DOI:10.13523/j.cb.2104018

纳米孔测序信号处理及其在DNA数据存储的应用

葛 奇[1] 张 鹏[1] 韩明哲[2,3] 杨晋生[1] 张大璐[4*] 陈为刚[1,3]

(1 天津大学微电子学院 天津 300072 2 天津大学化工学院 天津 300072)

(3 教育部合成生物学前沿科学中心 天津大学 天津 300072 4 中国生物技术发展中心 北京 100039)

摘要 随着高通量测序技术的不断更新,可以在单个分子水平读取核苷酸序列的第三代测序技术迅速发展,纳米孔测序技术是其具有代表性的单分子测序技术,该技术通过检测DNA单链分子穿过纳米孔时引起的跨膜电流信号的变化,实现碱基识别。纳米孔测序仪在便携性、碱基读取速度、测序读段长度等方面较传统的第一代与第二代测序技术都有明显优势。随着纳米孔测序技术的不断发展成熟,与其配套的各种信号处理与生物信息处理工具也迅速涌现,碱基识别和模型仿真是其中两个较为关键的研究方向。首先介绍纳米孔测序基本原理与信号处理流程,探讨其目前面临的挑战,归纳近年来在碱基识别与纳米孔模型仿真两个方面的主要进展与发展趋势,并用实测数据比较了不同碱基识别方法的性能。继而搭建了纳米孔测序集成仿真平台,为信号处理算法的评估提供支撑。进一步,随着全球数据量的爆发式增长,DNA数据存储正成为未来非常有潜力的海量数据存储方式,采用纳米孔测序读出是一种非常有效的方法。总结了纳米孔测序技术在DNA数据存储中的应用进展,分析了其可行性。分析了基于纳米孔测序实现的人工染色体数据存储的快速读出方法,探讨了与实际测序数据结合的纳米孔测序读段仿真在DNA数据存储中的应用,为开发适合DNA数据存储的方案提供参考。

关键词 纳米孔测序 碱基识别 纳米孔信号处理 DNA数据存储

中图分类号 Q819

采用纳米孔传感器对脱氧核糖核酸(DNA)测序是新一代DNA测序技术[1]。纳米孔作为第三代测序的核心器件,其设计不仅涉及分子生物学、合成生物学等知识,还包括基于微纳加工的异构集成的传感器设计以及配套的信号处理方法等,是典型的需多学科交叉创新的领域,引起了各领域研究者的广泛关注[2-3]。用于DNA测序的纳米孔主要分为两大类:以蛋白质构建的生物纳米孔;以半导体、合成材料等构建的固态纳米孔[4-16]。纳米孔测序具备测序读段长[17]、测序速度快、测序数据可以实时分析[18]以及无需聚合酶链反应(PCR)扩增[19]等诸多优点,在不同领域得到了广泛应用[20-21]。例如,用于检测胞嘧啶的四种碱基修饰,准确率可以达到92%~98%[22];临床实践中实时获取和分析DNA/RNA序列[23],用于疾病组织中的DNA结构变异的检测[24]、疫情监测[25-27]、病毒病原体结构检测[28]等。在上述应用的推动下,与纳米孔测序技术配套的信号处理与生物信息处理工具也迅速发展[29],其中包括序列对比工具[30-32]、从头组装工具[33-35]以及测序信号可视化工具等[36],建立了基于三代纳米孔测序的技术体系。但是,该技术框架与体系中仍存在很多有待完善的问题。

一方面,碱基识别(base calling)是指将测序输出的原始电流信号转换为碱基序列的处理过程,是分析与处理测序数据的基础[37]。以牛津纳米孔(Oxford nanopore technologies,ONT)为例,在被电解液包围的薄膜上施加电压差,带有负电荷的单链核苷酸在马达(Motor)蛋白的协助下经过蛋白质纳米孔穿过薄膜,核苷酸链的不同碱基会对孔的导电性产生变化,记录随时间变化的电流信号即可检测通过孔的不同碱基[38]。

