分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的方法, 所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,

所述方法包括:

使用输出一系列权重分布的机器学习技术来分析所述信号, 每个权重分布包括关于一组标签上的标签的权重, 所述一组标签包含表示所述可能类型的聚合物单元的标签; 以及从所述权重分布导出所述一系列聚合物单元的估计值,

其中导出所述一系列聚合物单元的估计值的步骤将表示是否允许或不允许标签之间的转换的转换矩阵考虑在内, 标签之间的至少一个转换被表示为不允许并且其它转换被表示为允许。

39. 一种用于分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的分析设备, 所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,

所述分析设备包括:

机器学习单元, 所述机器学习单元被布置成对所述信号执行机器技术并且输出一系列权重分布, 每个权重分布包括关于一组标签上的标签的权重, 所述一组标签包含表示所述可能类型的聚合物单元的标签; 以及

估计单元, 所述估计单元被布置成从所述权重分布导出所述一系列聚合物单元的估计值,

其中所述估计单元被布置成将表示是否允许或不允许标签之间的转换的转换矩阵考虑在内, 标签之间的至少一个转换被表示为不允许并且其它转换被表示为允许。

40. 一种分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的方法, 所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,

所述方法包括:

使用输出一系列权重分布的机器学习技术来分析所述信号, 每个权重分布包括关于一组标签上的标签的权重, 所述一组标签包含表示所述可能类型的聚合物单元的标签, 其中所述一系列聚合物单元中的相同类型的聚合物单元的连续实例以游程长度编码形式表示; 以及

从所述权重分布导出所述一系列聚合物单元的估计值。

41. 一种用于分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的分析设备, 所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,

所述分析设备包括:

机器学习单元, 所述机器学习单元被布置成对所述信号执行机器技术并且输出一系列权重分布, 每个权重分布包括关于一组标签上的标签的权重, 所述一组标签包含表示所述可能类型的聚合物单元的标签, 其中所述一系列聚合物单元中的相同类型的聚合物单元的连续实例以游程长度编码形式表示; 以及

估计单元, 所述估计单元被布置成从所述权重分布导出所述一系列聚合物单元的估计值。

使用机器学习技术分析纳米孔信号

技术领域

[0001] 本发明涉及对在聚合物(例如但不限于多核苷酸)相对于纳米孔易位期间从聚合物导出的信号的分析。

背景技术

[0002] 用于使用纳米孔估计聚合物中的聚合物单元的靶序列的测量系统是已知的,其中聚合物相对于纳米孔易位。系统的一些性质取决于纳米孔中的聚合物单元,并且进行所述性质的测量。性质取决于相对于纳米孔易位的聚合物单元的身份,并且因此,随时间变化的信号允许估计聚合物单元的序列。与孔的尺寸相比,每个聚合物单元可以非常小,从而允许多个聚合物单元在给定时间段内影响信号。由于聚合物链与纳米孔的相互作用、如缠绕或堆叠等链内性质或聚合物单元与用于控制其易位的任何系统之间的相互作用,也可以存在更长远的影响。信号形成必须被解码以估计潜在的聚合物单元的读段。

[0003] 这种纳米孔测量系统可以提供表示范围为数百到数十万(并且可能更多)个核苷酸的多核苷酸的连续长读段的信号。使用纳米孔的这种类型的测量系统具有相当大的前景,特别是在对如DNA或RNA等多核苷酸进行测序的领域中,并且已经成为最近发展的主题。

[0004] 然而,对聚合物单元的估计的准确度受到极其敏感的测量系统的限制。实际上,具有高准确度的估计需要机器学习技术,并且本发明涉及改善这种分析以增加估计准确度。

[0005] 早期分析技术使用了隐马尔可夫模型(Hidden Markov Model, HMM),所述模型显式地对包括连续聚合物单元组的可能的k聚体进行建模。最近已经开发了使用递归神经网络(RNN)的技术。RNN的使用可以通过将远程信息考虑在内来提高准确度。随着读取速度的提高,RNN尤其有用,结果是像HMM等显式信号建模方法所基于的假设不那么有效。举例来说,Teng等人,“手性子:使用深度学习将纳米孔原始信号直接翻译为核苷酸序列(Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning)”,《大数据科学(Gigascience)》,2018年5月1日;7(5)[参考1]公开了使用向其输入原始信号的RNN对聚合物核苷酸进行测序的方法。RNN输出一系列后验概率分布,所述一系列后验概率分布包括关于表示四种可能类型的碱基以及空白的标签的后验概率。通过根据后验概率估计最可能的聚合物单元由链结式时间分类解码器对这些后验概率分布进行解码,以导出一系列聚合物单元的估计值。

发明内容

[0006] 根据本发明的第一方面,提供了一种分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的方法,所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,所述方法包括:使用输出一系列权重分布的机器学习技术来分析所述信号,每个权重分布包括关于一组标签上的标签之间的转换的权重,所述一组标签包含表示所述可能类型的聚合物单元的标签;以及从所述权重分布导出所述一系列聚合物单元的估计值。

[0007] 所述一组标签可以包含表示空白和/或停留的标签。换句话说,可以说所述组表示

可能类型的聚合物单元。

[0008] 转换可以在一个标签与另一个标签之间。转换可以在连续的标签之间。

[0009] 因此,所述方法提供的权重是指表示可能类型的聚合物单元,而不是表示包括k个聚合物单元的k聚体的标签。然而,所述方法导出关于标签之间的转换的权重,而不是关于标签本身的权重。此类方法提供了优于比较方法的优点,所述比较方法导出关于一组标签上的标签的一系列权重,所述一组标签包含表示可能类型的聚合物单元的标签。通过提供关于所述一组标签上的标签之间的转换的权重,提供了另外的信息,所述另外的信息允许以更准确的方式估计一系列聚合物单元。这是因为权重提供有关标签可能路径的信息,而关于标签的权重却不提供。

[0010] 例如,存在以下情况:通过关于标签的权重预测的特定位置的标签不正确,而考虑通过所述位置的标签路径可能会预测出正确的不同标签。以此方式,另外的信息被馈送到估计值中,从而提高了准确度。

[0011] 举例来说,此技术允许更好地估计重复序列的区域,例如均聚物,包含重复一个或多个聚合物单元的短序列的区域。

[0012] 优选地,不允许标签之间的至少一个转换并且允许其它转换,所述权重分布各自包括关于被允许的转换的权重。在所述情况下,权重分布可以各自包括关于不被允许的转换的零权重,或者导出一系列聚合物单元的估计值的步骤可以考虑表示是否允许或不允许标签之间的转换的转换矩阵。

[0013] 在一种类型的表示中,所述一组标签可以包含关于每种类型的聚合物单元的第一标签和第二标签,所述第一标签表示所述类型的聚合物单元的实例的开始,并且所述第二标签表示所述类型的聚合物单元的所述实例中的停留,其中允许任何其它类型的聚合物单元从每个第一标签转换到所述第一标签,允许相同类型的聚合物单元从每个第一标签转换到所述第二标签,不允许任何其它类型的聚合物单元从每个第一标签转换到所述第二标签,允许相同类型的聚合物单元从每个第二标签转换到所述第一标签或允许任何其它类型的聚合物单元从每个第二标签转换到所述第二标签,并且不允许任何其它类型的聚合物单元从每个第二标签转换到所述第二标签。

[0014] “停留”表示其中所述方法确定标签不发生变化的情况,可以将其视为对应于聚合物单元的相同实例的两个权重分布。

[0015] 所述一组可能类型的聚合物单元可以包含总是以聚合物单元的已知序列出现的类型的聚合物单元,允许与所述已知序列一致的转换并且不允许与所述已知序列相反的转换。

[0016] 所述一系列聚合物单元中的相同类型的聚合物单元的标签连续实例可以编码形式表示。

[0017] 所述标签可以包含关于每种类型的聚合物单元的多个标签,例如两个标签,其中关于每种类型的聚合物单元的所述多个标签表示所述一系列聚合物单元中所述类型的聚合物单元的连续实例。

[0018] 用于每种类型的聚合物单元的所述多个标签可以具有预先确定的循环顺序,由此通过所述预先确定的循环顺序允许标签之间的一些转换并且通过所述预先确定的循环顺

序不允许之间的其它转换,所述权重分布包含关于通过所述预先确定的循环顺序允许的转换的权重。

[0019] 所述一系列聚合物单元中的相同类型的聚合物单元的连续实例以游程长度编码形式表示。

[0020] 所述标签可以包含关于每种类型的聚合物单元的不同游程长度的标签。

[0021] 所述标签可以包含关于每种类型的聚合物单元的标签,并且对于每种类型的聚合物单元,所述权重分布可以包括在相同类型的聚合物单元的连续实例的可能长度上的另外的权重。

[0022] 对于每种类型的聚合物单元,所述另外的权重可以包括在相同类型的聚合物单元的连续实例的一组可能长度上的权重的分类分布。

[0023] 对于每种类型的聚合物单元,所述另外的权重可以包括在相同类型的聚合物单元的连续实例的连续实例的可能长度上的参数化分布的参数。

[0024] 如果可能类型的聚合物单元包含具有未修饰形式和修饰形式的类型的聚合物单元,则所述一组标签可以包含表示具有未修饰形式和经修饰形式的所述类型的聚合物单元的标签,并且每个权重分布可以包括具有未修饰形式和经修饰形式的聚合物单元的所述至少一种类型中的每种的所述未修饰形式和经修饰形式的另外的权重。聚合物单元的未修饰形式可以描述为典型聚合物单元,并且聚合物单元的修饰形式可以描述为非典型聚合物单元。修饰(或非典型)聚合物单元通常影响与对应未修饰(典型)聚合物单元不同的信号。

[0025] 在一些实施例中,可以制备和随后分析包括一个或多个非典型聚合物单元的聚合物,如在2019年9月4日提交的国际专利申请号PCT/GB2019/052456中所详细描述,对其进行引用并且通过引用并入本文。在一个实例中,可以不确定性的方式,例如通过化学转化或通过酶促转化,将一定比例的典型聚合物单元(例如,氨基酸)转化为对应的非典型聚合物单元(例如,氨基酸)。在所述情况下,当导出一系列聚合物单元的估计值(“调用”)时,可以将非典型碱基估计(“调用”)为对应的典型碱基。以此方式,通过在分析中将非典型聚合物单元识别为典型聚合物单元,初始转换可以提供一种提供具有更多信息的信号的方式,例如,结果是信号分析中出现的任何错误将是非系统性的,从而带来估计准确度的提高。

[0026] 一组标签可以包含关于每种类型的聚合物单元的至少一个标签和关于一系列聚合物单元中的空白的至少一个标签。

[0027] 机器学习技术可以是包括至少一个递归层的神经网络,所述递归层可以是双向递归层。

[0028] 所述神经网络可以相对于通过所述一系列权重分布的所有路径应用所述权重分布的全局归一化。

[0029] 所述神经网络可以包含多个卷积层,所述多个卷积层布置在所述递归层之前并且执行信号的加窗部分的卷积。

[0030] 所述权重可以表示后验概率。

[0031] 从所述权重分布导出所述一系列聚合物单元的估计值的步骤可以使用链结式时间分类执行。

[0032] 从所述权重分布导出所述一系列聚合物单元的估计值的所述步骤可以包括导出关于每个权重分布的标签和对导出的标签进行游程长度压缩。

[0033] 从所述权重分布导出所述一系列聚合物单元的估计值的步骤可以包括估计按所述权重分布计通过所述一系列权重分布的最可能标签路径,从被估计为最可能的所述标签路径导出所述一系列聚合物单元的所述估计值。

[0034] 可替代地,从所述权重分布导出所述一系列聚合物单元的估计值的步骤可以包括估计关于每个权重分布最可能的所述标签,将通过所述一系列权重分布的前向和后向标签路径考虑在内,从被估计为最可能的所述标签导出所述一系列聚合物单元的所述估计值。

[0035] 根据本发明的第二方面,提供了一种分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的方法,所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,所述方法包括:使用输出一系列权重分布的机器学习技术来分析所述信号,每个权重分布包括关于一组标签上的标签的权重,所述一组标签包含表示所述可能类型的聚合物单元的标签;以及从所述权重分布导出所述一系列聚合物单元的估计值,其中导出所述一系列聚合物单元的估计值的步骤将表示是否允许或不允许标签之间的转换的转换矩阵考虑在内,标签之间的至少一个转换被表示为不允许并且其它转换被表示为允许。

[0036] 根据本发明的第三方面,提供了一种分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的方法,所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,所述方法包括:使用输出一系列权重分布的机器学习技术来分析所述信号,每个权重分布包括关于一组标签上的标签的权重,所述一组标签包含表示所述可能类型的聚合物单元的标签,其中所述一系列聚合物单元中的相同类型的聚合物单元的连续实例以游程长度编码形式表示;以及从所述权重分布导出所述一系列聚合物单元的估计值。

[0037] 第一方面的任何特征可以以任何组合应用于本发明的第二方面和第三方面。

[0038] 进一步地根据本发明,所述方法可以由在计算机设备中执行的计算机程序来实施,或者可以提供一种分析设备,所述分析设备被布置成实施与本发明的方面中的任何方面类似的方法。

[0039] 更进一步地根据本发明,可以提供一种纳米孔测量和分析系统,其包括与测量系统组合的此类分析设备,所述测量系统被布置成在聚合物相对于纳米孔易位期间从聚合物导出信号。

附图说明

[0040] 为了更好地理解,现在将参考附图通过非限制性实例的方式描述本发明的实施例,在附图中:

[0041] 图1是纳米孔测量和分析系统的示意图;

[0042] 图2是典型信号随时间变化的图;

[0043] 图3是分析系统中的神经网络的图;

[0044] 图4是信号的一部分的曲线图,其展示了神经网络的加窗部分的操作;

[0045] 图5是RNN的递归层的图;

[0046] 图6是非递归层的图;

[0047] 图7是单向层的图;

[0048] 图8是组合了“正向”和“反向”递归层的双向递归层的图;

[0049] 图9是以交替方式组合了“正向”和“反向”递归层的替代性双向递归层的图;

- [0050] 图10是权重分布表,其中权重关于表示四种类型的多核苷酸的标签之间的转换;
- [0051] 图11是权重分布表,其中权重关于表示四种类型的多核苷酸和空白的标签之间的转换;
- [0052] 图12是权重分布表,其中权重关于表示五种类型的多核苷酸和空白的标签之间的转换,所述五种类型的多核苷酸之一是甲基化-C
- [0053] 图13是权重分布表,其中权重关于标签之间的转换,所述标签包含关于四种类型的多核苷酸中的每种的两个标签;
- [0054] 图14是权重分布表,其中权重使用正位-翻转 (flip-flop) 表示来表示均聚物;
- [0055] 图15是使用6聚体信号模型和相对于读头和系统其它组件的近似位置的四个碱基的剩余电流的图;
- [0056] 图16是权重分布表,其中权重使用游程长度编码的表示来表示均聚物;
- [0057] 图17是权重分布表,其中权重使用游程长度编码的表示的不同表达方式来表示均聚物;
- [0058] 图18是权重分布的另外的权重的表,所述权重分布表示对于每种可能类型的均聚物的一组可能长度上的分类分布;
- [0059] 图19是权重分布的另外的权重的表,所述权重分布表示对于每种可能类型的均聚物的可能长度上的参数化分布;
- [0060] 图20是由均值和方差参数的不同值表示的两个分布的图;
- [0061] 图21是可以用于表示均聚物的可能分布的表;
- [0062] 图22是权重分布的另外的权重的表,所述权重分布表示对于每种可能的聚合物单元对的一组可能长度上的分类分布;
- [0063] 图23是权重分布的另外的权重的表,所述权重分布表示对于每种可能的聚合物单元三联体的一组可能长度上的分类分布;
- [0064] 图24是权重分布表,其中一组标签被扩展以包含关于修饰聚合物单元的标签;
- [0065] 图25是用于呈因子分解修饰表示的类型的聚合物单元的未修饰形式和经修饰形式的另外的权重的表;
- [0066] 图26是5-碱基表示的信号和从其估计的聚合物单元的图;
- [0067] 图27是由神经网络的解码器执行的方法的流程图;并且
- [0068] 图28到30是不同解码算法的定义;
- [0069] 图31是另外的解码算法的定义;
- [0070] 图32是用于构建用于正位-翻转表示的目标转换矩阵的算法的定义;
- [0071] 图33是在所有路径上用于训练的目标函数的定义;
- [0072] 图34是用于构建用于多停留表示的目标转换矩阵的算法的定义;
- [0073] 图35是用于构建用于游程长度编码的表示的目标转换矩阵的算法的定义;
- [0074] 图36是信号和从其估计的聚合物单元的图,其展示了长均聚物的实例;
- [0075] 图37是对于最佳路径用于训练的目标函数的定义;
- [0076] 图38是函子的表;
- [0077] 图39是信号和从其估计的聚合物单元的图,其展示了使用锐化训练的正位-翻转表示的实例;并且

[0078] 图40是展示了估计的一系列聚合物单元相对于参考的对准,以用于在没有和有锐化的情况下训练的表示的表。

具体实施方式

[0079] 图1展示了包括测量系统2和分析系统3的纳米孔测量和分析系统1。测量系统2在聚合物相对于纳米孔易位期间从所述聚合物导出信号,所述聚合物包括一系列聚合物单元。分析系统3执行分析信号的方法,以导出一系列聚合物单元的估计值。

[0080] 通常,聚合物可以是任何类型,例如多核苷酸(或核酸)、如蛋白质等多肽或多糖。聚合物可以是天然的或合成的。多核苷酸可以包括均聚物区域。均聚物区域可以包括5到15个核苷酸。

[0081] 在多核苷酸或核酸的情况下,聚合物单元可以是核苷酸。核酸通常是脱氧核糖核酸(DNA)、核糖核酸(RNA)、cDNA或本领域已知的合成核酸,如肽核酸(PNA)、甘油核酸(GNA)、苏糖核酸(TNA)、锁核酸(LNA)或具有核苷酸侧链的其它合成聚合物。PNA主链由通过肽键连接的重复N-(2-氨基乙基)-甘氨酸单元构成。GNA主链由通过磷酸二酯键连接的重复乙二醇单元构成。TNA主链由通过磷酸二酯键连接在一起的重复苏糖构成。LNA是由如上文所讨论的具有将核糖部分中的2'氧和4'碳连接的额外桥的核糖核苷酸形成。核酸可以是单链的、双链的或包括单链区域和双链区域两者。核酸可以包括与一条DNA链杂交的一条RNA链。通常,cDNA、RNA、GNA、TNA或LNA均是单链的。

[0082] 聚合物单元可以是任何类型的核苷酸。核苷酸可以是天然存在的或人工的。例如,所述方法可以用于验证制造的寡核苷酸的序列。核苷酸通常含有核碱基、糖和至少一个磷酸基。核碱基和糖形成核苷。核碱基通常是杂环的。合适的核碱基包含嘌呤和嘧啶,以及更具体地,腺嘌呤、鸟嘌呤、胸腺嘧啶、尿嘧啶和胞嘧啶。糖通常是戊糖。合适的糖包含但不限于核糖和脱氧核糖。核苷酸通常是核糖核苷酸或脱氧核糖核苷酸。核苷酸通常含有单磷酸、二磷酸或三磷酸。

[0083] 核苷酸可以是修饰的碱基,如受损的碱基或表观遗传的碱基。例如,核苷酸可以包括嘧啶二聚体。此类二聚体通常与紫外光引起的损坏相关联并且是皮肤黑色素瘤的主要病因。核苷酸可以被标记或修饰以充当具有不同信号的标记物。这一技术可用于识别碱基的缺失,例如多核苷酸中的脱碱基单元或间隔基。所述方法也可以应用于任何类型的聚合物。

[0084] 在多肽的情况下,聚合物单元可以是天然存在的或合成的氨基酸。

[0085] 在多糖的情况下,聚合物单元可以是单糖。

[0086] 特别是在测量系统2包括纳米孔且聚合物包括多核苷酸的情况下,多核苷酸可以是长的,例如至少5kB(千碱基),即至少5,000个核苷酸;或至少30kB(千碱基),即至少30,000个核苷酸;或至少100kB(千碱基),即至少100,000个核苷酸。

[0087] 测量系统2的性质和所得信号如下。

[0088] 测量系统2是包括一个或多个纳米孔的纳米孔系统。在简化类型中,测量系统2仅具有单个纳米孔,但是更实际的测量系统2通常在阵列中采用许多纳米孔,以提供并行的信息收集。

[0089] 通常通过纳米孔可以在聚合物相对于纳米孔易位期间记录信号。

[0090] 纳米孔是通常具有纳米级大小的孔,其可以允许聚合物通过其中。

[0091] 纳米孔可以是蛋白孔或固态孔。孔的尺寸可以使得一次仅有一种聚合物可以使孔易位。

[0092] 在纳米孔是蛋白孔的情况下,其可以具有以下性质。

[0093] 生物孔可以是跨膜蛋白孔。用于根据本发明使用的跨膜蛋白孔可以源自 β 桶孔或 α 螺旋束孔。 β 桶孔包括由 β 链形成的桶或通道。合适的 β 桶孔包含但不限于: β -毒素,如 α -溶血素、炭疽毒素和杀白细胞素;以及细菌的外膜蛋白/孔蛋白,如耻垢分枝杆菌孔蛋白(Msp),例如MspA、MspB、MspC或MspD、胞溶素、外膜孔蛋白F(OmpF)、外膜孔蛋白G(OmpG)、外膜磷脂酶A和奈瑟氏菌属自转运蛋白脂蛋白(Na1P)。 α 螺旋束孔包括由 α 螺旋形成的桶或通道。合适的 α 螺旋束孔包含但不限于内膜蛋白和 α 外膜蛋白,如WZA和ClyA毒素。跨膜孔可以源自Msp或来自 α -溶血素(α -HL)。跨膜孔可以源自胞溶素。W02013/153359中公开了源自胞溶素的合适的孔。W0-2012/107778中公开了源自MspA的合适的孔。孔可以源自如W0-2016/034591中所公开的CsgG。孔可以是DNA折纸孔。