收稿日期:2021-04-14 修回日期:2021-05-25

* 通讯作者,电子信箱:zhangdl@cncbd.org.cn

由于原始电流信号非常微弱，存在较多的噪声且具有随机性，碱基识别与下游分析成为了一个极具挑战的问题[35,39-40]。近年来，研究人员开发了多种碱基识别软件，如 Nanocall[41]、DeepNano[42]、BasecRAWller[43] 和 Chiron[44] 等。目前的纳米孔测序碱基识别准确度已经从75%提高到约 90%，但仍低于第二代测序技术的99%以上的读段精度[45-46]。纳米孔测序过程较为复杂，会导致原始电流信号的信噪比很低，使得后续测序数据的分析变得非常困难[41-44]。

另一方面，目前纳米孔商用产品种类少，开展实际纳米孔测序的成本较高，亟须一种仿真方法代替实际实验，支持基于纳米孔测序信号分析工具的开发。研究纳米孔测序电信号的仿真模型，可为测序数据分析方法与工具的开发提供可行的测试与验证平台[29,47-48]。与利用实际测序数据的分析评估相比，纳米孔测序信号仿真模型的开发可有效节约成本、降低数据分析难度并提高研发效率[49]。目前，已开发的常用仿真软件包括测序读段仿真软件（ReadSim[50]、SiLiCO[51] 和 NanoSim[52]）和测序信号仿真软件（DeepSimulator[53-54] 和 NanosigSim[55]）两种类型。

纳米孔测序可以支撑新一代数据存储方式 DNA 数据存储的快速读出，该领域亟须与数据存储系统研发相适应的模型仿真工具[56-58]。随着合成生物学的发展以及纳米孔测序碱基识别准确率的不断提升，将数字信息存储在合成 DNA 中，利用测序技术进行数据的读出已具有较好可行性，成为数据存储领域发展的热点[59-65]。三代测序技术的便携、快速读出等特点在 DNA 数据存储的读出中具有很好的应用前景。然而，DNA 数据存储的研究需建立 DNA 合成、处理与测序的完整流程，涉及多学科，设计与分析过程复杂，成本较高，使得 DNA 数据存储的研究较为困难，研究者较少。本文结合在 DNA 数据存储与测序读出等方面开展的工作，介绍基于仿真的 DNA 数据存储流程研究，为 DNA 数据存储的研究提供可行的仿真平台。

本文首先对纳米孔测序的碱基识别与仿真模型进行综述和分析，分析了纳米孔测序的基本原理以及存在的技术缺陷。重点介绍从传感器得到的电信号转变为碱基的识别方法、纳米孔测序仿真的方法，并在此基础上建立仿真平台，仿真生成纳米孔测序电信号并进行碱基识别方法的实际分析与验证。进一步，面向前沿的 DNA 数据存储领域，介绍了纳米孔测序信号处理与仿真流程，重点介绍了针对人工染色体数据存储的纳米孔测序读出信号的处理以及 DNA 数据存储的全流程仿真案例。

1 纳米孔测序原理及其信号处理面临的挑战

纳米孔测序的概念提出于20 世纪 80 年代[1]，不同的样本（包括生物样本或者 DNA 数据存储的样本），测序的流程均包括文库构建、纳米孔检测、信号读出以及序列分析等步骤（图 1）。首先，测序前需要对生物测序样本进行核酸文库的构建。以 DNA 测序为例，建库时，DNA 序列与 DNA 解旋酶混合在稀释液中，DNA 解旋酶的本质是 Motor 蛋白，DNA 片段的接头一端连接 Motor 蛋白，另一端连接 Tether 蛋白[5]。在腺苷三磷酸（ATP）供能下，Motor 蛋白作为推进器，将双链结构的 DNA 解旋为单链，并拉动单链分子通过纳米孔。Tether 蛋白用于锚定 DNA 链，防止在溶液中飘动，并使单链 DNA 进入纳米孔。其次，纳米孔测序技术是一种基于电流信号的 DNA 不同碱基的检测方法[66]，核心是一种由蛋白质构成的纳米孔，纳米孔被嵌入在一层具有超高电阻率的合成薄膜上，薄膜的两侧都浸没在含有离子的生理溶液中，纳米孔成为连接两个电解液室的通道[20]。当在薄膜的两侧施加不同的电位时，离子就会通过纳米孔从薄膜的一侧移动到另一侧，形成穿过纳米孔的稳态离子电流。在测序过程中，当单链核苷酸序列通过纳米孔穿过薄膜时，碱基分子会对离子的流动造成阻碍，引起穿过纳米孔的电流信号特征发生变化。由于不同的碱基对离子流动造成的阻碍大小是不同的，分析电流波动信号可以识别出正在通过该纳米孔的碱基，完成碱基序列的读取。