[0094] 蛋白孔可以是天然存在的孔或可以是突变孔。典型的孔描述于以下中:W0-2010/109197;Stoddart D等人,《美国国家科学院院刊(Proc Natl Acad Sci)》,12;106(19):7702-7;Stoddart D等人,《应用化学国际英文版(Angew Chem Int Ed Engl.)》2010;49(3):556-9;Stoddart D等人,《纳米快报(Nano Lett.)》2010年9月8日;10(9):3633-7;Butler TZ等人,《美国国家科学院院刊》2008;105(52):20647-52;以及W0-2012/107778。

[0095] 蛋白孔可以是W0-2015/140535中描述的类型蛋白孔之一并且可以具有其中所公开的序列。

[0096] 可以将蛋白孔插入到如生物膜等两亲层例如脂质双层中。两亲层是由具有亲水性和亲脂性两者的如磷脂等两亲分子形成的层。两亲层可以是单层或双层。两亲层可以是Gonzalez-Perez等人,《朗缪尔(Langmuir)》,2009,25,10447-10450或W02014/064444中所公开的共嵌段聚合物。可替代地,例如,如W02012/005857中所公开的,可以将蛋白孔插入到固态层中设置的孔中。

[0097] W0-2014/064443中公开了用于提供纳米孔阵列的合适的设备。纳米孔可以跨相应的孔提供,其中电极设置在每个相应的孔中,与ASIC电连接,以便测量流过每个纳米孔的电流。合适的电流测量设备可以包括如W0-2016/181118中所公开的电流感测电路。

[0098] 纳米孔可以包括形成在固态层中的孔,其可以被称为固态孔。孔可以是固态层中提供的阱、间隙、通道、沟槽或狭缝,分析物可以通过或进入所述固态层。这种固态层不是生物来源的。换言之,固态层不是从生物环境(如生物体或细胞)或合成制造形式的生物学可用结构中产生的,也不是从其中分离出来的。固态层可以由有机材料和无机材料形成,所述材料包含但不限于:微电子材料、如Si₃N₄、Al₂O₃和SiO₂等绝缘材料、如聚酰胺等有机聚合物和无机聚合物、如Teflon®等塑料或如二组分加成固化的硅橡胶等弹性体、以及玻璃。固态层可以由石墨烯形成。W0-2009/035647、W0-2011/046706或W0-2012/138357中公开了合适的石墨烯层。W0-2016/187519中公开了制备一排固态孔的合适方法。

[0099] 这种固态孔通常是固态层中的孔。可以通过化学方法或其它方式对孔进行修饰,以增强其作为纳米孔的性质。固态孔可以与提供聚合物的替代性或另外的测量结果的另外的组分结合使用,如隧穿电极(Ivanov AP等人,《纳米快报》2011年1月12日;11(1):279-85),或场效应晶体管(FET)装置(例如在W0-2005/124888中所公开的)。可以通过已知方法

形成固态孔,所述已知方法包含例如W0-00/79257中描述的方法。

[0100] 纳米孔可以是固态孔与蛋白孔的混合体。

[0101] 测量系统2进行一系列的性质测量,所述性质取决于相对于孔易位的聚合物单元。一系列测量形成信号。

[0102] 所述测量的性质可能与聚合物和孔之间的相互作用相关。这种相互作用可以发生在孔的收缩区域。

[0103] 在一种类型的测量系统2中,所测量的性质可能是流过纳米孔的离子电流。这些或其它电性质可以使用标准单通道记录设备来测量,如下所述:Stoddart D等人,《美国国家科学院院刊》,12,106(19):7702-7;Lieberman KR等人,《美国化学会志(J Am Chem Soc.)》2010;132(50):17961-72以及W0-2000/28312。可替代地,电性质的测量可以使用例如,如W0-2009/077734、W0-2011/067559或W0-2014/064443中所描述的多通道系统来进行。

[0104] 可以在膜或固态层的任一侧提供离子溶液,所述离子溶液可以存在于相应的隔室中。可以向膜的一侧添加含有所关注的聚合物分析物的样品并允许其相对于纳米孔移动,例如在电势差或化学梯度下。信号可以在聚合物相对于孔移动期间导出,例如在聚合物通过纳米孔易位期间进行的。聚合物可以部分地使纳米孔易位。

[0105] 为了在聚合物易位通过纳米孔时进行测量,可以通过聚合物结合部分控制易位速率。通常,所述部分可以使聚合物与施加的场一起或相对于施加的场移动通过纳米孔。所述部分可以是分子马达,在所述部分是酶的情况下,所述分子马达使用例如酶活性或作为分子刹车。在聚合物是多核苷酸的情况下,提出了许多用于控制易位速率的方法,包含使用多核苷酸结合酶。用于控制多核苷酸易位速率的合适的酶包含但不限于聚合酶、解旋酶、外切核酸酶、单链和双链结合蛋白以及拓扑异构酶(如旋转酶)。对于其它聚合物类型,可以使用与所述聚合物类型相互作用的部分。聚合物相互作用部分可以是以下文献中所公开的任何:W0-2010/086603、W0-2012/107778和Lieberman KR等人,《美国化学会志》2010;132(50):17961-72,并且用于电压门控方案(Luan B等人,《物理评论快报(Phys Rev Lett.)》2010;104(23):238103)。

[0106] 可以以多种方式使用聚合物结合部分来控制聚合物运动。所述部分可以使聚合物与施加的场一起或相对于施加的场移动通过纳米孔。所述部分可以用作分子马达,在所述部分是酶的情况下,所述分子马达使用例如酶活性或作为分子刹车。可以通过控制聚合物通过孔的运动的分子制动器来控制聚合物的易位。分子制动器可以是聚合物结合蛋白。对于多核苷酸,多核苷酸结合蛋白优选地是多核苷酸处理酶。多核苷酸处理酶是能够与多核苷酸相互作用并且修饰其的至少一个性质的多肽。酶可以通过切割多核苷酸以形成单独的核苷酸或较短核苷酸链如二核苷酸或三核苷酸来对多核苷酸进行修饰。所述酶可以通过将多核苷酸朝向或使其移动到特定位置来对多核苷酸进行修饰。多核苷酸操作酶并不需要显示酶活性,只要其能够结合靶多核苷酸并且控制其移动通过孔即可。例如,可以对酶进行修饰以移除其酶活性或可以在防止其充当酶的条件下使用。下文更详细地论述了此类条件。

[0107] 优选的多核苷酸处理酶是聚合酶、外切核酸酶、解旋酶和拓扑异构酶(如旋转酶)。核苷酸处理酶可以是例如W0-2015/140535或W0-2010/086603中描述的一种类型的多核苷酸处理酶。

[0108] 聚合物通过纳米孔的易位可以按以下方式发生:顺式到反式或反式到顺式,与施

加的电势一起或相对于施加的电势。可以在施加的电势下发生易位,所述施加的电势可以控制易位。

[0109] 在双链DNA上逐渐或逐步起作用的外切核酸酶可以在孔的顺式侧使用,以在施加的电势下供给剩余的单链或在反向电势下供给反式侧。同样,使双链DNA解旋的解旋酶还可以以类似的方式使用。还存在需要抵抗所施加的电势的链易位的测序应用的可能性,但是DNA必须首先在相反或无电势下由酶“捕获”。随着电势随后在结合后转回,链将以顺式到反式的方式穿过孔并且通过电流保持处于延长的构型。单链DNA外切核酸酶或单链DNA依赖性聚合酶可以充当分子马达,所述分子马达将最近易位的单链以逐步受控方式(反式到顺式,相对于施加的电势)牵拉回孔中。可替代地,单链DNA依赖性聚合酶可以充当减慢多核苷酸通过孔的移动的分子刹车。可以使用W0-2012/107778或W0-2012/033524中描述的任何部分、技术或酶来控制聚合物运动。

[0110] 然而,测量系统2可以是包括一个或多个纳米孔的替代性类型。

[0111] 类似地,所测量的性质可以是离子电流以外的类型。替代性性质的一些实例包含但不限于:电性质和光学性质。涉及荧光测量的合适的光学方法公开于《美国化学会志》2009,131 1652-1653。可能的电性质包含:离子电流、阻抗、隧穿性质,例如隧穿电流(例如,如公开于以下中:Ivanov AP等人,《纳米快报》2011年1月12日;11(1):279-85)和FET(场效应晶体管)电压(例如,如公开于W02005/124888中)。可以使用一种或多种光学性质,任选地与电性质组合(Soni GV等人,《科学仪器综述(Rev Sci Instrum.)》2010年1月;81(1):014301)。所述性质可以是跨膜电流,如流过纳米孔的离子电流。离子电流通常可以是DC离子电流,但是原则上替代方案使用AC电流(即,在施加AC电压下流动的AC电流的大小)。

[0112] 在一些类型的测量系统2中,所述信号可以被表征为包括来自一系列事件的测量结果,其中每个事件提供一组测量结果。图2展示了在电流测量的情况下此类信号10的典型实例。来自每个事件的测量结果组具有类似的水平,但是存在一些差异。这可以被认为是噪声步波,其中每个步与事件相对应。事件可以具有生物化学意义,例如由测量系统2的给定状态或相互作用引起。在某些情况下,这可能是由于以制动方式发生的聚合物通过纳米孔的易位引起的。然而,并非所有类型的测量系统都产生这种类型的信号,并且本文所描述的方法不取决于信号的类型。例如,当易位速率接近测量采样速率(例如,以聚合物单元的易位速率的1倍、2倍、5倍或10倍的速率获取测量结果)时,与较慢的测序速度或较快的采样速率相比,事件可能不那么明显或不存在。

[0113] 另外,在存在事件的情况下,通常不存在组中测量结果的数量先验知识,所述数量不可预测地变化。这些变化因素和缺乏对测量结果数量的了解可能使得难以区分一些组,例如在组是短的和/或两个连续组的测量结果水平彼此接近的情况下。

[0114] 与每个事件相对应的测量结果组通常具有在事件的时间尺度上一致的水平,但是对于大多数类型的测量系统2将在短时间尺度上发生变化。这种变化可能由测量噪声引起,例如由电路和信号处理产生,特别是在电生理学的特定情况下来自于放大器。由于测量小幅度的性质,这种测量噪声是不可避免的。这种变化还可能由测量系统2的基础物理或生物系统中的固有变化或扩散引起,例如可能由聚合物的构型变化引起的相互作用的变化。

[0115] 大多数类型的测量系统2将在更大或更小的程度上经历这种固有的变化。对于任何给定类型的测量系统2,两种变化源都可能有所贡献,或者这些噪声源中的一个可能占主

导地位。

[0116] 随着测序速率(即聚合物单元相对于纳米孔易位的速率)的增加,那么事件可能变得不那么明显,并因此更难鉴定或可能消失。因此,随着测序速率的增加,依赖于检测此类事件检测的分析方法可能变得更低效。

[0117] 然而,本文所公开的方法不依赖于检测此类事件。下文所描述的方法即使在相对高的测序速率下也是有效的,所述相对高的测序速率包含测序速率,在所述测序速率下,聚合物以每秒至少10个聚合物单元,优选地每秒100个聚合物单元,更优选地每秒500个聚合物单元或更优选地每秒1000个聚合物单元的速率易位。

[0118] 采样速率是信号中的测量速率。通常,采样速率高于测序速率。例如,采样速率可以在100Hz到30kHz的范围内,但这不是限制性的。实际上,采样速率可以取决于测量系统2的性质。

[0119] 在一些情况下,所述方法可以使用多个系列的测量结果,所述多个系列的测量结果是相关的一系列聚合物单元的测量结果。例如,所述多个系列的测量结果可以是具有相关序列的分离的聚合物的一系列测量结果,或者可以是具有相关序列的相同聚合物的不同区域的一系列测量结果。

[0120] 在多核苷酸的情况下,所述多个系列的聚合物单元可以通过互补而相关,使得一个系列的聚合物单元被称为模板,并且另一系列的与其互补的聚合物单元被称为补体。在这种情况下,可以使用任何合适的技术来测量模板和补体,例如使用多核苷酸结合蛋白或通过多核苷酸样品制备顺序地进行。合适的方法包含W0-2010/086622或W0-2013/014451中描述的那些方法。

[0121] 例如,通过使用W0-2010/086622或W0-2013/014451中描述的方法,本文所公开的涉及单个系列的聚合物单元的方法中的任何方法可以应用于所述多个系列的测量结果,如模板和补体。

[0122] 一系列测量结果形成被分析系统3分析的原始信号。原始信号可以在供应给分析系统2之前在测量系统2中进行预处理,或者作为分析系统3中的初始阶段进行预处理,例如经过滤波以减少噪声。在此类情况下,将对预处理的信号执行以下分析。

[0123] 分析系统3可以与测量系统2物理相关,并且还可以向测量系统2提供控制信号。在所述情况下,可以如W0-2008/102210、W0-2009/07734、W0-2010/122293、W0-2011/067559或W02014/04443中任一个所公开的布置包括测量系统2和分析系统3的纳米孔测量和分析系统1。

[0124] 可替代地,可以在单独的设备中实施分析系统3,在这种情况下,通过任何合适的装置(通常是数据网络)将所述一系列测量结果从测量系统2传送到分析系统3。例如,一种方便的基于云的实施方案是将分析系统3作为服务器,通过互联网向其提供输入信号11。

[0125] 分析系统3可以由执行计算机程序的计算机设备实施,或者可以由专用硬件装置或其任何组合实施。在任一种情况下,所述方法使用的数据储存在分析系统3的存储器中。

[0126] 在计算机设备执行计算机程序的情况下,计算机设备可以是任何类型的计算机系统,但通常是常规结构。可以用任何合适的编程语言编写计算机程序。计算机程序可以储存在任何类型的计算机可读存储介质上,例如:可插入到计算系统的驱动器中并且可以磁性地、光学地或光磁地储存信息的记录介质;计算机系统的固定记录介质,如硬盘驱动器;或

计算机存储器。

[0127] 在计算机设备由专用硬件装置实施的情况下,可以使用任何合适类型的装置,例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。在优选的实施例中,可以使用如图形处理单元(GPU)等适于并行化计算的硬件来实施计算机程序的各部分。

[0128] 如下执行使用纳米孔测量和分析系统1的方法。

[0129] 首先,使用测量系统2导出信号10。例如,使聚合物相对于孔(例如,通过孔)易位,并且在聚合物易位期间导出信号。可以通过提供允许聚合物易位的条件使聚合物相对于孔易位,因此易位可以自发发生。在易位期间,导出信号10。

[0130] 其次,分析系统3执行分析信号10的方法,现在将要对其进行描述。

[0131] 分析系统3使用神经网络20分析信号。神经网络20的参数在下文进一步描述的训练期间取值,并且因此递归神经网络不依赖于具有任何特定形式的测量结果或具有任何特定性质的测量系统2。例如,递归神经网络不依赖于取决于包括k个连续聚合物单元的k-聚体的测量结果。

[0132] 合适的神经网络20的实例示出在图3中并且包含加窗单元30,CNN(卷积神经网络)40、RNN(递归神经网络)50和解码器80,它们依次对信号进行如下处理。

[0133] 加窗单元30使信号10加窗,以导出信号10的连续加窗部分31,例如,如图4所展示的。加窗部分11被供应给CNN 40。

[0134] 加窗部分31具有长度32,以及在连续加窗部分31之间的步幅33,这两者都可以在时间上或在信号10的样品数上进行计数。步幅33可以是单个样品或多个样品。如果步幅33是单个样品,则忽略边缘效应,存在与信号10中的样品相同数量的加窗部分31。如果步幅33大于单个样品,则加窗单元30执行下采样并且存在比信号10中的样品少的加窗部分31。通常,步幅33小于长度33,使得加窗部分10在信号10中重叠。

[0135] 举例来说,长度32可以是4.75毫秒,并且步幅可以是0.5毫秒。

[0136] 通过另外的实例,长度可以是19个样品,并且步幅可以是2个样品。

[0137] CNN 40包括至少一个卷积层。至少一个卷积层对每个加窗部分11执行卷积,以导出关于每个加窗部分31的特征向量41。不管信号中是否可能出现任何事件,都可以这样做,并且因此同样适用于此类事件明显或不明显的情况下的信号,或在预处理期间提供事件的情况下的信号。特征向量41被供应给RNN 50。

[0138] CNN 40与RNN 50一起训练,如下文所讨论的。

[0139] CNN 40可以采用任何形式。

[0140] 在一个实例中,CNN 40可以是单个卷积层,由具有权重W和偏差b以及激活函数g的仿射变换定义。在本文中, $I_{t-j:t+k}$ 表示包括含有t-j到t+k测量结果的原始信号20的测量结果窗口,并且 O_t 是输出特征向量。

[0141] $y_t = AI_{t-j:t+k} + b$ 仿射变换

[0142] $O_t = g(y_t)$ 激活

[0143] 双曲正切是合适的激活函数,但是本领域已知更多的替代方案,包括但不限于:整流线性单元(ReLU)、指数线性单元(ELU)、softplus单元和S形单元。也可以使用多个卷积层。

[0144] 在另一个实例中,CNN 40可以采用与参考1中的CNN相同的形式。

[0145] 如所描述的,直接卷积网络的缺点在于,原始信号中检测到的特征的确切位置存在依赖性,并且这也意味着对特征之间的间隔的依赖性。通过将由第一个卷积产生的特征向量的输出序列用作对输入的阶统计量起作用的第二个“合并”网络中的输入,可以减轻依赖性。

[0146] 举例来说,在合并网络是单层神经网络的情况下,以下等式描述了输出如何与输入向量相关。将 f 设置为输入特征的指数,所以 A_f 是特征 f 的权重矩阵,并且将 $\mathcal{S}$ 设置为返回其输入的部分或全部阶统计量的函数:

$$\begin{aligned} y_t &= b + \sum_f A_f \mathcal{S}(I_{f,t-j:t+k}) && \text{仿射变换} \\ o_t &= g(y_t) && \text{激活} \end{aligned} \quad [0147]$$

[0148] 这种层的一个有用的且计算上有效的实例是返回特征向量,所述特征向量与输入特征的数量相同,其元素是针对每个相应特征获得的最大值。将函数 $\mathcal{S}_M$ 设置为仅返回最后阶统计量,作为其输入中获得的最大值,并且将 U_f 设置为完全由零而不是由其 $(f, 1)$ 元素处的单位值组成的(单列)矩阵:

$$\begin{aligned} y_t &= \sum_f U_f \mathcal{S}_M(I_{f,t-j:t+k}) && \text{仿射变换} \\ o_t &= y_t && \text{未应用激活} \end{aligned} \quad [0149]$$

[0150] 由于矩阵 U_f 非常稀疏,出于计算效率的原因,可以隐式地执行矩阵乘法:在本文中, $\sum_f U_f x_f$ 的作用是将输出特征向量的元素 f 设置为 x_f 。

[0151] 可以仅针对每第 n 个位置(n 的步幅)计算它们的输出并且因此对它们的输出进行下采样来执行卷积和/或合并。从计算角度来看,下采样可能是有利的,因为网络的其余部分必须处理更少的块(更快的计算)以实现类似的准确度。

[0152] 添加卷积层的堆栈解决了上文所描述的许多问题:通过卷积学习的特征检测既可以作为纳米孔特有的特征检测器,也可以作为汇总统计,而无需对系统做任何另外的假设;特征不确定性通过不同特征的相对权重传递到网络的其余部分,并且因此进一步处理可以将这一信息考虑在内,从而导致更精确的预测和对不确定性的量化。

[0153] RNN 50输出一系列权重分布。RNN 50包括至少一个递归层52,所述递归层或每个递归层之后是前馈层53。图5展示了单个递归层52的情况下的RNN,但是通常可以存在任何多数量的递归层52和随后的前馈层53。这提供了单元架构的灵活选择。所述层可以具有不同的参数,可以具有不同的大小,或者甚至可以由不同的单元类型组成。

[0154] 所述递归层或每个递归层52优选地是双向的,以允许每个输入特征向量的影响在两个方向上通过RNN传播。替代性优选的实施例包括在交替方向上布置的多个单向递归层,例如在反向、正向、反向、正向、反向的连续方向上布置的层。这些双向架构允许RNN 50以HMM不可用的方式来累积和传播信息。递归层的另外的优点是,它们不需要精确的信号缩放来建模(反之亦然),例如通过迭代程序。

[0155] 对于前馈层53中的下采样,将单独的仿射变换应用于每列的前向和后向层的输出向量,然后进行求和;这相当于将仿射变换应用于由输入和输出的串联形成的向量。然后,逐个元素地将激活函数应用于所得矩阵。

[0156] 递归层52可以使用几种类型的神经网络单元,现在将要对其进行描述。单元类型分为两大类,取决于其是否是“递归的”。虽然非-递归单元独立地处理序列中的每个步,但是递归单元被设计成在序列中使用并且将状态向量从一个步传递到下一个步。

[0157] 为了示意性地示出非递归单元与递归单元之间的差异,图6示出了非递归单元61的非递归层60并且图7到9示出了相应非递归单元64到66的三个不同层62到64。在图6到9的每个图中,箭头示出了向量通过的连接,被分割的箭头是复制的向量并且被组合的箭头是串联的向量。

[0158] 在图6的非递归层60中,非递归单元61具有不进行分割或串联的单独的输入和输出。

[0159] 图7的递归层62是单向递归层,其中递归单元65的输出向量被分割并单向传递到递归层中的下一个递归单元65。

[0160] 虽然图8和图9的双向递归层63和64本身不是分立单元,但其各自具有分别由更简单的递归单元66和67制成的重复单元状结构。

[0161] 在图8的双向递归层63中,双向递归层63由递归单元66的两个子层68和69组成,它们是具有与图7的单向递归层62相同的结构的前向子层68和具有与图7的单向递归层62相反的结构的后向子层69,好像时间被反转,从而将状态向量从一个单元66传递到前一单元66。前向子层68和后向子层69二者均接收相同的输入,并且其来自于对应单元66的输出被串联在一起以形成双向递归层63的输出。注意,前向子层68内的任何单元66与后向子层69内的任何单元之间不存在串联。

[0162] 类似地,图9的替代性双向递归层64由递归单元67的两个子层70和71组成,它们是具有与图7的单向递归层62相同的结构的前向子层68和具有与图7的单向递归层62相反的结构的后向子层69,好像时间被反转。再次,前向子层68和后向子层69接收相同的输入。然而,与图8的双向递归层63相反,前向子层68的输出是后向子层69的输入并且后向子层69的输出形成双向递归层64的输出(前向子层68和后向子层69可以反转)。

[0163] 图9中所示的双向递归层64的一般化将是由多个“前向”和“后向”递归子层组成的递归层的堆栈,其中每层的输出是下一层的输入。

[0164] RNN 50的双向递归层52可以采用图8和9的双向递归层63和64中的任一个的形式。通常,图3的双向递归层34可以由非递归层(例如,图6的非递归层60)或由单向递归层(例如,图7的递归层62)代替,但通过使用双向递归层34实现了改善的性能。

[0165] 现在将描述前馈层53。

[0166] 前馈层53包括处理相应向量的前馈单元54。前馈单元54是经典神经网络中的标准单元,即,将仿射变换应用于输入向量并且然后逐元素地应用非线性函数。前馈层53均使用非线性函数的双曲正切,但是可以使用许多其它内容,而在网络的总体准确度上几乎没有变化。

[0167] 如果步 t 处的输入向量是 I_t ,并且仿射变换的权重矩阵和偏差分别为 A 和 b ,则输出向量 O_t 为:

[0168] $y_t = AI_t + b$ 仿射变换

[0169] $O_t = \tanh(y_t)$ 非线性度

[0170] 将RNN 50的权重分布全局归一化。在下文中更详细地论述这一点。

[0171] 非递归单元62和递归单元65到67独立地处理每个事件,但是可以由具有现在将描述的形式的长短期记忆单元代替。

[0172] 长短期记忆 (LSTM) 单元在以下文献中被引入:Hochreiter和Schmidhuber,《长短期记忆 (Long short-term memory)》,《神经计算 (Neural Computation)》,9 (8):1735-1780,1997。LSTM单元是递归单元,并且因此将状态向量从序列中的一步传递到下一步。LSTM基于单元是存储器单元的符号:将含有存储器内容的隐藏状态从一步传递到下一步,并通过一系列控制存储器更新的门来对所述隐藏状态进行操作。一个门控制是否擦除(遗忘)存储器的每个元素,另一个门控制是否由新值替换,并且最终门测定是否从存储器读取和输出。概念计算机存储器单元的二进制开/关逻辑门由S形函数产生的概念概率代替并且存储器单元的内容代表预期值,这使得存储器单元是可微分的。

[0173] 首先描述了LSTM的标准实施方案,并且然后描述基本方法中实际使用的“窥视孔”修改。

[0174] 标准LSTM如下。

[0175] 与对LSTM单元的不同操作相关联的概率由以下方程组定义。将 I_t 设置为步骤t的输入向量, O_t 为输出向量,并将由x指数化的仿射变换设置为具有偏差 b_x 和分别用于输入和先前输出的权重矩阵 W_{x1} 和 W_{x0} ;σ是非线性S形变换。

$$f_t = \sigma(W_{f1}I_t + W_{f0}O_{t-1} + b_f) \quad \text{遗忘概率}$$

[0176] $u_t = \sigma(W_{u1}I_t + W_{u0}O_{t-1} + b_u) \quad \text{更新概率}$

$$o_t = \sigma(W_{o1}I_t + W_{o0}O_{t-1} + b_o) \quad \text{输出概率}$$

[0177] 给定上文定义的更新向量并设置_运算符表示逐元素 (Hadamard) 乘法,则用于更新内部状态 S_t 和测定新输出的等式是:

$$v_t = \tanh(W_{v1}I_t + W_{v0}O_{t-1} + b_v) \quad \text{要更新的值}$$

[0178] $S_t = S_{t-1} \circ f_t + v_t \circ u_t \quad \text{更新存储器单元}$

$$O_t = \tanh(S_t) \circ o_t \quad \text{从存储器单元读取}$$

[0179] 窥视孔修改如下。

[0180] “窥视孔”修改 (Gers和Schmidhuber,2000) 为LSTM架构添加了一些另外的连接,所述架构允许遗忘概率、更新概率和输出概率“窥视”存储器单元的隐藏状态 (或由其通知)。网络的更新等式如上所述,但将 P_x 设置为长度等于隐藏状态的“窥视”向量,概率向量的三个等式变为:

$$f_t = \sigma(W_{f1}I_t + W_{f0}O_{t-1} + b_f + P_{f-} \circ S_{t-1}) \quad \text{遗忘概率}$$

[0181] $u_t = \sigma(W_{u1}I_t + W_{u0}O_{t-1} + b_u + P_{u-} \circ S_{t-1}) \quad \text{更新概率}$

$$o_t = \sigma(W_{o1}I_t + W_{o0}O_{t-1} + b_o + P_{o-} \circ S_t) \quad \text{输出概率}$$

[0182] 非递归单元62和递归单元65到67可以可替代地由具有如下形式的门控递归单元代替。

[0183] 已经发现门控递归单元 (GRU) 运行得更快,但最初发现其产生较差的准确度。GRU的架构不像省去隐藏状态与输出之间的分离并且还组合了“遗忘”和“输入门”的LSTM那样

直观。

$$\begin{aligned}
 o_t &= \sigma(W_{oI}I_t + W_{oS}S_{t-1} + b_o) && \text{输出概率} \\
 u_t &= S_{t-1} \circ \sigma(W_{uI}I_t + W_{uS}S_{t-1} + b_u) && \text{从状态更新} \\
 v_t &= \tanh(W_{vI}I_t + W_{vR}u_t + b_v) && \text{要更新的值} \\
 S_t &= (1 - o_t) \circ S_{t-1} + o_t \circ v_t && \text{更新状态}
 \end{aligned}$$

[0185] 虽然存在与事件相同数量的列输出,但假设用网络输入中的单个事件鉴定每列是不正确的,因为由于双向层的存在,每列的内容可能由整个输入事件集通知。输入事件与输出列之间的任何对应关系都是通过用训练集中的符号对其进行标记来实现的。

[0186] 在另一个实例中,RNN 50可以采用与参考1中的RNN相同的形式。

[0187] 现在将讨论由RNN 50输出的一系列权重分布51。

[0188] 输出关于连续的时间步长的权重分布51,以形成一系列的权重分布。时间步长原则上可以与信号10的样品周期具有相同的长度,但是由于神经网络20中的过采样,所述时间步长通常比信号10的样品周期更长。然而,时间步长是规则的长度,例如对应于加窗单元30的步幅13,这与执行事件调用的系统相反,并且因此连续权重分布51之间的时间步长对应于可变的连续检测到的事件的长度。

[0189] 通常,权重分布51以比连续的聚合物单元相对于纳米孔易位的速率更高的速率输出,即,权重分布51比聚合物单元多。与所述系列中的每个聚合物单元相对应的多个权重分布51是先验未知的。

[0190] 每个权重分布51包括多个权重。所述权重表示后验概率。权重可以是实际的后验概率,或更通常地,可以是不是实际概率但仍然表示后验概率的权重。通常,在权重不是实际概率的情况下,原则上可以根据权重的归一化确定后验概率。

[0191] RNN 50输出关于一组标签上的标签之间的转换的权重,所述一组标签包含表示可能类型的聚合物单元的标签。因此,关于转换的权重表示那些转换的后验概率。因为权重分布51比聚合物单元多,所以要理解,在一些表示中,允许从标签到相同标签的转换,并且因此权重分布51包含关于这种转换的权重,即单词“转换”既不意味着必须更改标签,也不意味着必须发出另外的聚合物单元。

[0192] 下文给出由RNN 50输出的权重分布51的各个实例。这些实例中的每个是指聚合物单元是多核苷酸并且聚合物单元的类型是四个碱基A,C,G和T的情况。如上文所讨论的,本发明方法同样适用于更多类型的多核苷酸和/或非核苷酸的聚合物单元,因此可以相应地概括这些实例。在所述实例的每个中,权重分布51包含表示标签之间的转换的权重。因此,权重被标记为 w_{ij} ,其中i是从其进行转换的标签的指数,并且j是转换进行到的标签的指数。因此权重 w_{ij} 是从标签i到标签j转换的权重。在附图中的每个中,行对应于从其发生转换的标签i,并且列对应于转换到其的标签j。

[0193] 这种配置的一个实例是在其输出(最终)层中配置有等于要产生的权重分布中的权重数的数量的前馈元素的RNN。图7-9还提供了示出了来自RNN单元的许多输出的实例;应当理解,在RNN内可以存在这些配置中的任何一个或多个,使得输出的数量等于要产生的权重分布中的权重的数量。

[0194] 图10和11示出了可以由RNN 50输出的权重分布51的两个实例。

[0195] 在图10的实例中,关于示出为A、C、G和T的四个碱基,存在单个标签。允许所有转换,因此关于从标签中的每个到标签中的每个的16个转换总共存在十六个权重 w_{ij} 。

[0196] 图10的实例没有提供均聚物的良好表示,所述均聚物是一系列聚合物单元内的相同类型的多个聚合物单元的连续。这是因为从标签到相同标签的转换不能区分给定类型的聚合物单元的相同实例和给定类型的聚合物单元的另外的实例。结果,从标签到相同标签的一系列转换表示一系列任何数量(一个或多个)的聚合物单元(即,具有任何长度的单个聚合物单元和相同类型的聚合物单元的均聚物两者)的实例。

[0197] 然而,图11是通过扩展图10的表示改善均聚物的表示的实例,使得所述一组标签包含(i)单个标签,每个标签表示四个碱基中的一个不同碱基,以及(ii)表示一系列聚合物单元中空白的标签。在图11的实例中示出了所有允许的转换,因此关于从标签中的每个到标签中的每个的25个转换,总共有25个权重 w_{ij} 。在此表示中,空白标签表示所述系列中的碱基(聚合物单元)的两个实例之间的分隔,即使所述实例是相同类型。

[0198] 换句话说,在聚合物单元的序列中,可能的情况是分析了数据测量的窗口,但在所述数据窗口中不存在聚合物单元之间的转换。在此情况下,转换可以表示为从先前标签到“空白”标签的转换,这表示没有新的聚合物单元的实例转换到窗口中。

[0199] 在一些实施例中,可以将空白视为强制性的,因为在确定的聚合物单元序列中必须存在空白,以便将空白任一侧上的聚合物单元视为单独的聚合物单元。例如,在以下生成的标签序列的情况中,空白用“-”表示:A A A - A,这将解析为聚合物单元的实际序列=A A A。“A”标签的前三个实例中的每个被视为相同的实际聚合物单元“A”的实例,而最后一个“A”标签被视为不同的,因为它与前三个“A”之间用两个空白标签隔开。

[0200] 在一些实施例中,空白可以被视为任选的,因为空白表示聚合物单元与标签的重复之间的间隔子。例如,在以下生成的标签序列的情况中,空白用“-”表示:A A A - A,这将解析为聚合物单元的实际序列=A A A A。“A”标签的前三个实例中的每个被视为不同的聚合物单元,并且空白标签充当这些单元与最终“A”标签之间的间隔子。

[0201] 在RNN 50输出中使用关于标签之间的转换的权重的此表示与参考1形成对比,其中RNN输出关于一组标签上的标签的后验概率(权重的特定实例),所述一组标签由表示四种类型的多核苷酸(即碱基C、G、A和T)中的每一种的四个标签和表示空白的标签组成。使用关于标签之间的转换的权重的表示优于使用关于标签的权重的表示,因为提供了另外的信息,所述信息提高了估计一系列聚合物单元的准确度。这是因为配重提供了关于通过一系列聚合物单元的可能路径的信息,而出于进一步分析的目的,关于标签本身的权重会丢失关于与其它标签的关系的信息。因此,将另外的信息提供给估计聚合物单元的步骤,这提高了解码的准确度。

[0202] 另外,所述表示允许表示允许和不允许的转换。即,标签可以不允许标签之间的转换中的一个或多个转换并且允许其它转换的方式表示可能类型的聚合物单元。在所述情况下,权重分布51包括关于被允许的转换的权重。所述权重分布51可以包括关于不被允许的转换的零权重。

[0203] 零权重可以是在由RNN 50输出的权重分布51中不存在权重。在附图所示出的实例中,通过不存在权重来展示零权重,但是可以替代地应用下文的替代方案。

[0204] 可替代地,零权重可以是存在于由RNN 50输出的权重分布51中的权重,以易于实

施RNN 50,但是具有标称值。此类标称值可以是具有零值或无关紧要大小的值,使得所述标称值确实影响由解码器80执行的估计,如下文所描述的。可替代地,此类标称值可以是存在于由RNN输出的权重分布51中但被解码器80忽略的值,例如通过使用如下文所描述的转换矩阵。

[0205] 这一点的一些实例如下。

[0206] 发生允许的转换和不允许的转换的第一实例是一组可能类型的聚合物单元包含总是以聚合物单元的已知序列出现的类型的聚合物单元。在此情况下,允许与已知序列相一致的转换,并且不允许与已知序列相反的转换。对于多核苷酸,这样的实例是,脊椎动物中的5-甲基胞嘧啶仅出现在鸟嘌呤之前的胞嘧啶(“CpG”)上,并且这可以用于进一步限制可能的转换,并且因此,需要来自RNN50的较少权重。即,CpG甲基化导致甲基化C(其在本文中将为表示为 C^M)始终在G之前,因此 C^M 始终以已知序列 C^MG 出现。图12是用于表示这一点的权重分布51的实例。权重分布51改编自图11的权重分布,以将表示甲基化C的标签添加到表示四种类型的多核苷酸(即,碱基C、G、A和T)的四个标签和表示空白的标签。在此情况下,不允许从 C^M 到A、C或T的转换,因此,对于这些转换,存在零权重,即权重分布中的权重 w_{61} 、 w_{62} 和 w_{64} 为零。这使RNN提供有关甲基化C碱基的更好信息,从而提高了对甲基化C碱基的估计的准确度。

[0207] 任选地,从 C^M 到 C^M 的转换的权重可以为零。这可以是在停留的情况下。这是因为尽管可以在测量期间将其鉴定出来,但由于CpG甲基化导致甲基化C总是在G之前,即序列 C^MG ,所述转换不形成序列的一部分。这一点的另外的实例是下文所描述的正位-翻转表示,其中允许从经修饰正位或经修饰翻转到鸟嘌呤或经修饰翻转标签的转换,从而将RNN 50所需的权重数从60减少到52(参考:所有可能的转换需要100个权重)。除了减少所需的网络输出量之外,将转换限制到可能的转换防止所述方法在不可能的背景中产生带有修饰的类型的聚合物单元的估计值,这既是估计值中的误差,又是错误的肯定修饰调用。

[0208] 第二实例是其中所述一组标签被修饰的表示,使得每种类型的聚合物由多个标签而不是单个标签表示。例如,所述一组标签可以包含关于每种类型的聚合物单元的第一标签和第二标签,其中所述第一标签表示所述类型的聚合物单元的实例的开始,并且所述第二标签表示所述类型的聚合物单元的所述实例中的停留。如上文所提及的,“停留”表示其中所述方法确定与连续权重分布相关的标签不发生变化的情况,可以将其视为对应于聚合物单元的相同实例的两个权重分布。在本文中,此实例将被称为“多停留”。由于停留由不同的标签表示,因此这改善了表示。这改善了聚合物单元的估计准确度。

[0209] 这种多停留表示的结果是允许一些转换,并且一些转换是不允许的。例如,关于相同类型的聚合物单元,仅允许第一标签例如仅允许“A”转换为第二标签,例如关于相同类型的聚合物单元的 A^S ,或转换为不同类型的聚合物单元的第一标签。更具体地,以下转换是允许的和不允许的:

[0210] a) 允许任何其它类型的聚合物单元从每个第一标签转换到第一标签并且允许相同类型的聚合物单元从每个第一标签转换到第一标签;

[0211] b) 允许相同类型的聚合物单元从每个第一标签转换到所述第二标签;

[0212] c) 不允许任何其它类型的聚合物单元从每个第一标签转换到所述第二标签;

[0213] d) 允许相同类型的聚合物单元从每个第二标签转换到所述第一标签或允许任何

其它类型的聚合物单元从每个第二标签转换到所述第一标签；

[0214] e) 允许相同类型的聚合物单元从每个第二标签转换到所述第二标签；并且

[0215] f) 不允许任何其它类型的聚合物单元从每个第二标签转换到所述第二标签。

[0216] 可以与图11的方案类似的方式，将上文说明性的多停留表示方案视为“强制性”方案，其中如上文所描述的，空白可以被认为是强制性的或任选的。这样，应当理解，对于多停留表示，可以设想与上文类似的方案，其中允许第一标签转换为相同的第一标签。此类方案可以被认为是“任选的”多停留表示方案。

[0217] 图13展示了权重分布51的实例，所述权重分布改编自图10的权重分布以实施这种类型的表示。因此，在图13中，所述一组标签包含关于示出为A、C、G和T的四种类型的碱基的四个第一标签，以及关于示出为 A^S 、 C^S 、 G^S 和 T^S 的四种类型的碱基的四个第二标签。在本文中，上标S（用于“停留”）用于区分关于相同类型的碱基的第二标签和第一标签，并表示停留。如图13所示，鉴于允许和不允许的转换，以下权重存在或为零：

[0218] a) 允许任何其它类型的聚合物单元（例如，C、G和T）从每个第一标签（例如，A）转换到第一标签并且允许相同类型的聚合物单元从每个第一标签（例如，A）转换到第一标签（例如，A），因此左上象限中的所有权重都存在；

[0219] b) 允许相同类型的聚合物单元标签（例如， A^S ）从每个第一标签（例如，A）转换到第二标签，因此右上象限中的 w_{15} 、 w_{26} 、 w_{37} 和 w_{48} 都存在；

[0220] c) 不允许任何其它类型的聚合物单元从每个第一标签（例如，A）转换到第二标签（例如， C^S 、 G^S 和 T^S ），因此右上象限中除了 w_{15} 、 w_{26} 、 w_{37} 和 w_{48} 的权重都为零；

[0221] d) 允许相同类型的聚合物单元从每个第二标签（例如， A^S ）转换到第一标签（例如，A）或允许任何其它类型的聚合物单元（例如，C、G和T）从每个第二标签转换到第一标签，因此左下象限中的所有权重都存在；

[0222] e) 允许相同类型的聚合物单元从每个第二标签（例如， A^S ）转换到第二标签（例如， A^S ），因此右下象限中的 w_{55} 、 w_{66} 、 w_{77} 和 w_{88} 都存在；并且

[0223] f) 不允许任何其它类型的聚合物单元从每个第二标签（例如， A^S ）转换到第二标签（例如， C^S 、 G^S 和 T^S ），因此右下象限中除了 w_{55} 、 w_{66} 、 w_{77} 和 w_{88} 的权重都为零。

[0224] 可以将多停留表示与上文所阐述的甲基化C的表示组合，或者实际上与总是以聚合物单元的已知序列出现的一种类型的聚合物单元的任何类似表示组合。

[0225] 现在将考虑均聚物的表示。均聚物是一系列聚合物单元中的相同类型的聚合物单元的连续实例的序列。

[0226] 均聚物通过上文所讨论的多停留表示适当地表示，因为对于相同类型的聚合物单元（例如，A），从第二标签（例如， A^S ）到第一标签的转换表示相同类型的聚合物单元的第二实例。例如，一系列标签 $AA^S A^S AA^S AA^S A^S AA^S A^S A^S$ 表示长度为四个聚合物单元的均聚物，连续标签A或 A^S 的数量是任意的并且在实践中变化。然而，可以通过调整表示来提高估计的准确度，使得标签表示呈编码形式的均聚物，例如如下。

[0227] 呈编码形式的均聚物的第一表示将被称为“正位-翻转”表示并且如下。

[0228] 使分析方法的输出为重叠的固定长度片段的好处之一是，可以使用重叠量来确定是否发生了聚合物单元的易位以及发生了多少易位。依赖重叠的分析方法在聚合物（如均聚物）的低复杂度区域中失败，其中重叠可以是不明确的（例如，AAA-→AAA可以为A同聚物的

零个、一个、两个或更多个易位),并且不同的表示是期望的。在正位-翻转表示中,对于每种类型的聚合物单元通过包含多个标签,标签表示均聚物,其中用于每种类型的聚合物单元的所述多个标签表示所述一系列聚合物单元中所述类型的聚合物单元的连续实例。通常,每种类型的聚合物单元有两个标签,为便于参考,可以将所述标签称为“正位”和“翻转”。