最常用的纳米孔测序设备为 ONT 公司开发的测序仪[67]。ONT 于 2008 年获得了纳米孔测序技术核心专利的授权，并于 2010 年开始进行核苷酸链的测序工作。基于纳米孔测序的原理，利用高度集成电路技术，ONT 公司开发了第一款商用的高通量纳米孔测序平台 MinION。MinION 是一个重量仅为 90 g 的便携式测序设备，核心部分为一种称为“flow cell”的测序芯片，其上有多达 2 048 个可单独寻址的纳米孔。在“flow cell”上有 512 个传感器，其中每个传感器连接了 4 个纳米孔，并通过专用集成电路进行控制。MinION 中每个纳米孔的碱基通过速度大约为 450 bp/s，每个“flow cell”芯片可以生成 10 ~ 30 Gb 的测序数据[67]。

MinION 测序平台具有体积小、重量低、便携性强、

连接电脑即可使用、不需要额外的设备且不局限于实验室环境等优点。但是也存在尚未解决的缺点,测序时,感测电极无法识别长均聚物通过纳米孔时的电流变化。目前,纳米孔测序的准确率低于二代测序技术,达到 90% ~95%,主要原因是纳米孔测序过程中的两个步骤会出现误差,导致测序后数据中包括碱基的插入(insertion)、删节(deletion)与替代(substitution)错误。一方面,通过纳米孔的 DNA 或 RNA 的四种碱基结构较为相似,导致不同碱基产生的原始电流信号差异较小从而产生低信噪比的原始电流信号。另一方面,在将原始电流信号转换为碱基序列的过程中也会出现错误,目前采样率参数以及碱基序列的信息实际上已存在于原始电流信号中,但是由于该信号模型与传统信号处理的模型存在较大差别,需要进一步研究[68]。

DNA 由四种碱基构成,四种碱基 A、T、G、C 的排列顺序包含了存储信息。k 个碱基的任意一个排列被视为一个 k-mer,其中 k 指的是这个字符串的长度。在纳米孔测序中,Motor 蛋白引导 DNA 片段进入孔道,并伴随着电流通过。当 DNA 序列穿过纳米孔时,其电压特性的变化可以被观察到,位移的幅度和持续时间等参数被记录下来。由于纳米孔的厚度大于 DNA 碱基的尺度,存在于孔中的碱基个数一般为 5 ~6 个,被认为是一个特殊的 k-mer 序列。当下一个碱基进入纳米孔时,一个新的 k-mer 产生并引起电流的变化[46]。当前版本的 MinION 测序仪,单个核苷酸链以平均 450 bp/s 的速度穿过蛋白质纳米孔,而原始电流信号的采样频率为 4 kHz,这种速率上的差距意味着平均每个 k-mer 有 9 个离散的测量信号值,有利于碱基的正确识别[46]。但是,测序过程中有以下几个因素会导致低信噪比的原始电流信号:(1) DNA/RNA 的四种碱基结构差异性较小,导致不同碱基通过纳米孔时引起的原始电流信号差异也较小;(2) 原始电流信号主要受到同时占据纳米孔的 5 或 6 个碱基的影响,电流测量值会对应 4^5 或 4^6 个可能的 k-mer 值,给识别造成较大困难;(3) 牵引 DNA 或 RNA 序列移动的 Motor 蛋白的非理想特性会导致碱基序列的移速不均匀;(4) 当一个均聚物超过 k-mer 长度时,一个 k-mer 穿过纳米孔,另一个 k-mer 进入纳米孔时的变化很难确定,通过纳米孔时的原始电流信号没有发生变化,从而导致具有多个相同碱基的长链被错误解释[46]。上述几个因素使得纳米孔测序技术产生的测序读段精度较低,导致测序后的数据发生错误,且总错误率达到 10% 以上。例如,在文献[58]我们开展的三代纳米孔测序实验中,实际测序读段的插入、删节错误的概率高达 7.2%,替代错误的概率为 3.59%[58]。