[0229] 因此,与解码到固定长度片段相反,解码的正位-翻转方法将聚合物单元的序列表示为“正位”和“翻转”标签的序列,所述标签具有以下限制:均聚物必须以“正位”标签开始,并且然后在“正位”和“翻转”标签之间交替,直到它们终止。正位-翻转表示确保没有标签与其相邻标签相同,因此一个单元与均聚物的易位(从正位到翻转的改变,反之亦然)总是有区别于没有易位的(正位到正位或翻转到翻转)。举例来说,所述一系列的聚合物单元CAATACCTTTAAAAAAGAACTTTTAGCTC被表示为 $CAA^F TACC^F TT^F TAA^F AA^F AA^F GAA^F ACTT^F TT^F AGCTC$,其中聚合物单元X的正位标签由X表示并且对应的翻转标签由 X^F 表示。在正位-翻转编码下,始终可以将一个易位与没有易位区分开来;大量聚合物单元的易位可能仍然是不明确的。因此,就由连续权重分布51表示的标签而言,如果碱基A的两个标签是A(正位)和 A^F (翻转),则一系列标签 $AAAAAA^F A^F A^F AAA$ 表示长度为三个聚合物单元的均聚物,连续的标签A或 A^F 的数量是任意的并且在实践中变化。对于每种类型的聚合物单元,原则上可以有两个以上的标签,但是两个标签是足够的。

[0230] 每种类型的聚合物单元的多个标签可以具有预先确定的循环顺序。在每种类型的聚合物单元的两个标签(正位和翻转)的实例中,预先确定的循环顺序可以是第一聚合物单元总是正位,然后正位和翻转交替。因此,预先确定的循环顺序允许标签之间的一些转换,并且预先确定的循环顺序不允许标签之间的其它转换。在权重分布中,存在通过预先确定的循环顺序不允许的转换的零权重,而当然,存在通过预先确定的循环顺序允许的转换的权重。

[0231] 在预先确定的循环顺序是第一聚合物单元始终是正位,然后正位和翻转交替的上述实例中,不允许从任何给定类型的聚合物单元的正位到任何其它类型的聚合物的翻转的转换,并且类似地,不允许从任何给定类型的聚合物单元的翻转到任何其它类型的聚合物的翻转的转换。

[0232] 图14展示了用于这种类型的正位-翻转表示的权重分布51的实例。因此,在图14中,所述一组标签包含关于示出为A、C、G和T的四种类型的碱基的四个第一标签(正位),以及关于示出为 A^F 、 C^F 、 G^F 和 T^F 的四种类型的碱基的四个第二标签(翻转)。如图14所示,鉴于允许和不允许的转换,以下权重存在或为零:

[0233] a) 允许所有类型的聚合物单元从每个第一标签(正位,例如,A)到第一标签(例如,A、C、G和T)(正位)的转换,因此左上象限中的所有权重都存在;

[0234] b) 允许相同类型的聚合物单元(翻转,例如, A^F)从每个第一标签(正位,例如,A)到第二标签的转换,因此右上象限中的 w_{15} 、 w_{26} 、 w_{37} 和 w_{48} 都存在;

[0235] c) 不允许任何其它类型的聚合物单元从每个第一标签(正位,例如,A)到第二标签(例如, C^F 、 G^F 和 T^F)的转换,因此右上象限中除了 w_{15} 、 w_{26} 、 w_{37} 和 w_{48} 的权重都为零;

[0236] d) 允许所有类型的聚合物单元从每个第二标签(翻转,例如, A^F)到第一标签(正位,例如,A、C、G和T)的转换,因此左下象限中的所有权重都存在;

[0237] e) 允许相同类型的聚合物单元从每个第二标签(翻转,例如, A^F)到第二标签(翻

转,例如, A^F)的转换,因此右下象限中的 w_{55} 、 w_{66} 、 w_{77} 和 w_{88} 都存在;并且

[0238] f) 不允许任何其它类型的聚合物单元从每个第二标签(翻转,例如, A^F)到第二标签(翻转,例如, C^F 、 G^F 和 T^F)的转换,因此右下象限中除了 w_{55} 、 w_{66} 、 w_{77} 和 w_{88} 的权重都为零。

[0239] 根据相对于聚合物单元的易位速度进行测量的速率,当聚合物在测量之间易位多次时,可以观察到一个以上单元的明显易位。在可能发生的情况下,可以添加每个聚合物单元的另外的多余标签(“flap”、“flup”、“flep”等),使得可以表示存在另外的单元,例如,从正位到flap的序列意味着中间翻转标签的存在。

[0240] 呈编码形式的均聚物的第二表示将被称为游程长度编码的表示并且如下。

[0241] 正位-翻转表示可以通过长均聚物进行调用,但必须作为交替标签的路径并进行多次连接调用。对于较长的均聚物,所观察到的信号变平可能意指不再存在清晰的时间,因为聚合物相对于纳米孔易位时,信号发生变化,并且标签中每个变化的位置变得更加任意。图15示出了示例区域的这种特异性丧失的实例,在所述示例区域中,权重在T-正位或T-翻转之间分配,尽管两者的累积证据都很高。

[0242] 因此,与将均聚物表示为交替标签的序列相反,替代地整个均聚物可以由关于聚合物单元的类型标签表示。因此,与训练RNN 50来调用典型序列或其正位-翻转编码相反,训练RNN 50来调用一系列聚合物单元的游程长度编码。例如,典型序列TAATTCAAACCTTTTTTCTGATAAGCTGGT的游程长度编码为 $TA^2T^2CA^3CT^7CTGATA^2GCTG^2T$,其中游程长度遵循碱基,并且一个的长度是隐含的。始终采用尽可能长的游程,因此没有游程与具有相同碱基的游程相邻。

[0243] 在游程长度编码表示的第一表达方式中,标签包含每种类型的聚合物单元的不同游程长度的标签。图16展示了此类权重分布的实例。在此实例中,关于示出为A,C,G和T的四个碱基以及关于示出为 A^2 、 A^3 等的每个碱基的均聚物存在单个标签。这是难处理的,因为存在大量标签以适应所有可能的均聚物长度,并且允许所有转换,除了从关于一种类型的碱基均聚物的标签转换到相同类型的碱基但长度不同的均聚物之外,因此,关于标签之间的大多数转换的权重 w_{ij} 的数量几乎等于标签数量的平方(其它可能的转换方案可以可替代地实施)。

[0244] 大基因组中的长均聚物发生的频率比偶然预期的要高,因此表示常规测序期间可能遇到的所有均聚物长度所需的标签数量非常高。由于网络输出的权重明确地参数化了均聚物标签之间的转换,因此训练数据成为问题,这既是因为需要训练的参数数量众多,又是因为它们耦接较弱。将标签的标签混排(例如, $A6 \rightarrow A3$ 、 $T2 \rightarrow T7$ 、 $G8 \rightarrow G1$)会产生等效模型,可以对所述等效模型进行训练以达到相同的性能,因此训练长度为4和6的均聚物的实例不会告知所述模型长度为5的那些均聚物。

[0245] 游程长度编码的替代性和优选表达方式是将权重分布51因子分解为若干个相关分布。因此,标签包含关于每种类型的聚合物单元的标签,并且权重分布51除了关于转换的权重之外,包括对于每种类型的聚合物单元的游程长度压缩的均聚物的可能长度上的另外的权重。RNN 50发出转换权重,以描述游程长度压缩序列的分布,即所有长度下降的游程长度编码序列,以及在给定聚合物单元的情况下针对游程长度的一组单独的条件分布。

[0246] 在游程长度编码的此优选表达方式中,由RNN输出的权重分布51可以包含呈图10所示出形式的权重,以表示不同类型的聚合物单元之间的转换。如上文所讨论的,在此情况

下,从标签到相同标签的一系列转换表示一系列任何数量的聚合物单元(即具有任何长度的单个聚合物单元或相同类型的聚合物单元的均聚物)的实例。

[0247] 作为对此(游程长度编码的此优选表达方式)的替代方案,可以在一组标签上定义RNN输出的权重分布51,其中每种类型的聚合物由第一标签和第二标签而不是单个标签表示,例如关于第一类型的聚合物单元的标签A和A^H。因此,上标H用于关于相同类型的聚合物单元区分第二标签与第一标签并且有效地表示“保持(hold)”。

[0248] 除了以下之外,这类似于图13中所示出的多停留表示。如上文所描述的,在多停留表示中,允许从第二类型的标签到第一类型的标签的转换(例如,从A^S到A),并且表示出现了相同类型的聚合物单元的另外的实例。结果,均聚物由一系列标签表示,其中所述第一类型的所述标签被重复,如在上文实例中AA^SAA^SAA^SAA^SAA^SAA^SAA^S表示长度为三个聚合物单元的均聚物。与之相反,在当前表示中,允许的转换不同,使得不允许从第二类型的标签到第一类型的标签(例如,从A^H到A)的转换。结果,一种类型的聚合物单元的单个实例和相同类型的聚合物单元的任何长度的均聚物全部由一系列标签表示,所述一系列标签包括第一类型的标签和任意数量的所述第二类型的标签。例如,AA^HAA^HAA^HAA^HAA^HAA^HAA^H可以表示单个碱基A或碱基A的均聚物。具体地,这通过以下方式实现:

[0249] a) 允许任何其它类型的聚合物单元从每个第一标签转换到第一标签,但不允许相同类型的聚合物单元从每个第一标签转换到第一标签;

[0250] b) 允许相同类型的聚合物单元从每个第一标签转换到所述第二标签,

[0251] c) 不允许任何其它类型的聚合物单元从每个第一标签转换到所述第二标签;

[0252] d) 不允许相同类型的聚合物单元从每个第二标签转换到所述第一标签;

[0253] e) 允许任何其它类型的聚合物单元从每个第二标签转换到所述第一标签;

[0254] f) 允许相同类型的聚合物单元从每个第二标签转换到所述第二标签;并且

[0255] g) 不允许任何其它类型的聚合物单元从每个第二标签转换到所述第二标签。

[0256] 图17展示了此类权重分布51的实例,所述权重分布改编自图10的权重分布以实施这种类型的表示。因此,在图17中,所述一组标签包含关于示出为A、C、G和T的四种类型的碱基的四个第一标签以及关于示出为A^H、C^H、G^H和T^H的四种类型的碱基的四个第二标签。如图17所示,鉴于允许和不允许的转换,以下权重存在或为零:

[0257] a) 允许任何其它类型的聚合物单元从每个第一标签(例如,A)转换到第一标签(例如,C、G和T),但不允许相同类型的聚合物单元从每个第一标签(例如,A)转换到第一标签(例如,A),因此除了为零的w₁₁、w₂₂、w₃₃和w₄₄之外,两个左象限中的权重都存在;

[0258] b) 允许相同类型的聚合物单元从每个第一标签(例如,A)转换到第二标签(例如,A^H),因此右上象限中的权重w₁₅、w₂₆、w₃₇和w₄₈存在;

[0259] c) 不允许任何其它类型的聚合物单元从每个第一标签(例如,A)转换到第二标签(例如,C^H、G^H、T^H),因此,右上象限中除了w₁₅、w₂₆、w₃₇和w₄₈之外的权重都为零;

[0260] d) 不允许相同类型的聚合物单元从每个第二标签(例如,A^H)转换到第一标签(例如,A),因此,左下象限中的权重w₅₁、w₆₂、w₇₃和w₈₄为零;

[0261] e) 允许任何其它类型的聚合物单元从每个第二标签(例如,A^H)转换到第一标签(例如,C、G和T),因此,左下象限中除了w₅₁、w₆₂、w₇₃和w₈₄之外的权重都存在;

[0262] f) 允许相同类型的聚合物单元从每个第二标签(例如,A^H)转换到第二标签(例如,

A^H), 因此, 右下象限中的权重 w_{55} 、 w_{66} 、 w_{77} 和 w_{88} 存在; 并且

[0263] g) 不允许任何其它类型的聚合物单元从每个第二标签 (例如, A^H) 转换到第二标签 (例如, C^H 、 G^H 、 T^H), 因此, 右下象限中除了 w_{55} 、 w_{66} 、 w_{77} 和 w_{88} 之外的权重都为零。

[0264] 因此, 给定类型的聚合物单元的一系列标签总是以第一标签的单个实例, 然后以第二标签的一个或多个实例开始。例如, 一系列标签 A 、 AA^H 、 $AA^H A^H$ 等 (具有任意数量的标签 A^H) 中的任何一种表示一系列任何数量的聚合物单元实例 (即, 具有任何长度的单个聚合物单元或相同类型的聚合物单元的均聚物)。

[0265] 如上文所提及的, 图10的实例不提供均聚物的良好表示, 并且图17的实例也是如此。然而, 均聚物由在游程长度压缩的均聚物的可能长度上的另外的权重表示。现在将描述此类另外的权重的若干种概率, 所述另外的权重中的每个都可以与呈图10的形式或图17的形式的权重组合应用。

[0266] 另外的权重的第一概率是, 对于每种可能类型的聚合物单元, 它们包括在一组均聚物的可能长度上的权重的分类分布。可能长度是一个类别并且RNN 50输出为每个类别分配权重。通常, 每个类别可以表示单个均聚物长度, 或者类别的一些或全部可以表示一定范围的均聚物长度。类别可以包含表示大于给定长度的所有均聚物的类别。类别不必均匀间隔。

[0267] 图18示出了根据此第一概率的此类另外的权重的实例。在此实例中, 对于四个碱基A、C、G、T的每一个的每个可能的长度, 存在权重 l_{ij} , 碱基由指数i指数化, 并且长度由指数j指数化。在此实例中, 每个类别对应于单个长度, 但是可替代地, 每个类别可以对应于一定范围的长度以减少类别的数量。图18中所示出的另外的权重连同用于标签之间的转换的权重一起形成权重分布51的一部分, 其可以采用如上文所描述的形式, 例如, 如图10到13中的任何一个所示出的。

[0268] 与完全指定所有均聚物标签之间的转换相比, 分类分布需要更少的参数, 并且允许估计潜在的游程长度压缩的基因组, 但仍然存在弱耦接的问题, 所述问题会导致训练数据使用率低, 并且使长均聚物难以训练。

[0269] 另外的权重的第二概率是, 对于每种可能类型的聚合物单元, 它们包括在均聚物的可能长度上的参数化分布的参数。此类参数可以用于计算给定聚合物单元的均聚物为任何给定长度的概率。

[0270] 图19示出了根据此第二概率的此类另外的权重的实例。在此实例中, 对于四种类型的碱基中的每种 (示出为A、C、G、T并且由指数i指数化) 存在权重 p_{ij} 。权重指示分布的j个参数 P_1 、 P_2 、 $\dots$ 、 P_j , 这些参数由指数j指数化。参数可以是表示分布的任何参数。通常, 根据分布, j可以具有任何复数值。图19中所示出的另外的权重连同用于标签之间的转换的权重一起形成权重分布51的一部分, 其可以采用如上文所描述的形式, 例如, 如图10到13中的任何一个所示出的。

[0271] 举例来说, 图20给出了分别由两个参数 (均值和方差) 的不同值表示的均聚物长度的两种不同分布的实例。

[0272] 在均聚物长度上使用参数化分布的优势在于, 分布可以解释为均聚物长度的后验分布, 从而将置信度置于估计的长度中。例如, 在图20中, 两个分布都给出了均聚物长度的相同后验平均估计值, 但是却给出了不同的置信度, 方差较大 (左) 的分布的置信度比方差

较小(右)的分布的置信度低。

[0273] 由于对不同均聚物长度的预测都是通过同一组网络输出进行的,因此它们比以前更加紧密地耦接在一起,并且允许网络将一种均聚物的实例推广到具有类似长度的那些。

[0274] 许多不同的概率分布可以与网络的输出结合使用。选择能够表示可能出现的任何均聚物长度的分布是有利的,因此所述分布应该支持较大或甚至半无限的潜在长度集。还期望存在表示给定均聚物长度中高置信度(低方差)和低置信度(高方差)的参数值。可以使用负二项式或几何分布,并且无法区分高置信度情况和低置信度情况。

[0275] 几何分布的方差是均值的函数,负二项式具有另外的自由度,并且其方差必须始终大于均值。满足这些标准的分布可以通过离散化支持 $[0, \infty]$ 的连续分布来找到。离散化的一种方式是将长度为L的均聚物的概率设置为从L到L+1,可替代地L-0.5到L+0.5的密度函数的积分,其中适当的处理为L=0。

[0276] 优选地,离散的分布具有显式的累积密度函数。此类密度的实例是但不限于威布尔分布(Weibull distribution)、对数逻辑分布、对数正态分布、 γ 分布。如果存在使用的参数分布或其离散对应项的均值、众数和方差的明确表达式,则是有利的,但不是必需的。

[0277] 图21展示了用于表示均聚物长度的一些合适的离散分布,所有都支持 $\mathbb{N}$ 。 $\Gamma(\alpha)$ 是 γ 函数, $\gamma(\alpha, \beta)$ 是不完全 γ 函数,并且 $\Phi(x)$ 是标准正态分布的累积分布。

[0278] 在第一概率和第二概率中的每一个中,对于聚合物单元的每种可能类型,即可能类型的均聚物的聚合物单元,定义另外的权重。尽管这是有效的,但是可以通过修饰来提供进一步的改进,其中(a)对于给定聚合物单元的类型和先前聚合物单元的类型的可能对、(b)对于给定聚合物单元的类型和后续聚合物单元的类型的可能对或(c)对于给定聚合物单元的类型、先前聚合物单元的类型和后续聚合物单元的类型的可能三联体的另外的权重。

[0279] 通过此修饰,权重采用相同的形式,例如,根据第一概率,在一组均聚物的可能长度上的权重的绝对分布,或者根据第二概率,在均聚物的可能长度上的参数化分布的参数,但是增加了权重数量。对于情况(a)和(b),权重数量增加了三倍,以便为每个可能对而不是聚合物单元的每种可能类型定义分布,例如对于12对碱基{(A,C)、(A,G)、(A,T)、(C,A)、(C,T)、(C,G)、(G,A)、(G,C)、(G,T)、(T,A)、(T,C)、(T,G)}而不是4种类型的碱基{A,C,G,T}。举例来说,图22示出了此类另外的权重的实例,所述另外的权重包括对于每对类型的聚合物单元定义的均聚物的可能长度上的参数化分布的参数。这对应于情况(a)和(b),在情况(a)下,所述对是给定聚合物单元和先前聚合物单元的类型,并且在情况(b)下,所述对是给定聚合物单元的类型和后续聚合物单元的类型。参数本身的形式与图19相同,并且可以相同的方式用于计算给定聚合物单元的均聚物为任何给定长度的概率。

[0280] 类似地,对于情况(c),权重数量增加了九倍,以便为每个可能的三联体定义分布,例如为36个三联体的碱基而不是4种类型的碱基。举例来说,图23示出了此类另外的权重的实例,所述另外的权重包括对于每个三联体类型的聚合物单元定义的均聚物的可能长度上的参数化分布的参数。这对应于情况(c),三联体是给定聚合物单元、先前聚合物单元的类型和后续聚合物单元的类型。参数本身的形式与图19相同,并且可以相同的方式用于计算给定聚合物单元的均聚物为任何给定长度的概率。

[0281] 基于区分长均聚物的边缘的能力可以根据先前和/或后续聚合物单元而变化的认

识,此修饰提高了准确度。例如,从碱T到碱A的均聚物的转换比从碱C到碱A的均聚物的转换更容易区分。因此,提供表示各种对或三联体的分布的不同的另外的权重会提供可以更准确地估计聚合物单元的代表。

[0282] 将权重分布51类似地因子分解为若干个相关分布可以用于表示聚合物的其它性质。一个实例是具有未修饰形式和经修饰形式的一种类型的聚合物单元的代表,例如可以包含一种类型的碱基和同一碱基的修饰类型的多核苷酸。

[0283] DNA的天然链含有修饰的碱基,例如5-甲基胞嘧啶或6-甲基腺嘌呤,并且可使用一系列纳米孔测量来检测它们的存在和位置。正位-翻转和其它表示很容易概括为能够通过从碱基A、C、G和T扩展一组标签以包含表示修饰的碱基的另外的标签(例如, C^M 表示修饰C)来调用修饰。

[0284] 图24示出权重分布的实例,其中一组标签被扩展以另外包含关于修饰的碱基的标签 C^M 。类似地,可以将另外的标签 C^M 添加到图10、12到14或16所示出的权重分布51中的任何权重分布中的一组标签。

[0285] 标签的字母的这种扩展还可以与本领域中描述的先前方法一起使用,所述方法假定在特定时间的信号可以由固定长度的碱基片段表示,但是由于网络必须针对固定长度的碱基之间的每个可能的转换具有输出,因此随着所考虑的修饰数量的增加,这些标签缩放较差。例如,对于长度为5的片段,存在1024 (4^5) 个可能组合,其由四个典型碱基组成,如果允许另外的修饰的碱基,则为3125 ($=5^5$),并且如果允许两个修饰,则为7776 ($=6^5$)。RNA中已知有一百多种修饰,因此基于片段的模型需要快速增加的处理量。

[0286] 聚合物单元的未修饰形式可以描述为典型聚合物单元,并且聚合物单元的修饰形式可以描述为非典型聚合物单元。修饰(或非典型)聚合物单元通常影响与对应未修饰(典型)聚合物单元不同的信号。

[0287] 通过引用并入本文的2019年9月4日提交的国际专利申请号PCT/GB2019/052456含有与可以应用于本文所公开的本发明方法中的任何一个的典型和非典型碱基有关的教导。

[0288] 国际专利申请号PCT/GB2019/052456公开了可以应用于本发明方法中的任何一个的非典型碱的实例。

[0289] 国际专利申请号PCT/GB2019/052456还公开了制备和分析包括一种或多种非典型聚合物单元的聚合物的方法,其可以与本发明方法中的任何一个组合使用。

[0290] 通过非限制性实例,国际专利申请号PCT/GB2019/052456中公开的可以与本发明本发明方法中的任何一个组合的一种方法以不确定性的方式,例如通过化学转化或通过酶促转化,将一定比例的典型聚合物单元(例如,氨基酸)转化为对应的非典型聚合物单元(例如,氨基酸)。在所述情况下,当导出一系列聚合物单元的估计值(“调用”)时,可以将非典型碱基估计(“调用”)为对应的典型碱基。这包含参考以下的图18b-18k所描述的方法:国际专利申请号PCT/GB2019/052456。