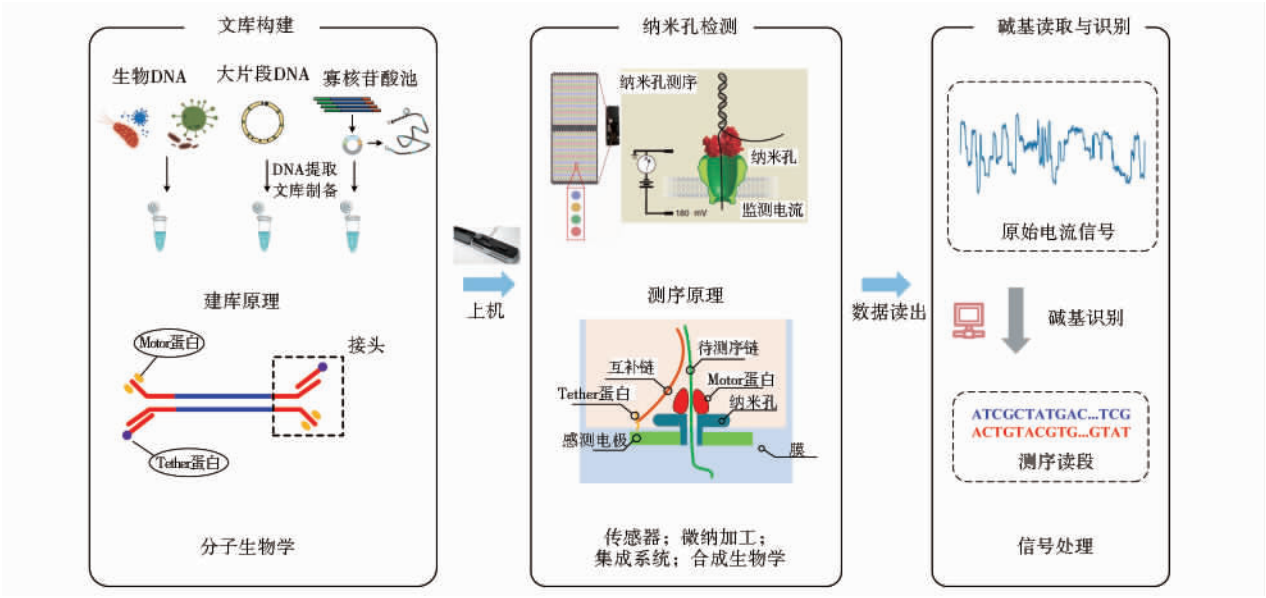

图 1　纳米孔测序的基本流程与基本原理

Fig. 1　The workflow and principles of the nanopore sequencing paradigm

2 碱基识别: 从纳米孔传感电流信号到碱基序列

三代测序仪 MinION 的发布标志着纳米孔测序技术进入新发展阶段,同时也带来一个新挑战,即如何将原始电流信号转换为碱基序列。基于前文提及的纳米孔传感器的特性,碱基识别成为了一个极具挑战性的信号处理与识别问题。在最早发布 MinION 时,用户只能通过 ONT 的云计算平台 Metrichor 来完成碱基识别工作,Metrichor 平台的两个主要缺点限制了 MinION 的使用。首先,Metrichor 必须依赖于互联网的接入,互联网受限情况下的测序任务(如在偏远地区进行的现场测序) 无法正常进行。其次,Metrichor 平台碱基识别精度较低,对下游的读段组装与分析造成较大困难[69]。因此,提高碱基识别准确度的同时扩展纳米孔测序仪的应用领域成为非常重要的研究问题,ONT 和许多研究团队开发了多种碱基识别方法(表 1) 。根据输入数据类型,开发的碱基识别方法可分为两类: 基于分段事件的识别方法、基于原始电流信号的识别方法。

表 1 纳米孔测序的碱基识别软件比较

Table 1 The comparison of base calling software for nanopore sequencing

碱基识别软件	是否开源	程序语言	输入数据类型	核心计算模型	开发团队
Nanocall	是	C + +	分段事件	隐马尔科夫模型	David 等[41]
Nanonet	是	C + +	分段事件	循环神经网络	ONT 公司
DeepNano	是	Python	分段事件	循环神经网络	Boza 等[42]
BasecRAWller	否	–	原始电流信号	循环神经网络	Stoiber 等[43]
Chiron	是	Python	原始电流信号	循环神经网络	Teng 等[44]
Albacore	否	Python	原始电流信号	循环神经网络	ONT 公司
Guppy	否	–	原始电流信号	循环神经网络	ONT 公司