[0291] 由于非确定性地将典型和非典型聚合物单元并入靶聚合物中,因此聚合物单元的基础序列是未知的,并且将在链到链的基础上变化。即使每个链含有替代性聚合物单元,但是仍然存在相关的典型序列,并且所关注的是直接调用这而不是尝试推断任何替代方案的类型和位置。换言之,尽管在靶聚合物中存在另外的聚合物单元,但分析仅将典型值归属于信号,使得所测定的序列由来自A、C、G和T的组的碱基组成。以此方式,通过在分析中将非典

型聚合物单元识别为典型聚合物单元,初始转换可以提供一种提供具有更多信息的信号的方式,例如,结果是信号分析中出现的任何错误将是非系统性的,从而带来估计准确度的提高。

[0292] 正位-翻转和类似的表示更易于处理,因为在每个时间点从RNN 50输出的权重数量需要使用修饰数量而不是等于片段长度的幂对转换权重标度进行二次参数化(对于4个典型碱基40个输出,对于一个另外的修饰碱基60个输出,对于两个碱基84个输出等)。

[0293] 当神经网络10使用正位-翻转表示时,执行训练以使正确序列的概率最大化,对于每次读取,其产生条件随机场,必须对所述条件随机场进行进一步解码以产生估计序列。所使用的解码方法可能会在最终调用中引入不必要的偏差,所述偏差会以大量量度显示自身,例如称为其组成的读段或汇总统计信息的碱基总数。当将来自具有相同序列或含有共同子序列的链的读段的估计序列综合考虑时,另外的偏差可能明显。

[0294] 为了减少这个问题,可以将惩罚项并入到训练的神经网络10中,调整其输出以提高所关注的量度的性能:例如,从对应于不发射新的聚合物单元(相同碱基中的正位-正位或翻转-翻转转换)的所有权重中减去常数将增加调用的聚合物单元的数量,而通过将常数添加到结束发射具有所述标识的新的聚合物单元的所有转换,可以增加特定聚合物单元的比例。

[0295] 可以通过针对网格值上的一组代表性读段计算所关注的量度来调整使用的惩罚项的值,可替代地可以使用更正式的优化方法(例如,单纯形法或本领域已知的许多其它方法)。惩罚项可以是关于读段的先验信息的函数,而不是固定的常数。

[0296] 可以在任何层将惩罚项并入到神经网络10中,但是优选的是在可能的情况下将它们并入到最终层中,直接影响所发射的转换权重,因为这样做的优点是可以直观地影响最终的估计序列,从而指导惩罚的形式。

[0297] 为了保持将神经网络10的输出解释为概率模型,期望但并非必须的是,在执行“全局归一化”之前并入惩罚。

[0298] 通常,准确地确定典型碱基的序列和任何修饰的存在都是所关注的,并且不期望尝试估计修饰对基础典型序列的估计产生不利影响。这如何发生的一个实例是典型胞嘧啶与5-甲基胞嘧啶之间的权重分割,因此另一种碱成为最可能的估计。

[0299] 为了防止权重分割行为,可以将RNN输出的权重分布51因子分解成两个相关分布。在此情况下,第一分布是权重分布51,所述权重分布采用上文所描述的形式中的任何一种,表示具有未修饰形式和经修饰形式的类型的聚合物单元的单个标签,并且第二分布是条件分布,所述条件分布包括未修饰形式和修饰形式的另外的权重。对于任何数量的修饰形式和可能类型的聚合物单元中的任何一种的修饰形式可以扩展此表示。

[0300] 图25示出了用于表示碱基C的未修饰形式和相同碱基 C^M 的修饰形式的另外的权重的实例。在此情况下,另外的权重是关于碱基C的未修改形式的权重 m_1 和针对碱基 C^M 的修饰形式的权重 m_2 。这可以代替图24所示出的类型的权重分布51而被应用。另外的权重连同用于标签之间的转换的权重一起形成权重分布51的一部分,其可以采用如上文所描述的形式,例如,如图10到14或16中的任何一个所示出的。

[0301] 这种因子分解式表示意指可以确定典型序列,就好像不存在修饰一样,然后可以确定任何修饰的位置。修饰的条件分布本身可能是因子分解的,可能反映了先前的生物学

期望。例如,一种分布可以表示胞嘧啶是否被修饰,并且另一种分布可以表示在存在修饰的情况下,所述修饰是5-甲基胞嘧啶还是5-羟甲基胞嘧啶。

[0302] 作为实例,图26示出了当采用已经以此方式扩展的四个碱基的正位-翻转表示来检测修饰的碱基5mC时,由RNN 50的输出预测的碱基。在此实例中,在与外部预测一致的位置处的三个位置处估计了修饰的碱基5mC。

[0303] 如上文所提及的,将RNN 50的权重分布全局归一化。此类全局归一化可以通过一系列权重分布在标签的所有路径上,使得所有可能路径上的总和是一。全局归一化可以在输出空间上,使得权重可以被认为是后验概率。

[0304] 全局归一化严格意义上比局部归一化更具表达性,并且避免了本领域中称为“标签偏差问题”的问题。

[0305] 使用全局归一化优于局部归一化的优点类似于条件随机场(Lafferty等人,《条件随机场:分段和标记序列数据的概率模型(Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data)》,《国际机器学习会议论文集(Proceedings of the International Conference on Machine Learning)》,2001年6月)具有优于最大熵马尔可夫模型(McCallum等人,《用于信息提取和分割的最大熵马尔可夫模型(Maximum Entropy Markov Models for Information Extraction and Segmentation)》,《ICML会议论文集(Proceedings of ICML)》,2000,591-598。斯坦福,加利福尼亚州,2000)的优点。标签偏差问题影响标签之间允许稀疏的转换矩阵的模型,如聚合物序列的扩展。

[0306] 全局归一化通过在整个序列上进行归一化来缓解这个问题,从而允许不同时间处的转换相互交换。全局归一化对于避免均聚物和其它低复杂性序列的偏差估计特别有利,因为与其它序列相比,这些序列可具有不同数量的允许的转换(取决于模型,其可以更多或更少)。

[0307] 现在将考虑解码器80。

[0308] 解码器80从所述权重分布51中导出所述一系列聚合物单元的估计值。这可以通过使用链结式时间分类来完成,例如在以下中公开的:Graves等人,“用递归神经网络标记未分割的序列数据的链结式时间分类(Connectionist temporal classification labelling unsegmented sequence data with recurrent neural networks)”,《第23届国际机器学习会议论文集(Proceedings of the 23rd international conference on Machine learning)》中,369-376(ACM,2006)。

[0309] 解码器80如下执行如图27所示出的三个步骤。

[0310] 在步骤S1中,导出关于相应权重分布51的标签的估计值。此估计在下文中进一步讨论。

[0311] 在步骤S2中,对在步骤S1中导出的标签进行游程长度压缩,以导出一系列聚合物单元的估计值(也可以被称为解码)。这是需要的,因为权重分布51比聚合物单元多。游程长度压缩产生聚合物单元的估计值,因为如上文所描述的,在RNN 50固有的聚合物表示中,相同标签的连续序列表示相同聚合物单元。

[0312] 步骤S2还考虑了其中多个标签用于表示给定类型的聚合物单元的代表。例如,在上文所描述的多停留表示中,第二标签被压缩成关于相同类型的聚合物单元的第一标签。

类似地,在上文所描述的正位-翻转表示中,第一标签(正位)的连续实例被压缩成单个聚合物单元,并且第一标签(翻转)的连续实例被压缩成另一个单个聚合物单元,依此类推,从而提供均聚物的估计值。

[0313] 例如,在图11的方案中,可以在步骤S2中执行解码空白以区分相同聚合物单元的实例。如上文所讨论的,对于空白可以考虑“任选”和“强制”方案,因此,步骤S2可以根据以下两个方案中的哪一个将标签序列: A AA--A 解码为 A AAA 或 A A。

[0314] 在正位-翻转方案的情况下,步骤S2可以包括将相同标签的多个游程折叠到单个对应的聚合物单元上。例如,可以在步骤S2中将标签序列 $CAA^F TACC^F TT^F$ 解码为一系列聚合物单元 CAATACCTT。

[0315] 关于多停留方案,步骤S2可以包括通过将相同标签的连续序列识别为相同类型的不同聚合物单元来进行解码。例如,可以在步骤S2中将标签序列 $AA^S A^S TT^S CAA^S A^S$ 解码为一系列聚合物单元 ATCA。

[0316] 关于游程长度编码方案,步骤S2可以包括通过折叠相同标签的游程来进行解码(并且如果在所述方案中必要则丢弃空白)。例如,在步骤S2中标签序列 $TA^2 T^2 CA^3$ 可以表示一系列聚合物单元 TAATTCAAA。

[0317] 在权重分布51被因子分解为相关分布的情况下执行步骤S3,否则省略。在此情况下,使用关于转换的权重执行步骤S1和S2,并且在步骤S3中,使用另外的权重来估计由此表示的聚合物单元的质量。例如,在上文所描述的游程长度编码的表示中,另外的权重用于估计均聚物的长度。类似地,在上文所描述的修饰形式的因子分解式表示中,另外的权重用于估计聚合物单元是未修饰形式还是修饰形式。

[0318] 现在将讨论步骤S1中的标签估计。由于权重表示相应转换的后验概率,因此权重可以用于通过权重分布51导出标签的任何给定路径的后验概率,即对于对应于所讨论的路径的一系列转换,通过组合由权重表示的后验概率。这意指权重允许考虑不同路径的可能性,从而提高了估计的准确度。因此,步骤S1应用了一种技术,所述技术基于对通过权重分布51关于标签路径的转换的组合权重的考虑。

[0319] 在不允许一个或多个转换的情况下(如上文所讨论的),解码器在步骤S1中执行的估计可以考虑表示标签之间的转换是否被允许或不允许的转换矩阵。

[0320] 两种不同的方法是可能的,所述方法将被称为“最佳路径”和“最佳标签”。

[0321] 在最佳路径方法中,基于权重分布51通过一系列权重分布51的最可能标签路径。在此情况下,在步骤S1中导出的关于相应权重分布51的标签是所述最可能路径的标签。

[0322] 由于权重分布51是转换上的权重,因此解码以估计序列的一种方法是找到具有最大权重之和的路径。此类路径可以有效的从转换权重中找到,例如使用动态程序算法。可以使用维特比算法(Viterbi algorithm)。

[0323] 例如,图28展示了最佳路径算法,其中在框i处,RNN 50将权重 w_{jk}^i 输出到从标签j到标签k的转换。向量 t^i 存储了回溯信息,所述回溯信息是来自给定当前标签的最佳标签,并且用于确定评分S和最佳路径P。

[0324] 对于游程长度编码,找到的最佳路径是针对游程长度压缩序列的,并且需要根据由RNN 50输出的适当的条件分布来确定每个游程的长度。在最佳路径示出了新的聚合物单元的情况下,可以根据对应于所述聚合物单元的条件分布来估计游程长度。进行此估

计的适当方法包含：找到条件分布的均值（四舍五入）、众数或中值；在合适的先验条件下，还可以使用具有最大贝叶斯因子 (Bayes factor) 的长度。在网络输出表示可能存在的可能的碱基修饰的条件分布的情况下，以它们的存在来标记最佳路径的过程以类似地进行，尽管后验均值和中值不是合理的估计量，因为修饰是分类的，而不是有序的。

[0325] 对于游程长度编码，可以应用游程长度偏差校正。由于所述模型是根据实际读段训练的，因此存在学习并且并入模型的权重中的游程长度的一些先前分布。对于从随机链或实际（例如，基因组）链导出的读段，训练数据将含有的不同长度的游程的比例存在显著的偏差，例如，长游程极为罕见。这影响了所述方法调用长游程的能力。游程的长度存在歧义，通常调用短游程比调用长游程更正确，并且因此单读段准确度的最大化趋于导致调用短游程。这样，对相对较短的游程长度应用偏差校正是有利的。

[0326] 找到最可能的路径后，必须导出一系列典型碱基。对于正位-翻转表示，合并标签的相邻重复，因为它们在其他CTC类模型中起与空白标签相同的间隔作用，然后去除每个标签的正位或翻转标识以保留典型碱基。对于游程长度编码，删除空白标签，并且将每个游程扩展为适当数量的碱基。

[0327] 现在将讨论最佳标签方法，注意到在正确的标签不在最可能的路径上的情况下，最佳路径方法可能会错误地估计一些特定的标签。来自RNN 50的权重分布51有效地定义了标签的所有可能路径上的概率分布，为位置分配标签的一致方法，并且每个路径对应于一系列标签，因此对应于聚合物单元，尽管这种对应关系不是唯一的（可能存在给出相同序列的许多路径）。最佳标签方法通过估计最可能的一系列标签（以及因此聚合物单元）来改进最佳路径方法。也就是说，通过对满足此条件的所有路径求和，而不是找到最佳路径，可以得出在时间步长 i 之后路径在标签 j 中的后验概率。这可以考虑通过一系列权重分布51的标签的前向和后向路径。在此情况下，在步骤S1中导出的关于相应的权重分布51的标签是由此最有可能导出的标签。

[0328] 动态的时间前向和后向编程允许使用类似于最佳路径的递归的递归以有效的方式执行此计算；在最佳路径算法可以看作是维特比解码的形式的情况下，后验概率的计算可以看作是前向和后向算法的形式。类似地，可以通过对满足此转换的所有路径求和来计算在时间步长 i 处标签发生变化的后验概率；此计算还可以以有效的方式执行。

[0329] 尽管后验概率提供了有关每个位置处可能的标签的信息，但通过选择最可能的标签进行解码可能会导致路径不一致，从而导致序列不一致。通过定义从一个标签到另一个标签的转换矩阵 T （根据是否允许转换，所述标签的条目为一或零），可以将最佳路径解码算法应用于这些后验概率，以从所有一致的路径中找到最大化其标签的后验概率之和的路径。

[0330] 作为此的实例，图29展示了应用于在位置 i 处在标签 k 中的后验概率 p_k^i 的算法。向量 t_i 存储了回溯信息，所述回溯信息是来自给定当前标签的最佳标签，并且用于确定评分 S 和最佳路径 P 。

[0331] 可替代地，可以将最佳路径算法应用于后验概率的对数，以找到在所有一致路径上最大化其标签的对数后验概率之和的路径。这等同于找到在所有一致路径上最大化其标签的后验概率乘积的路径。

[0332] 作为此的实例，图30展示了应用于在位置 i 处在标签 k 中的对数后验概率 p_k^i 的最佳

路径算法。向量 t_i 存储了回溯信息,所述回溯信息是来自给定当前标签的最佳标签,并且用于确定评分S和最佳路径P。

[0333] 可替代地,由于权重分布51是在转换上定义的,因此前向和后向算法可以用于计算位置之间转换的后验概率,而不是每个位置处的标签。

[0334] 作为此的实例,图31展示了计算在所有路径上求和的后验概率。由于这些权重是在转换上的,因此所述权重具有与转换矩阵相同的形状,并且它们的对数可以馈送到图28中定义的方程中而不是转换权重以找到一致路径。

[0335] 从覆盖基因组相同区域的许多信号中产生共有序列的更成功的方法之一被称为“抛光(polishing)”,并且已经在一些出版物中进行了描述。抛光共有序列是迭代过程,其中根据所有读段与候选变化的匹配程度,对草案共有序列的候选更改进行评分,并且保持较高的评分更改,允许由一种读段引起的错误由其它读段来校正;重复此程序,直到找不到更高的评分更改为止。

[0336] 不明显的是,抛光还可以有益地应用于单个读段。先前小节中所描述的所有估计聚合物单元的方法都旨在通过网络输出找到好的路径,可以从中提取出碱基序列,但是无配准培训目标针对给定序列的所有路径求和,而不是将单个路径标识为好。为了与训练标准一致,理想地,应该通过找到最可能的序列,对导致相同序列的所有路径而不是最可能的路径求和,来对来自RNN 50的输出进行解码。给定序列的所有路径的总和是抛光用来评估候选变化是否良好的标准,因此可以将抛光视为迭代启发法(贪婪爬山的变体),以找到最可能的序列。

[0337] 在分析作为相关的一系列聚合物单元的测量的多个测量系列的情况下,所述方法基本上是相同的,但是来自多个测量系列的测量被视为布置为多个相应的尺寸。这增加了维度,但是神经网络10的形式在其它方面与上文所描述的相同。在此情况下适用的一些另外的考虑如下。

[0338] 当使用惩罚项时,作为不发射的惩罚针对所有转换不变的替代方案,惩罚可以根据转换取不同的值或完全不存在。例如,一些转换不会导致状态发生变化,并且可以是自由的,或具有小的惩罚,因为所述转换并不意味着其它读段中丢失的状态。

[0339] 对于每个读段,所使用的一种或多种惩罚不必相同,并且对于两个读段可能具有不同的特性可能有好的生物物理原因。例如,一个读段可能来自在马达上方的双链的分子,而另一个则是单链的;可替代地,两个读段可以是具有不同马达的链;一个读段可以是DNA,而另一个可以是RNA;可替代地,两个读段可以是同一正向-反向-互补链的第一部分和第二部分,并且在测序期间两者之间的杂交改变了动力学。

[0340] 所使用的一种或多种惩罚可以与时间有关。所使用的一种或多种惩罚可以取决于读段的本地统计信息。这样的实例包含:速度、失速的存在或噪音。所使用的一种或多种惩罚可以取决于使用其它模型或技术的读段分析的输出,例如,预测滑动的可能性(缺失碱基)。

[0341] 正位-翻转表示和RLE表示的状态转换模型都具有时间顺序,并且颠倒状态的顺序可移不是有效的状态序列。也就是说,在RLE表示中,必须在停留之前发射碱基,并且正位-翻转表示要求任何重复的第一碱基必须是“正位”。这样的结果是,在读段之一来自链(或链的一部分)的情况下,所述链是另一条链的反向补体或反向链,在分析之前反转读段之一并

应用与两个正向读段相同的程序是不够的。

[0342] 虽然可以使用更复杂的程序组合两个不同方向上的读段,从而跟踪作为一对的两个读段的状态,但是有利的是在一个读段上使用标准模型并且在另一个读段上使用已被“后向”训练的模型——在训练期间,来自读段的信号和靶序列是相反的(并且可能是补充的)。使用此类一对模型确保正向和反向读段以相同的顺序通过模型的状态,因此可以将所述读段组合起来,就好像它们都是正向读段一样。

[0343] 可以使用常规技术来训练神经网络10,例如如下。

[0344] 神经网络10输出表示权重的分布,所述权重表示在标签的路径上的概率(用标签对度量的一致标记),然后将其解码为聚合物单元序列的估计值。用旨在确保此估计值具有低比例的误差的标准来训练神经网络10。

[0345] 使用转换权重定义路径上的概率分布的重要方面是,必须对权重进行归一化使得所有路径上的总和为1。给定一组转换权重,可以使用动态编程通过应用如上文所讨论的后验概率计算中使用的前向算法(或后向算法)来计算归一化因子。由于所有可能路径的总和,而不是在每个时间点将网络的输出归一化为1,因此此技术被称为全局归一化,并且确保每个路径的评分具有概率(其对数的)的解释。具有一致标签的每条路径对应于概率,并且这些概率形成了所有路径上的分布。

[0346] 与全局归一化相反,归一化神经网络10使得每个时间点的输出总和为1被称为局部归一化。可以计算每条路径的评分并且其具有概率的形式,但是由于总的概率质量小于1,因此它们不形成分布。局部归一化将概率分配给所有标签序列,无论它们是否形成一致路径。

[0347] 序列标记的训练需要训练实例,即输入信号对及其对应的标签序列,以及用于优化训练实例的目标函数。由于纳米孔测量结果与聚合物单元序列之间的真正配准是未知的,因此,无配准的训练方法如Graves等人所描述(2006)是优选的。如果配准培训方法需要对测量序列的每个元素进行标记,则免配准的方法仅需要知道聚合物单元的真实序列。可以通过在纳米孔装置中测量已知序列的聚合物,或将读段与参考序列或具有已知序列的一组测量结果进行比较,来确定用于读段的聚合物单元的真实序列。

[0348] 测量已知序列的实例可以包含小的基因组,其中有可能在单个读段(如 λ 噬菌体(50千碱基))中对完整的基因组进行测序。也可以使用限制性酶切,并且通过其长度鉴定片段。另一个实例涉及将已知片段按顺序添加到游程中,因此可以通过片段出现在数据中的时间来识别它们。显而易见的是,可以使用可以将序列分配给信号读段的任何方法。

[0349] 在训练神经网络10时,在各种背景下以及在各种实验中进行跨每个聚合物单元的测量是有益的,因此所网络已暴露于正常运行状况下将遇到的整个变化范围的大部分。理想地,神经网络10使用完整的读段来训练,即纳米孔读取的覆盖全长聚合物的信号和序列对。然而,出于实际考虑(计算时间、内存),通常在较小的信号和序列块上进行操作。

[0350] 递归、卷积和注意力神经网络单元具有时间顺序的概念,并且训练中呈现的测量窗口的大小限制了可以从中学习的背景。由于每个聚合物单元可能具有很大的影响范围,因此为神经网络10呈现较大的测量窗口以进行训练是有益的。所使用的窗口的大小是呈现足够大的一系列测量结果之间的平衡,以使神经网络10可以在孔隙、聚合物链和其它系统组件之间的相互作用以及可用的计算能力的数量上创建足够的内部表示。理想地,将使用

每个读段的全部内容,但在实践中,固定大小的度量块呈现良好的折衷方案。足够的块的大小取决于纳米孔以及链的易位速率,但是已证明,对应于约200到约300个碱基的块大小是足够的。例如,已证明这对于CsgG纳米孔是足够的。

[0351] 示例训练集大小可以包括~1百万组~300个信号和序列的碱基块。仅几千个块的较小训练集可能就足够了,并且>1百万个块的较大训练集可以为训练提供更多多样性。

[0352] 训练神经网络的许多技术或其它机器学习方法是本领域已知的,并且可以在此应用。由于所述方法概述不同的实验运行和聚合物序列的能力得益于大量的训练数据,因此寻求最大化目标函数方向通常是不切实际的,因为优选的是在内存受限的图形处理单元(GPU)或其它专用硬件上执行计算。与直接在整个数据集上最大化目标函数相反,优选的是使用随机梯度下降法(SGD)或相关技术,以迭代方式使用整个训练集的子集(“小批量”)来近似地最大化目标函数。优选的小批量大小取决于所使用的计算装置上的可用内存以及小批量的每个元素中的度量数量。

[0353] 随机梯度下降(SGD)的许多变体是本领域已知的,例如:SGD、具有动量的SGD、具有Nesterov动量的SGD、RMSprop、AdaMax、Adam。Adam的修饰“Adamski”,其中迭代N的动量以从0到最大值 μ 的动量斜坡因子 r 增加:其中 $\mu_N = \mu(1 - e^{-rN})$,是优选的。Adamski具有学习速率、两个平滑参数(在本领域中通常称为衰减1和衰减2)和动量斜坡速率。这些参数的许多选择都是有益的。优选的参数化具有初始学习速率 10^{-3} 、平滑参数0.9和0.999以及动量斜坡因子0.005。平滑参数0.95和0.99也已被证明对于完善已经训练的模型有效,因为初始学习速率已降至 10^{-4} 。

[0354] SGD和相关技术以迭代方式进行,每次迭代由以下步骤组成:

[0355] 1. 选择完整训练数据的子集。

[0356] 2. 计算此子集的目标函数

[0357] 3. 使用反向传播计算所有网络参数的梯度

[0358] 4. 使用SGD或变体更新网络参数

[0359] 5. 转到1(下一个迭代的开始)

[0360] 步骤4中更新的大小按称为学习速率的因子进行缩放。高学习速率意指参数可以快速变化,因此最大化可以更快地进行,但是每个小批量的影响可能较大,这意指模型接近收敛时,更新可以由小批量到小批量可变性来控制。优选的是缓慢降低迭代到迭代的学习速率;这种降低可以是动态的,根据目标函数在批量到批量之间的变化和可变性,或者根据某些预先确定的时间表来调整学习速率。优选地,使用双曲线衰减,其中对于某个初始学习速率 R 和小批量的数量 K ,第 N 个小批量的学习速率为 $R/(1 + (N/K))$ 。

[0361] 虽然已经使用求和来将小批量的每个成员的评分组合到小批量的评分中,但是其它组合方法也是可能的。求和得出的小批量评分与其组成元素的评分的均值成比例,对应于其它集中趋势度量的组合也具有良好的性质。如中值、修剪均值或加权均值等组合器或拟合M估计器的组合器可以用于更改物镜对具有异常值的小批量元素的敏感性。

[0362] 小批量的每个元素对总评分的作用是在所有一致路径上求和的真实序列的后验概率(的对数)。对于正位-翻转表示,正位到正位或翻转到翻转的转换表示停留在序列的相同位置,而所有其它转换涉及位置的移动。给定表示在每个时间点从RNN 50输出的标签之间的转换的权重,可以将这些权重转换为已知序列位置之间的转换权重。

[0363] 图32示出了对于标签 $S_1, S_2, \dots, S_N$ 的正位-翻转编码序列,如何构建每个时间点 i 的目标转换矩阵 m_i 的元素。图33中所描述的目标函数使用此目标转换矩阵来计算小批量的每个元素的评分。

[0364] 由于目标函数的转换矩阵非常稀疏,对角线(停留)和超对角线(在位置上移动)上仅具有非零元素,因此从真实序列的长度(从二次到线性)的角度,此计算的优选实施例仅忽略了零元素并且降低每个步骤的表现复杂性。

[0365] 多停留表示的目标函数在结构上与正位-翻转目标类似,但表示停留在相同位置的状态不同。从停留或非停留状态到任何非停留状态的转换意味着位置的改变;停留状态的任何转换不意味着位置的改变。对于多停留表示,表示停留在新位置(碱基到停留转换)和停留在旧位置(停留到停留转换)的转换是有区别的,并且目标函数的有效计算需要使用一组重复的真实序列的“剩余”位置: $S_1, R_1, S_2, R_2, \dots, S_N, R_N$ 。

[0366] 图34示出了对于此实例如何构建目标转换矩阵的元素。出于形成目标转换矩阵的目的,原始位置被列举为 $1 \dots N$,而对应的重复位置被列举为 $N+1 \dots 2N$ 。

[0367] 图33中所描述的目标函数使用此目标转换矩阵来计算小批量的每个元素的评分。目标转换矩阵是稀疏的,并且目标计算的优选实施例利用了这种稀疏性。

[0368] 每个评分可以乘以权重,然后用于目标函数,并且此权重可以表示对应于训练过程的小批量的元素的值。例如,对于具有异常序列组成的元素或已知与碱基调用误差有关的元素,权重可能更大,这可以在先前训练的网络的测试期间发现。确定小批量的元素的权重的一种方法是将其设置为等于其最稀有的均聚物的频率的倒数,所述频率是根据整个训练数据集或根据其它外部参考确定的。

[0369] 每当转换了新的序列位置以表示通过网络输出的对应条件分布对游程长度的预测程度如何,游程长度编码的目标的定义类似于多停留模型的目标除了并入的另外的因子。游程长度压缩序列上的目标转换矩阵的形式与许多停留目标的形式相同,在位置之间允许的转换中,没有碱基可以遵循隐含的相同碱基的限制,但是在给定其组成的每个位置处,网络分配给均聚物长度的对数概率的另外组成。

[0370] 当已知训练数据的均聚物含量存在偏差时,在许多应用中网络学习此偏差可能是不期望的,因为网络可能无法表示其它数据集。代替直接在训练目标中在给定其组成的每个位置处使用网络分配给均聚物长度的对数概率,它可以首先与另一个分布合并;可以通过从训练数据中列出均聚物的频率来获得这种其它分布(“训练先验分布”)。通过以此方式进行训练,网络必须学习分配克服对训练先验分布的期望的对数概率。

[0371] 出于碱基调用的目的,可以使用标准方法如贝叶斯定理(Bayes theorem)将训练数据的先验分布或均聚物长度的任何其它期望与网络分配的对数概率组合,以产生通过有关均聚物长度的外部信息得出新的对数概率;可替代地,可以将来自网络的对数概率直接用于无偏调用。

[0372] 图35示出了对于此实例如何构建目标转换矩阵的元素。将网络在测量的时间步长 i 处分配给长度为 L_j 的游程长度的概率的对数为 $r^i S_j : L_j$,游程具有序列位置 j 的组成 S_j 。图33中所描述的目标函数使用此目标转换矩阵来计算小批量的每个元素的评分。

[0373] 尽管以免配准的方式训练碱基调用模型的优点众多,但在呈现的大多数解码算法与使用的训练目标之间存在断开。用于模型训练的目标函数最大化真实碱基序列的概率,

对可以表示其的所有各个路径的概率求和,而除如上文所描述的抛光外的所有解码例程都在寻找具有高评分的路径。图36示出了此断开连接所引起的问题之一。具体地,对于正位-翻转表示的实例,图36示出了在时间上处于特定标签中的信号(顶部)和后验概率(底部),在时间2410与2600之间的长均聚物区域的情况下,即大约模型停留在T翻转状态(红色虚线)而不是与T正位状态(红色实线)交替的情况。输入了长均聚物后,围绕所述区域的开始和结束进行了估计,但正位状态和翻转状态很快变得不那么明显,并且后验概率均匀。存在穿过所述区域的多条路径,其中正位和翻转碱基的配准略有不同,并且后验概率反映了此集合的平均值。

[0374] 一种可能的替代方案是使用最佳路径的评分作为训练目标,而不是对所有路径求和,并且这仍然是免配准的方法,因为没有明确定义配准,并且与标记不同,最佳配准可能会随着模型而改变。虽然训练最佳路径似乎很直观,但是从头开始训练模型时,这种方法却大为失败,因为最初的不良模型具有不良的最佳路径,并且训练过程对此有所强化。

[0375] 锐化是将训练集中朝向单个路径的方法,而不必事先指定配准,同时仍要考虑所有其它可能性。首先,考虑用于计算所有路径(图33)和最佳路径(图37)的总和评分的算法。这些都应用了函子, $\log \sum_j \exp$ 和 $\max_j$ 分别将转换权重和先前的前向向量组合在一起。锐化的目标是用仍然对所有可能路径求和但对评分较高的那些向上加权的函子代替此函子。

[0376] 图38示出了一些函子,所述函子可以是用于将前向向量和转换或映射权重组合在一起的函数。优选在图38中被称为“锐化的所有路径”的函子,但是可以使用许多其它函子,并且实际上,可以将它们组合在一起以创建新的函子。

[0377] 与从一开始就启用锐化进行训练相反,已发现使用所有路径目标函数开始训练,然后一旦找到了良好的模型,将锐化因子(a)从1增加到更高的值是有利的,其中可能会以更高的锐化值重复。此多阶段过程还使模型可以使用最佳路径目标进行训练。首先对所有路径目标进行训练找到良好的模型,使得最佳路径是良好的,然后通过进一步的训练来强化此路径。

[0378] 图39示出了在如图36所示出的相同示例区域上锐化正位-翻转表示的作用。图39示出了对于此实例随时间处于特定状态但使用锐化训练的信号(顶部)和后验概率(底部)。在时间2400与2620之间大约有一个长的区域,在所述区域中发生均聚物,并且模型在T翻转状态与T正位状态之间交替,以调用碱基序列。各个调用更不同,并且可以看出在整个均聚物区域中,T正位和T翻转之间交替。

[0379] 对此模型进行解码产生比未锐化模型更好的聚合物单元的估计。这在图40所示出的实例中进行了说明,其中将来自未锐化和锐化模型的聚合物单元(碱基调用)的估计值与参考序列进行了比较。尽管未锐化的调用仅调用8个T碱基,锐化的调用同意在参考中找到的27个T碱基。

[0380] 尽管使用最佳路径或锐化代替了训练异议,但它们也可以用来增强它并且训练网络远离测试期间已发现的不期望行为。一种此类不期望行为可能是向下调用均聚物长度的趋势,这可能会在训练数据严重偏向短均聚物时发生,并且可以通过对训练异议添加惩罚来校正。可以通过使用最佳路径找到调用均聚物的位置并且将其真实长度与基于网络在所述位置分配的对数概率的估计值进行比较,来找到一种此类惩罚;可以使用绝对差之和进行比较;可以使用平方差之和进行比较;许多其它比较方法是本领域已知的。可以将惩罚添

加到训练目标中;惩罚可以另外通过预先确定的因子加权,以改变其相对于训练目标的重要性。

[0381] 代替预先确定的,对惩罚项进行加权的因子可以被视为拉格朗日乘数(Lagrange multiplier)。在寻找拉格朗日乘数的固定点的同时,通过优化训练目标进行训练。在这些点处或附近,惩罚近似为零,并且网络已经受惩罚条件保持下的训练;对于惩罚是真实长度与估计长度之间的绝对差之和的实例,平均而言,网络调用将是正确的长度。

[0382] 可以使用多个惩罚项来增强训练目标,例如,每个均聚物长度一个惩罚项;每个惩罚可以通过预先确定的因子加权或被视为拉格朗日乘数。

[0383] 上文的描述考虑了权重分布51表示一组标签之间的转换的情况。作为替代方案,本文所描述的方法可以适用于权重分布51表示一组标签内的标签的情况。

[0384] 在权重分布51表示一组标签内的标签的情况下,解码器80可以使用转换矩阵来表示是否允许或不允许标签之间的转换。转换矩阵可以具有与权重分布51中的权重矩阵类似的形式,但是二进制元素指示允许或不允许的转换。转换矩阵可以将至少一个转换表示为不允许,并且将其它转换表示为允许。考虑到根据转换矩阵通过允许的标签的不同路径的可能性,解码器80可以使用此转换矩阵从表示标签的权重分布51导出一系列聚合物单元的估计值。

[0385] 同样在这种权重分布51表示一组标签中标签的情况下,可以以如上文所描述的编码形式来表示一系列聚合物单元中的相同类型的聚合物单元的连续实例,例如,使用正位-翻转-表示或游程长度编码表示。

[0386] 尽管上文的描述涉及包含RNN 50的神经网络10,但是具有上文所描述的形式和解码的权重分布可以等同地应用于任何其它形式的机器学习技术,例如HMM。

[0387] 根据本发明的第二方面,提供了一种如以下条款中所定义的方法。

[0388] 条款1.一种分析在聚合物相对于纳米孔易位期间从所述聚合物导出的信号的方法,所述聚合物包括属于一组可能类型的聚合物单元的一系列聚合物单元,所述方法包括:使用输出一系列权重分布的机器学习技术来分析所述信号,每个权重分布包括关于一组标签上的标签的权重,所述一组标签包含表示所述可能类型的聚合物单元的标签;以及从所述权重分布导出所述一系列聚合物单元的估计值,其中导出所述一系列聚合物单元的估计值的步骤将表示是否允许或不允许标签之间的转换的转换矩阵考虑在内,标签之间的至少一个转换被表示为不允许并且其它转换被表示为允许。

[0389] 条款2.根据条款1所述的方法,其中不允许标签之间的至少一个转换并且允许其它转换,所述权重分布各自包括关于被允许的标签的权重。

[0390] 条款3.根据条款2所述的方法,其中所述权重分布各自包括关于不被允许的标签的零权重。

[0391] 条款4.根据条款2或3所述的方法,其中导出所述一系列聚合物单元的估计值的步骤将表示是否允许或不允许标签之间的转换的转换矩阵考虑在内。

[0392] 条款5.根据条款2到4中任一项所述的方法,其中所述一组标签包含关于每种类型的聚合物单元的第一标签和第二标签,所述第一标签表示所述类型的聚合物单元的实例的开始,并且所述第二标签表示所述类型的聚合物单元的所述实例中的停留,其中允许任何其它类型的聚合物单元从每个第一标签转换到所述第一标签,允许相同类型的聚合物单元

从每个第一标签转换到所述第一标签,允许相同类型的聚合物单元从每个第一标签转换到所述第二标签,不允许任何其它类型的聚合物单元从每个第一标签转换到所述第二标签,允许相同类型的聚合物单元从每个第二标签转换到所述第一标签或允许任何其它类型的聚合物单元从每个第二标签转换到所述第一标签,并且允许相同类型的聚合物单元从每个第二标签转换到所述第二标签,并且不允许任何其它类型的聚合物单元从每个第二标签转换到所述第二标签。

[0393] 条款6.根据条款2到5中任一项所述的方法,其中所述一组可能类型的聚合物单元包含总是以聚合物单元的已知序列出现的类型的聚合物单元,允许与所述已知序列一致的转换并且不允许与所述已知序列相反的转换。

[0394] 条款7.根据条款2到6中任一项所述的方法,其中所述一系列聚合物单元中的相同类型的聚合物单元的连续实例以编码形式表示。

[0395] 条款8.根据条款7所述的方法,其中所述标签包含关于每种类型的聚合物单元的多个标签,其中关于每种类型的聚合物单元的所述多个标签表示所述一系列聚合物单元中所述类型的聚合物单元的连续实例。

[0396] 条款9.根据条款8所述的方法,其中用于每种类型的聚合物单元的所述多个标签具有预先确定的循环顺序,由此通过所述预先确定的循环顺序允许标签之间的一些转换并且通过所述预先确定的循环顺序不允许之间的其它转换,所述权重分布各自包含关于通过所述预先确定的循环顺序允许的标签的权重。

[0397] 条款10.根据条款8或9所述的方法,其中用于每种类型的聚合物单元的所述多个标签是用于每种类型的聚合物单元的两个标签。

[0398] 条款11.根据条款7所述的方法,其中所述一系列聚合物单元中的相同类型的聚合物单元的连续实例以游程长度编码形式表示。

[0399] 条款12.根据条款11所述的方法,其中所述标签包含关于每种类型的聚合物单元的不同游程长度的多个标签。

[0400] 条款13.根据条款11所述的方法,其中所述标签包含关于每种类型的聚合物单元的标签,并且对于每种类型的聚合物单元,所述权重分布包括在相同类型的聚合物单元的连续实例的可能长度上的另外的权重。

[0401] 条款14.根据条款3所述的方法,其中对于每种类型的聚合物单元,所述另外的权重包括在相同类型的聚合物单元的连续实例的一组可能长度上的权重的分类分布。

[0402] 条款15.根据条款13所述的方法,其中对于每种类型的聚合物单元,所述另外的权重包括在相同类型的聚合物单元的连续实例的连续实例的可能长度上的参数化分布的参数。

[0403] 条款16.根据条款2到15中任一项所述的方法,其中所述可能类型的聚合物单元包含具有未修饰形式和经修饰形式的类型的聚合物单元。

[0404] 条款17.根据条款16所述的方法,其中所述一组标签包含关于具有未修饰形式和经修饰形式的所述类型的聚合物单元的标签。

[0405] 条款18.根据条款17所述的方法,其中对于具有所述未修饰形式和所述经修饰形式的所述类型的聚合物单元中的每个聚合物单元的所述未修饰形式和所述经修饰形式,每个权重分布包括另外的权重。

[0406] 条款19.根据条款2到18中任一项所述的方法,其中所述一组标签包含至少一个表示每种类型的聚合物单元的标签。

[0407] 条款20.根据前述条款中任一项所述的方法,其中所述一组标签进一步包含至少一个表示所述一系列聚合物单元中的空白和/或停留的标签。

[0408] 条款21.根据前述条款中任一项所述的方法,其中所述机器学习技术是包括至少一个递归层的神经网络。

[0409] 条款22.根据条款21所述的方法,其中所述至少一个递归层是双向递归层。

[0410] 条款23.根据条款21或22所述的方法,其中所述神经网络相对于通过所述一系列权重分布的所有标签路径应用所述权重分布的全局归一化。

[0411] 条款24.根据条款21到23中任一项所述的方法,其中所述神经网络包含至少一个卷积层,所述至少一个卷积层布置在所述至少一个递归层之前并且执行所述信号的加窗部分的卷积。

[0412] 条款25.根据前述条款中任一项所述的方法,其中所述权重表示后验概率。

[0413] 条款26.根据前述条款中任一项所述的方法,其中从所述权重分布导出所述一系列聚合物单元的估计值的步骤使用链结式时间分类执行。

[0414] 条款27.根据前述条款中任一项所述的方法,其中从所述权重分布导出聚合物单元的估计值的步骤包括导出关于相应的权重分布的标签和对导出的标签进行游程长度压缩。

[0415] 条款28.根据前述条款中任一项所述的方法,其中从所述权重分布导出所述一系列聚合物单元的估计值的步骤包括估计按所述权重分布计通过所述一系列权重分布的最可能标签路径并且从被估计为最可能的所述标签路径导出所述一系列聚合物单元的所述估计值。

[0416] 条款29.根据前述条款中任一项所述的方法,其中从所述权重分布导出所述一系列聚合物单元的估计值的步骤包括估计关于每个权重分布最可能的所述标签、将通过所述一系列权重分布的前向和后向标签路径考虑在内并且从被估计为最可能的所述标签导出所述一系列聚合物单元的所述估计值。

[0417] 条款30.根据前述条款中任一项所述的方法,其中所述纳米孔是蛋白孔。

[0418] 条款31.根据前述条款中任一项所述的方法,其中所述聚合物是多核苷酸,并且所述聚合物单元是核苷酸。

[0419] 条款32.根据前述条款中任一项所述的方法,其中所述信号从以下性质中的一个或多个的测量结果导出:离子电流、阻抗、隧穿性质、场效应晶体管电压和光学性质。