Scrappie	是	C	分段事件或原始电流信号	循环神经网络	ONT 公司
Flappie	是	C	原始电流信号	循环神经网络	ONT 公司

2.1 基于分段事件的碱基识别方法

在碱基识别之前可以将原始电流信号分段成离散的事件,该方法可以减小输入数据集的大小,并将冗余测量结果合并为一个更可靠的、基于事件的信号[70]。目前常用的分段算法,首先在原始电流信号上计算两对相邻滑动窗口的统计量,然后根据统计信息确定事件的边界,从而实现信号分段。经过分段得到的每个事件包含了该事件所对应信号的平均值、标准差和信号长度。为将事件序列正确地解释为碱基序列,ONT 提供了孔模型和缩放参数,其中孔模型包含了每个 *k*-mer 的预期信号值,而缩放参数则用于校正不同测序运行过程中可能出现的差异。

早期碱基识别方法采用隐马尔科夫模型(hidden Markov model,HMM) ,根据事件序列、孔模型和缩放参数来预测碱基序列,如 Nanocall[41]。在碱基识别的隐马尔科夫模型中,以输入的分段事件序列作为观测状态,隐藏状态表示所有可能的 *k*-mer,初始概率基于孔模型计算得到,而状态转移概率则是根据训练数据确定的。在碱基识别过程中,维特比译码算法用于计算最大可能生成观测事件序列的隐藏状态序列,并根据连续两个隐藏状态的最大重叠量对隐藏状态序列进行合并,从而得到最终的碱基序列。尽管隐马尔科夫模型非常善于表示短程依赖关系,但难以捕获到碱基识别任务中的长程依赖关系,即无法检测到长度大于 *k*-mer 大小的均聚物,导致识别准确度较低,限制了其在碱基识别中的应用[71]。

考虑到基于隐马尔科夫模型的碱基识别方法存在无法检测长均聚物的弊端,基于循环神经网络(recurrent neural network, RNN) 处理分段事件序列的方法被提出。该方法可显著提高碱基识别准确度[44],如 ONT 公司的 Nanonet 与 Boza 团队研发的 DeepNano。循环神经网络是一种用于序列标记的人工神经网络,给定一个输入向量序列,其预测一个输出向量序列。在碱基识别中,输入向量由每个事件的平均值、标准差和信号长度组成,输出向量则给出了每个事件的碱基概率分布,根据概率分布可以计算得到该事件序列所