[0420] 条款33.根据前述条款中任一项所述的方法,所述方法在计算机设备中执行。

[0421] 条款34.根据前述条款中任一项所述的方法,其进一步包括在所述聚合物相对于纳米孔易位期间从所述聚合物导出所述信号。

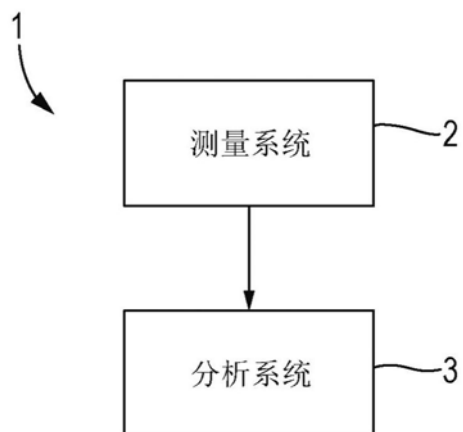


图1

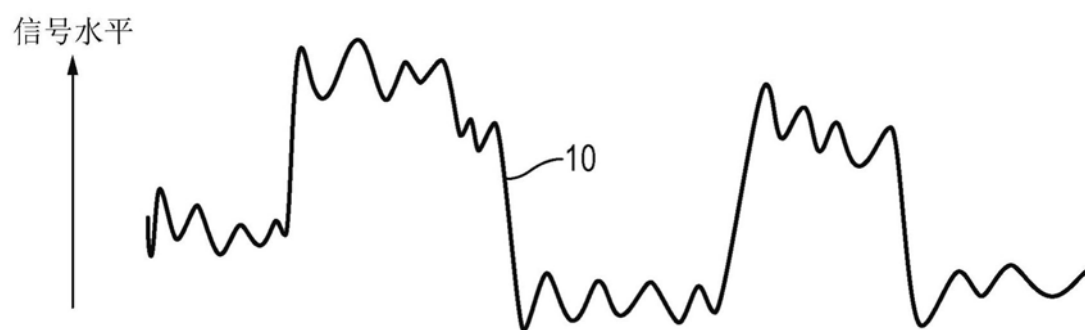


图2

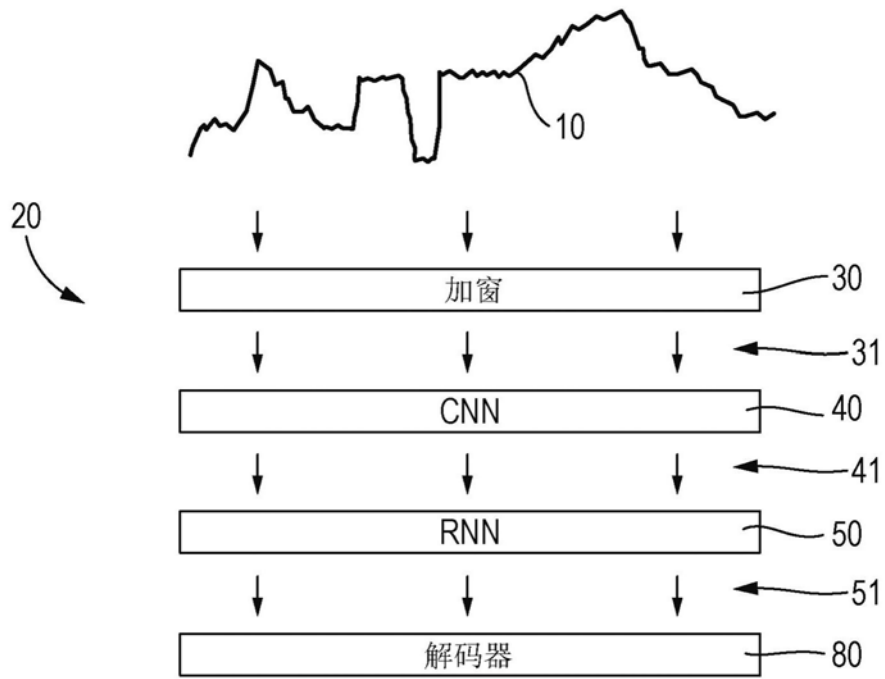


图3

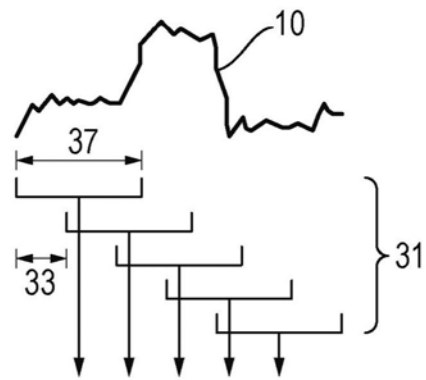


图4

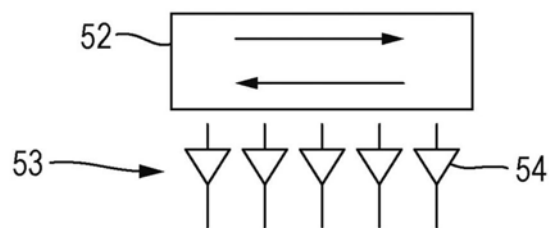


图5

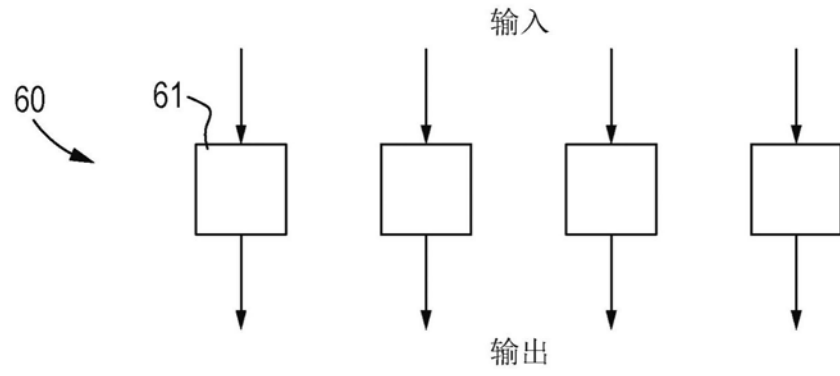


图6

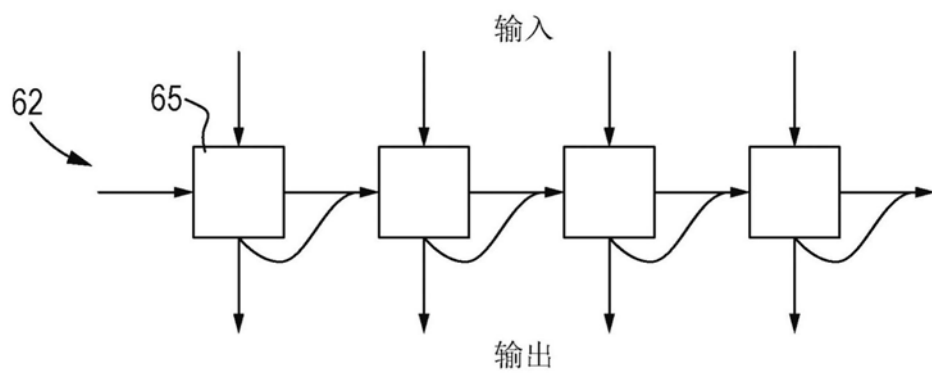


图7

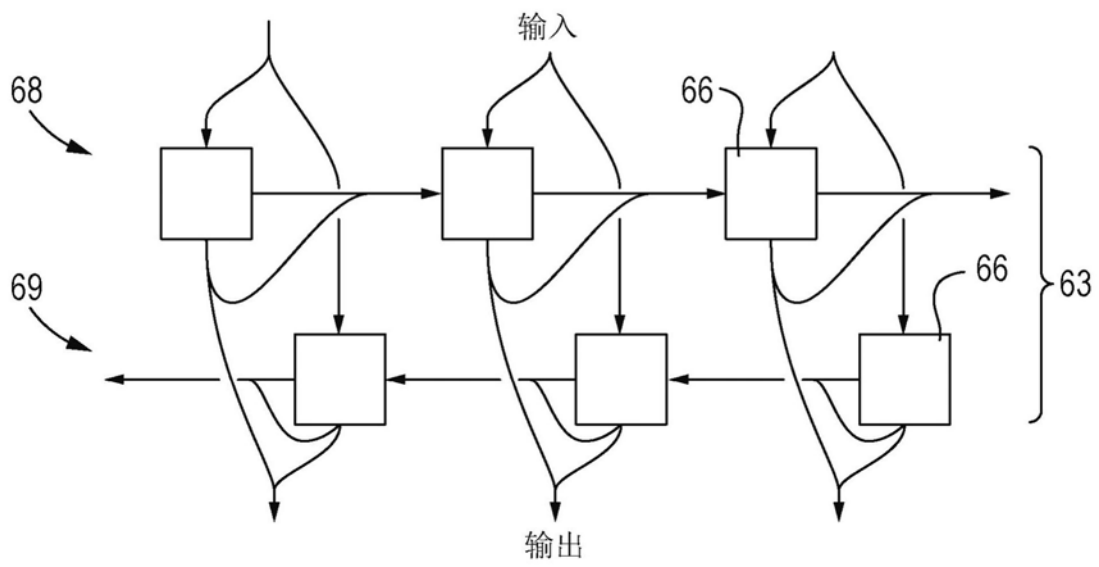


图8

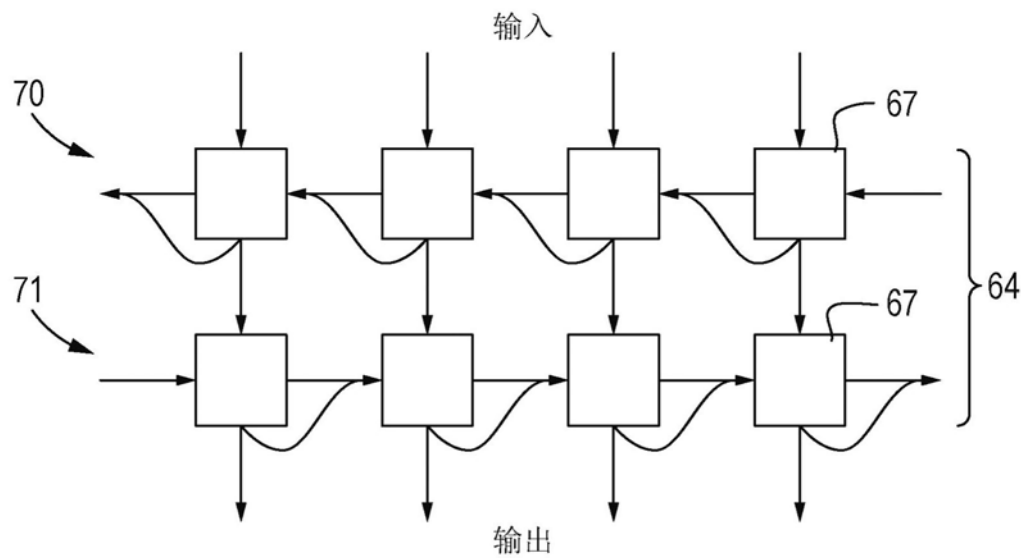


图9

基础

	A	C	G	T
A	W_{11}	W_{12}	W_{13}	W_{14}
C	W_{21}	W_{22}	W_{23}	W_{24}
G	W_{31}	W_{32}	W_{33}	W_{34}
T	W_{41}	W_{42}	W_{43}	W_{44}

图10

空白

	A	C	G	T	B
A	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}
C	W_{21}	W_{22}	W_{23}	W_{24}	W_{25}
G	W_{31}	W_{32}	W_{33}	W_{34}	W_{35}
T	W_{41}	W_{42}	W_{43}	W_{44}	W_{45}
B	W_{51}	W_{52}	W_{53}	W_{54}	W_{55}

图11

CpG甲基化

	A	C	G	T	B	C^m
A	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}	W_{16}
C	W_{21}	W_{22}	W_{23}	W_{24}	W_{25}	W_{26}
G	W_{31}	W_{32}	W_{33}	W_{34}	W_{35}	W_{36}
T	W_{41}	W_{42}	W_{43}	W_{44}	W_{45}	W_{46}
B	W_{51}	W_{52}	W_{53}	W_{54}	W_{55}	W_{56}
C^m	—	—	W_{63}	—	W_{65}	W_{66}

图12

多停留

	A	C	G	T	A ^s	C ^s	G ^s	T ^s
A	W ₁₁	W ₁₂	W ₁₃	W ₁₄	W ₁₅	—	—	—
C	W ₂₁	W ₂₂	W ₂₃	W ₂₄	—	W ₂₆	—	—
G	W ₃₁	W ₃₂	W ₃₃	W ₃₄	—	—	W ₃₇	—
T	W ₄₁	W ₄₂	W ₄₃	W ₄₄	—	—	—	W ₄₈
A ^s	W ₅₁	W ₅₂	W ₅₃	W ₅₄	W ₅₅	—	—	—
C ^s	W ₆₁	W ₆₂	W ₆₃	W ₆₄	—	W ₆₆	—	—
G ^s	W ₇₁	W ₇₂	W ₇₃	W ₇₄	—	—	W ₇₇	—
T ^s	W ₈₁	W ₈₂	W ₈₃	W ₈₄	—	—	—	W ₈₈

图13

正位-翻转

	A	C	G	T	A ^f	C ^f	G ^f	T ^f
A	W ₁₁	W ₁₂	W ₁₃	W ₁₄	W ₁₅	—	—	—
C	W ₂₁	W ₂₂	W ₂₃	W ₂₄	—	W ₂₆	—	—
G	W ₃₁	W ₃₂	W ₃₃	W ₃₄	—	—	W ₃₇	—
T	W ₄₁	W ₄₂	W ₄₃	W ₄₄	—	—	—	W ₄₈
A ^f	W ₅₁	W ₅₂	W ₅₃	W ₅₄	W ₅₅	—	—	—
C ^f	W ₆₁	W ₆₂	W ₆₃	W ₆₄	—	W ₆₆	—	—
G ^f	W ₇₁	W ₇₂	W ₇₃	W ₇₄	—	—	W ₇₇	—
T ^f	W ₈₁	W ₈₂	W ₈₃	W ₈₄	—	—	—	W ₈₈

图14

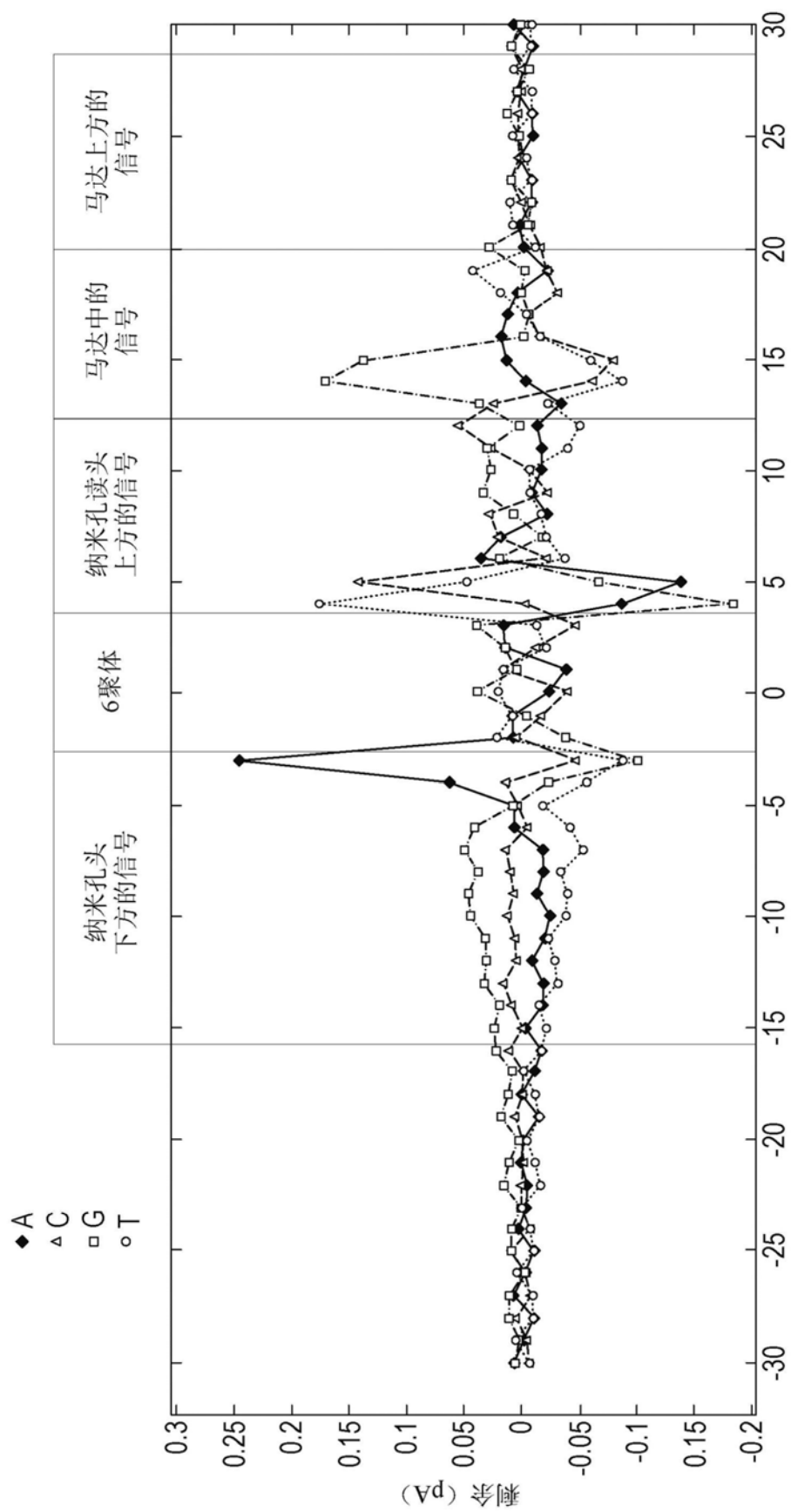


图15

RLE-离散										
	A ¹	C ¹	G ¹	T ¹	A ²	C ²	G ²	T ²	A ³	
A ¹	W ₁₁	W ₁₂	W ₁₃	W ₁₄	—	W ₁₆	W ₁₇	W ₁₈	—	
C ¹	W ₂₁	W ₂₂	W ₂₃	W ₂₄	W ₂₅	—	W ₂₇	W ₂₈	W ₂₉	
G ¹	W ₃₁	W ₃₂	W ₃₃	W ₃₄	W ₃₅	W ₃₆	—	W ₃₈	W ₃₉	
T ¹	W ₄₁	W ₄₂	W ₄₃	W ₄₄	W ₄₅	W ₄₆	W ₄₇	—	W ₄₉	
A ²	—	W ₅₂	W ₅₃	W ₅₄	W ₅₅	W ₅₆	W ₅₇	W ₅₈	W ₅₉	
C ²	W ₆₁	—	W ₆₃	W ₆₄	W ₆₅	W ₆₆	W ₆₇	W ₆₈	W ₆₉	
G ²	W ₇₁	W ₇₂	—	W ₇₄	W ₇₅	W ₇₆	W ₇₇	W ₇₈	W ₇₉	
T ²	W ₈₁	W ₈₂	W ₈₃	—	W ₈₅	W ₈₆	W ₈₇	W ₈₈	W ₈₉	
A ³	—	W ₉₂	W ₉₃	W ₉₄	—	W ₉₆	W ₉₇	W ₉₈	W ₉₉	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

图16

RLE-保持									
	A	C	G	T	A ^H	C ^H	G ^H	T ^H	
A	—	W ₁₂	W ₁₃	W ₁₄	W ₁₅	—	—	—	
C	W ₂₁	—	W ₂₃	W ₂₄	—	W ₂₆	—	—	
G	W ₃₁	W ₃₂	—	W ₃₄	—	—	W ₃₇	—	
T	W ₄₁	W ₄₂	W ₄₃	—	—	—	—	W ₄₈	
A ^H	—	W ₅₂	W ₅₃	W ₅₄	W ₅₅	—	—	—	
C ^H	W ₆₁	—	W ₆₃	W ₆₄	—	W ₆₆	—	—	
G ^H	W ₇₁	W ₇₂	—	W ₇₄	—	—	W ₇₇	—	
T ^H	W ₈₁	W ₈₂	W ₈₃	—	—	—	—	W ₈₈	

图17

RLE-分类

	1	2	3	4	
A	l_{11}	l_{12}	l_{13}	l_{14}	
C	l_{21}	l_{22}	l_{23}	l_{24}	
G	l_{31}	l_{32}	l_{33}	l_{34}	
T	l_{41}	l_{42}	l_{43}	l_{44}	