对应的碱基序列。在测序过程中的原始电信号主要受到同时占据纳米孔的多个碱基的影响,有关碱基序列的信息包含在当前事件的上游和下游事件中,通常需要使用双向循环神经网络对每个方向的事件进行预测,并将每个事件的两个预测值进行合并。循环神经网络不明确依赖于 k-mer 的长度,并且能将更大范围的信息考虑在内,有效解决了碱基识别的长程依赖关系。

2.2　基于原始电流信号的碱基识别方法

将纳米孔测序原始电流信号分段成离散的事件,然后采用相关的算法将分段事件序列转换为碱基序列是一种简单但容易出错的碱基识别方法。分段算法根据一个窗口内信号值的急剧变化来确定两个事件之间的边界,其中窗口的大小由核苷酸序列在纳米孔中预期的移动速度决定。然而,在测序过程中核苷酸序列的移动速度是可变的,再加上低信噪比的原始电流信号,往往导致分段精度较低。现有的分段算法通常对窗口的大小进行保守估计,使得计算得到的事件数量多于实际测序的碱基数量。尽管可以采用动态规划算法将多个分段事件合并为一个 k-mer 来解决此问题,但仍然会影响最终的预测精度。除了使用基于分段事件的碱基识别方法之外,采用神经网络模型直接将原始电流信号转换为碱基序列也是常用的一种碱基识别方法。

采用神经网络模型处理原始电流信号的一种思路是使用两个独立的循环神经网络来预测碱基序列,如软件 BasecRAWller[43]。第一个循环神经网络将归一化的原始电流信号作为输入,经过一系列计算过程,该网络输出每个信号值作为信号分段边界的概率,同时输出每个信号值所对应的碱基概率。随后根据信号分段边界概率将原始电流信号分段为离散的事件序列,其中每个事件包含了该事件所对应分段信号的碱基概率平均值。然后,第二个循环神经网络将前一循环神经网络计算得到的事件序列作为输入向量,经过处理之后得到的输出向量即为最终预测的碱基序列。虽然上述方法使用原始电流信号作为输入数据类型,但其仍然在第一个循环神经网络之后执行信号分段的步骤。

除了上述思路之外,还可以采用将卷积神经网络(convolutional neural networks,CNN)、循环神经网络和连接时序分类(connectionist temporal classification,CTC)译码器相结合这一新颖的神经网络模型来预测碱基序列[44]。该网络模型可以直接对原始电流信号进行建模,避免了易出错的信号分段步骤。

2.3　碱基识别方法分析比较

为比较不同碱基识别方法的性能,本文对由 ONT 开发的 Albacore、Guppy 和 Scrappie 以及由独立的研究团队开发的 Nanocall、DeepNano 和 Chiron 进行了测试。Albacore 和 Guppy 均采用基于原始电流信号的碱基识别方法,这两个软件只提供给 ONT 的客户使用。而 Scrappie 是一个开源的碱基识别软件,ONT 将其称为“技术演示者”,它是 ONT 尝试新技术的第一个碱基识别软件,随后将新技术融入 Albacore 和 Guppy。Scrappie 实际上是 Scrappie-events 和 Scrappie-raw 两个软件的组合,其中 Scrappie-events 使用神经网络模型将分段事件序列转换为碱基序列,而 Scrappie-raw 则使用神经网络模型直接将原始电流信号转换为碱基序列。Nanocall 和 DeepNano 采用了基于分段事件的碱基识别方法,两者的不同之处在于 Nanocall 使用隐马尔科夫模型,而 DeepNano 使用循环神经网络。Chiron 软件使用了将卷积神经网络、循环神经网络和连接时序分类译码器相结合这一新颖的神经网络模型来分析原始电流信号进而预测碱基序列。