图18

RLE-参数化

	P_1	P_2	
A	P_{11}	P_{12}	
C	P_{21}	P_{22}	
G	P_{31}	P_{32}	
T	P_{41}	P_{42}	

图19

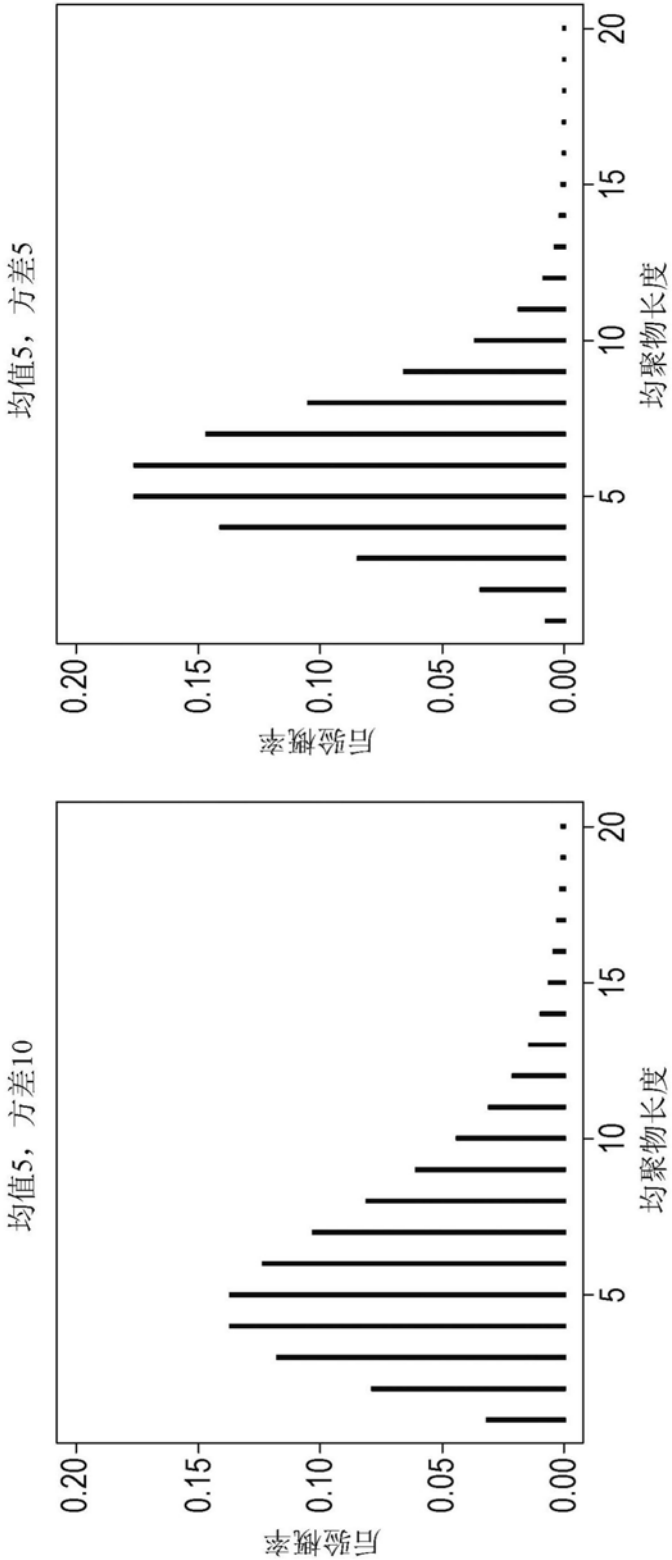


图20

分布 γ	参数 形状 $\alpha > 0$; 标度 $\beta > 0$	$\mathbb{P}(X = x), \text{ 方案1}$ $\frac{1}{\Gamma(\alpha)} [\gamma(\alpha, \frac{x+1}{\beta}) - \gamma(\alpha, \frac{x}{\beta})]$	$\mathbb{P}(X = x), \text{ 方案2}$ $\frac{1}{\Gamma(\alpha)} [\gamma(\alpha, \frac{x+1/2}{\beta}) - \gamma(\alpha, \frac{x-1/2}{\beta})]$
对数逻辑	形状 $\alpha > 0$; 标度 $\beta > 0$	$\frac{1}{1 + (\frac{x+1}{\alpha})^{-\beta}} - \frac{1}{1 + (x/\alpha)^{-\beta}}$	$\frac{1}{1 + (\frac{x+1/2}{\alpha})^{-\beta}} - \frac{1}{1 + (\frac{x-1/2}{\alpha})^{-\beta}}$
对数正态	对数-均值 $\mu \in \mathbb{R}$; log-SD $\sigma > 0$	$\frac{1}{2} \Phi(\frac{\log(x+1) - \mu}{\sigma\sqrt{2}}) - \frac{1}{2} \Phi(\frac{\log x - \mu}{\sigma\sqrt{2}})$ $e^{-(x/\beta)^\alpha} - e^{-(\frac{x+1}{\beta})^\alpha}$	$\frac{1}{2} \Phi(\frac{\log(x+1/2) - \mu}{\sigma\sqrt{2}}) - \frac{1}{2} \Phi(\frac{\log(x-1/2) - \mu}{\sigma\sqrt{2}})$ $e^{-(\frac{x-1/2}{\beta})^\alpha} - e^{-(\frac{x+1/2}{\beta})^\alpha}$
威布尔	形状 $\alpha > 0$; 标度 $\beta > 0$		

图21

	P ₁	P ₂	
A→C	P ₁₁	P ₁₂	
A→G	P ₂₁	P ₂₂	
A→T	P ₃₁	P ₃₂	
C→A	P ₄₁	P ₄₂	
C→G	P ₅₁	P ₅₂	
C→T	P ₆₁	P ₆₂	
G→A	P ₇₁	P ₇₂	
G→C	P ₈₁	P ₈₂	
G→T	P ₉₁	P ₉₂	
T→A	P ₁₀₋₁	P ₁₀₋₂	
T→C	P ₁₁₋₁	P ₁₁₋₂	
T→G	P ₁₂₋₁	P ₁₂₋₂	

图22

	P ₁	P ₂	
C→A→C	P ₁₁	P ₁₂	
G→A→G	P ₂₁	P ₂₂	
T→A→T	P ₃₁	P ₃₂	
G→A→C	P ₄₁	P ₄₂	
T→A→G	P ₅₁	P ₅₂	
T→A→T	P ₆₁	P ₆₂	
T→A→C	P ₇₁	P ₇₂	
C→A→G	P ₈₁	P ₈₂	
G→A→T	P ₉₁	P ₉₂	
A→C→A	P ₁₀₋₁	P ₁₀₋₂	
G→C→G	P ₁₁₋₁	P ₁₁₋₂	
T→C→T	P ₁₂₋₁	P ₁₂₋₂	
G→C→A	P ₁₃₋₁	P ₁₃₋₂	
T→C→G	P ₁₄₋₁	P ₁₄₋₂	
A→C→T	P ₁₅₋₁	P ₁₅₋₂	
T→C→A	P ₁₆₋₁	P ₁₆₋₂	
A→C→G	P ₁₇₋₁	P ₁₇₋₂	
G→C→T	P ₁₈₋₁	P ₁₈₋₂	
T→G→T	P ₁₉₋₁	P ₁₉₋₂	
A→G→A	P ₂₀₋₁	P ₂₀₋₂	
C→G→C	P ₂₁₋₁	P ₂₁₋₂	
A→G→T	P ₂₂₋₁	P ₂₂₋₂	
C→G→A	P ₂₃₋₁	P ₂₃₋₂	
T→G→C	P ₂₄₋₁	P ₂₄₋₂	
C→G→T	P ₂₅₋₁	P ₂₅₋₂	
T→G→A	P ₂₆₋₁	P ₂₆₋₂	
A→G→C	P ₂₇₋₁	P ₂₇₋₂	
A→T→A	P ₂₈₋₁	P ₂₈₋₂	
C→T→C	P ₂₉₋₁	P ₂₉₋₂	
G→T→G	P ₃₀₋₁	P ₃₀₋₂	
C→T→A	P ₃₁₋₁	P ₃₁₋₂	
G→T→C	P ₃₂₋₁	P ₃₂₋₂	
A→T→G	P ₃₃₋₁	P ₃₃₋₂	
C→T→A	P ₃₄₋₁	P ₃₄₋₂	
G→T→C	P ₃₅₋₁	P ₃₅₋₂	
A→T→G	P ₃₆₋₁	P ₃₆₋₂	

图23

	A	C	G	T	B	C ^M
A	W ₁₁	W ₁₂	W ₁₃	W ₁₄	W ₁₅	W ₁₆
C	W ₂₁	W ₂₂	W ₂₃	W ₂₄	W ₂₅	W ₂₆
G	W ₃₁	W ₃₂	W ₃₃	W ₃₄	W ₃₅	W ₃₆
T	W ₄₁	W ₄₂	W ₄₃	W ₄₄	W ₄₅	W ₄₆
B	W ₅₁	W ₅₂	W ₅₃	W ₅₄	W ₅₅	W ₅₆
C ^M	W ₆₁	W ₆₂	W ₆₃	W ₆₄	W ₆₅	W ₆₆

图24

	C	C ^M
A	M ₁	M ₂

图25

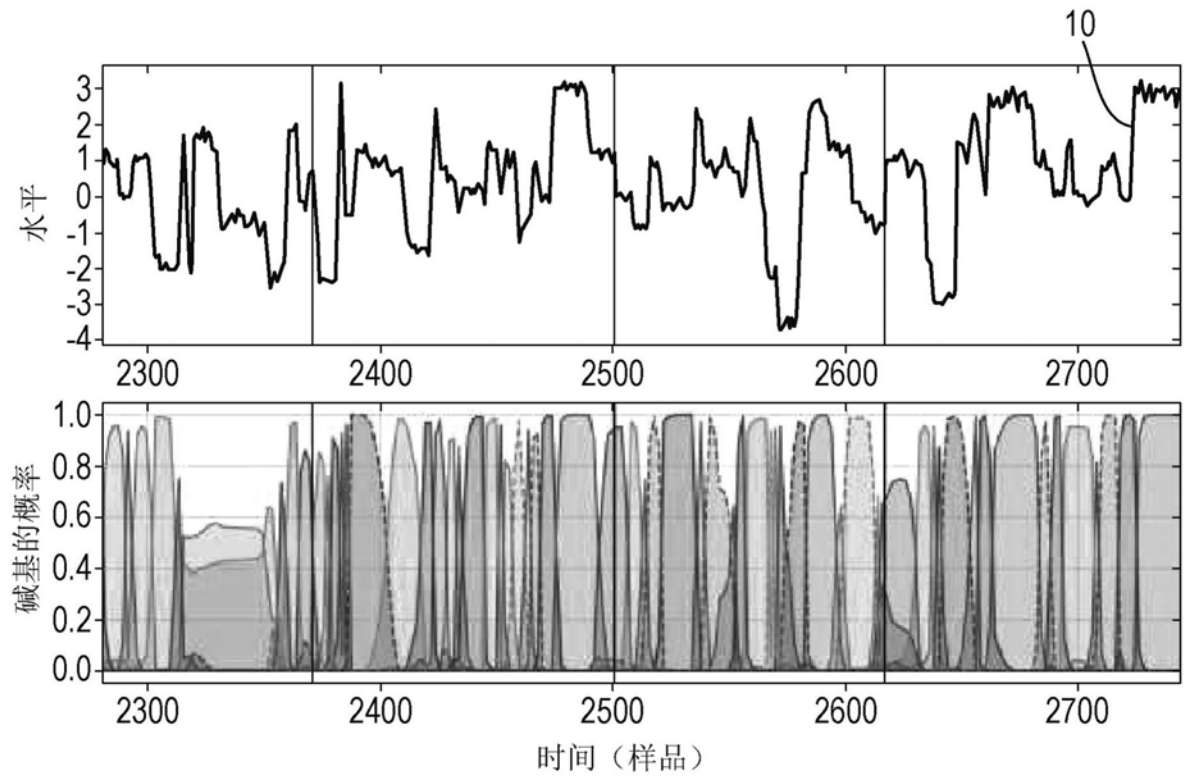


图26

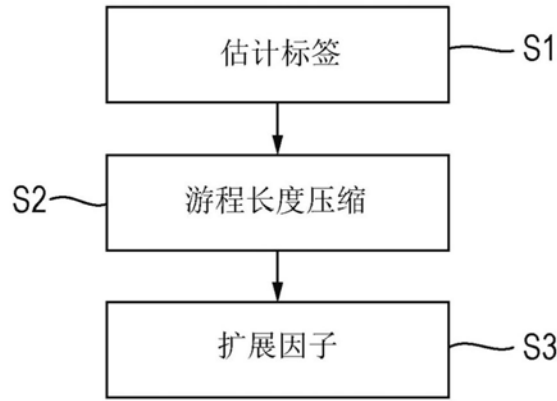


图27

f_k^0	$= 0 \quad \forall k$	初始化前向向量
f_k^{i+1}	$= \max_j f_j^i + w_{jk}^i$	最佳路径更新
t_k^{i+1}	$= \operatorname{argmax}_j f_j^i + w_{jk}^i$	回溯更新
S	$= \max_j f_j^n$	计算评分
P_n	$= \operatorname{argmax}_j f_j^n$	最终位置处的最高评分状态
P_{i-1}	$= t_{P_i}^i$	回溯以找到最佳路径

图28

f_k^0	$= p_k^0$	初始化前向向量
f_k^{i+1}	$= \max_j f_j^i + T_{jk} p_k^{i+1}$	最佳路径更新
t_k^{i+1}	$= \operatorname{argmax}_j f_j^i + T_{jk} p_k^{i+1}$	回溯更新
S	$= \max_j f_j^n$	计算评分
P_n	$= \operatorname{argmax}_j f_j^n$	最终位置处的最高评分状态
P_{i-1}	$= t_{P_i}^i$	回溯以找到最佳路径

图29

f_k^0	$= \log p_k^0$	初始化前向向量
f_k^{i+1}	$= \max_j f_j^i + T_{jk} \log p_k^{i+1}$	最佳路径更新
t_k^{i+1}	$= \operatorname{argmax}_j f_j^i + T_{jk} \log p_k^{i+1}$	回溯更新
S	$= \max_j f_j^n$	计算评分
P_n	$= \operatorname{argmax}_j f_j^n$	最终位置处的最高评分状态
P_{i-1}	$= t_{P_i}^i$	回溯以找到最佳路径

图30

f_k^0	$= 0 \quad \forall k$	初始化前向向量
f_k^{i+1}	$= \log \sum_j e^{f_j^i + w_{jk}^i}$	前向更新
b_k^N	$= 0 \quad \forall k$	初始化后向向量
b_k^{i-1}	$= \log \sum_j e^{b_j^i + w_{kj}^i}$	后向更新
S	$=$	评分（归一化因子）
p_j^i	$= \frac{1}{S} e^{f_j^i + b_j^i}$	后验概率（标签j）
P_{jk}^i	$= \frac{1}{S} e^{f_j^i + w_{jk}^i + b_k^{i+1}}$	后验概率（转换j → k）

图31

$m_{j \ j+1}^i$	$= w_{S_j}^i S_{j+1}$	从状态 S_j 移动到 S_{j+1}
$m_{j \ j}^i$	$= w_{S_j}^i S_j$	停留在状态 S_j
$m_{j \ k}^i$	$= 0$	其它方式

图32

$$\begin{aligned}
 f_k^0 &= 0 \quad k = 1; -\infty \quad \text{否则} && \text{初始化前向向量} \\
 f_k^{i+1} &= \log \sum_j e^{f_j^i + m_{jk}^i} && \text{所有路径更新} \\
 S &= f_N^n && \text{计算评分}
 \end{aligned}$$

图33

$$\begin{aligned}
 &\text{从位置 } j \text{ 移动到 } j+1 \\
 &\text{从已保持的位置 } j \text{ 移动到 } j+1 \\
 &\text{保持在位置 } j \text{ (新到达)} \\
 &\text{保持在位置 } j \text{ (先前到达)} \\
 &\text{其它方式} \\
 \\
 &m_{jj+1}^i = w_{S_j S_{j+1}}^i \\
 &m_{j+N+1, j+1}^i = w_{R_j S_{j+1}}^i \\
 &m_{j+N+1}^i = w_{S_j R_j}^i \\
 &m_{j+N+1, j+N+1}^i = w_{R_j R_j}^i \\
 &m_{jk}^i = 0
 \end{aligned}$$

图34

$$\begin{array}{lcl}
m_{j+1}^i & = & w_{S_{j+1}}^i + r_{S_{j+1}:L_{j+1}}^i \\
m_{j+N+1}^i & = & w_{R_{j+1}}^i + r_{S_{j+1}:L_{j+1}}^i \\
m_{j+N+1}^i & = & w_{S_j}^i \\
m_{j+N+1}^i & = & w_{R_j}^i \\
m_{j+k}^i & = & 0
\end{array}$$

从位置 j 移动到 $j+1$
 从已保持的位置 j 移动到 $j+1$
 保持在位置 j （新到达）
 保持在位置 j （先前到达）
 其它方式

图35

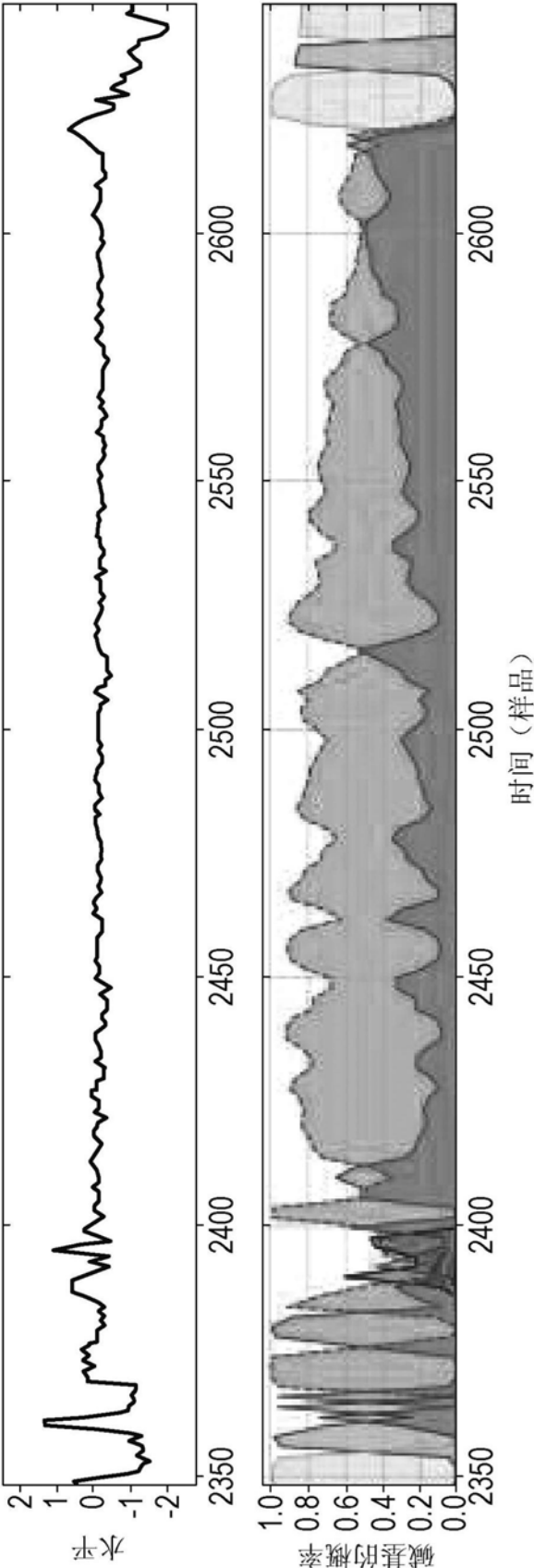
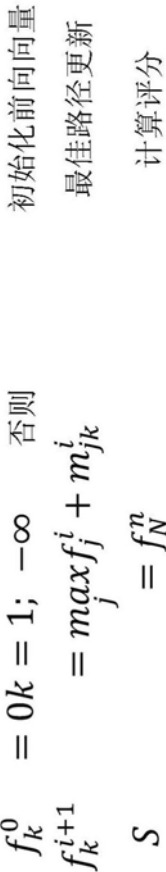


图36



否则

$$= 0k = 1; -\infty$$
$$= \max_j f_j^i + m_{jk}^i$$
$$= f_N^n$$

f_k^0

f_k^{i+1}

S

图37

函子		
$\log \sum_j \exp$	所有路径	L_1 归一化
$\max_j$	最佳路径	L_∞ 归一化
$\frac{1}{a} \log \sum_j \exp a$	锐化的所有路径	L_a 归一化
$\log \sum_j \exp - \max_j$	除了最佳路径之外的所有路径	
$p \max_j + (1 - p) \log \sum_j \exp$	向上加权的最佳路径	

图38

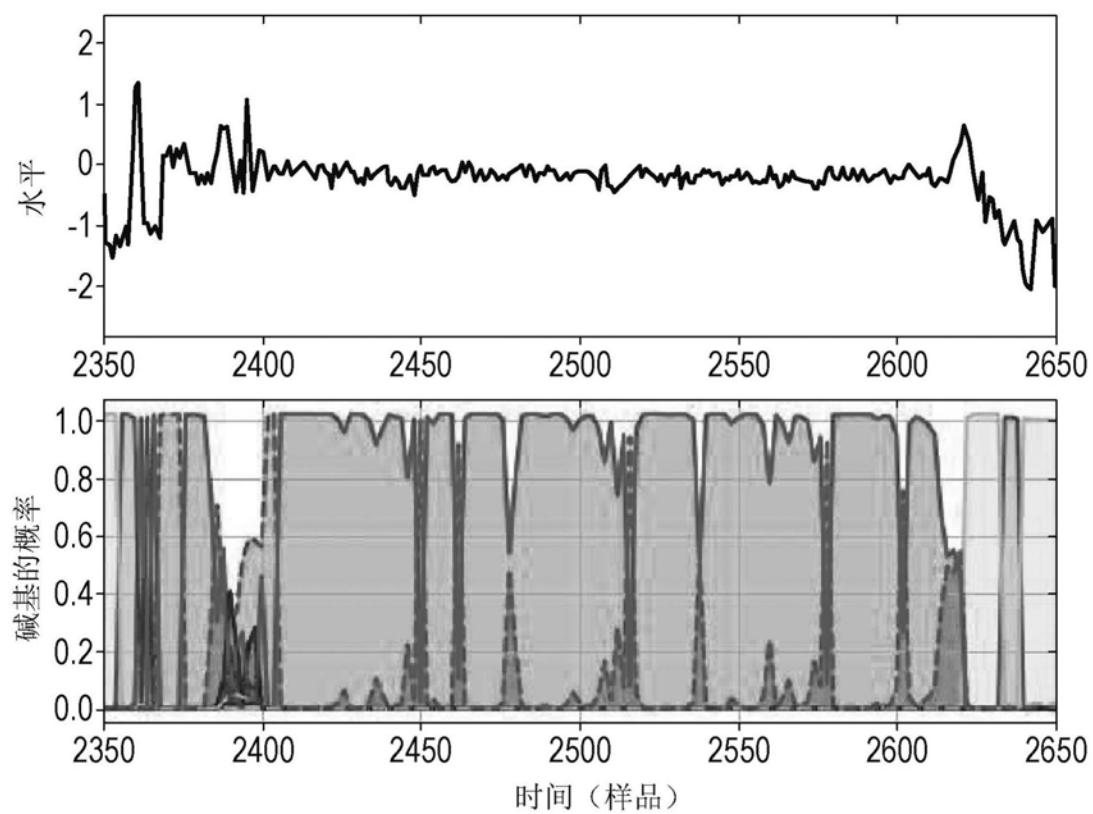


图39

模型	调用与参考的比对	
未锐化	<pre> GATACTAT--CACGATATTTTCTCTTTT-----GAGACAG-GTC . GATACTATACGATATTTCTTTTTTTTTTTTTGTGACGGAGTC </pre>	调用 参考
锐化	<pre> GATACTAT--CACGATATTTTCTTTTTTT-----GAGACAG-GTC GATACTATACGATATTTCTTTTTTTTTTGTGACGGAGTC </pre>	调用 参考

图40