评估碱基识别软件性能优劣的一个重要指标是比较由不同软件识别出的测序读段的错误率,其中包括插入错误率、删节错误率和替代错误率。本文使用由 MinION(R9.4 测序芯片)测序得到的大肠杆菌数据集(http://gigadb.org/dataset/100425)进行性能测试实验,该数据集包含了 4 000 条大肠杆菌测序数据(图 2)。由测试结果可知,对于基于分段事件的碱基识别方法来说,采用循环神经网络的处理方法较隐马尔科夫模型在碱基识别准确度方面有了很大的改进,这主要是因为循环神经网络可以对隐马尔科夫模型无法捕获到的长程依赖关系进行高效地处理。测序读段精度最高的四个碱基识别软件 Guppy、Albacore、Scrappie-raw 和 Chiron 都使用了神经网络模型直接将原始电流信号转换为碱基序列。与基于分段事件的碱基识别方法相比,基于原始电流信号的碱基识别方法将测序读段的错误率平均降低了 7% 左右。碱基识别准确度的提高是由于在信号分段过程中出现的错误很难在后续的数据处理过程中纠正,并且将原始电流信号简化为平均值、标准差和信号长度三个特征时会丢失掉许多重要的信息,而使用神经网络模型直接对原始电流信号进行建模则可以避免该问题。

2.4　插入删节错误建模及其理论基础

三代纳米孔测序以及碱基识别存在的插入删节错

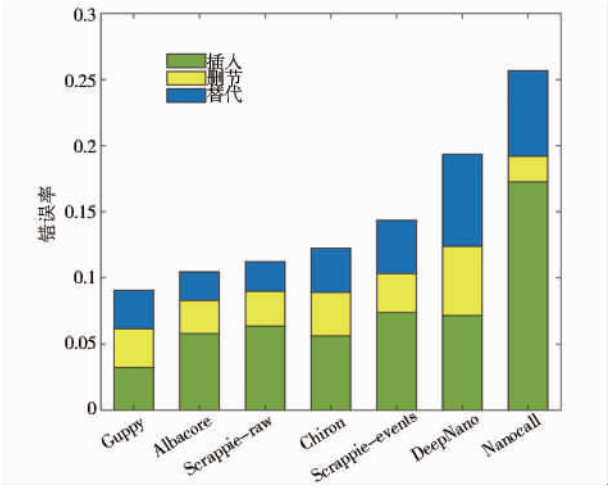

图 2 碱基识别软件性能比较

Fig. 2 The performance comparison of base calling software

误,成为信息理论研究中非常重要的内容。根据碱基识别错误率的统计分析结果可以建立序列的有记忆信道模型。Davey 和 MacKay[72] 提出了比特插入、删节与替代(IDS)有记忆错误模型(图 3)。在第三代纳米孔测序错误场景中,基于“核苷酸空间”A = {A,T,G,C}的测序错误的 IDS 错误模型十分契合图 3 所示的有记忆模型,测序数据通过碱基信道产生的错误表现为生物碱基在合成扩增测序等过程中发生的碱基插入、碱基丢失与碱基替代错误。对于每个传入的碱基 t_i,发生三个事件中的一个。(1)一个随机碱基以概率 P_i 被插入到 t_i 前(最多可连续发生 I 次)。(2) t_i 以概率 P_d 被删掉。(3) t_i 以概率 $P_t = 1 - P_i - P_d$ 等待传输,进一步,以概率 P_s 被其他三种碱基等概率替代。

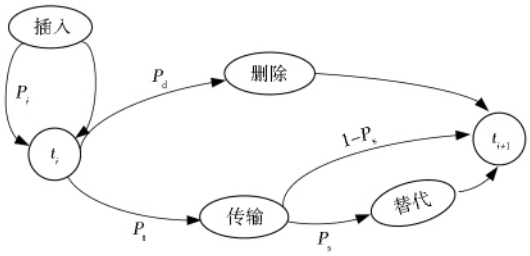

图 3 IDS 有记忆错误模型

Fig. 3 The memory error model for IDS

三代纳米孔测序过程中存在插入与删节错误,需要对三代测序的各个要素进行专门的优化设计。DNA 高通量测序中,可以将多个样本复用进行并行多路测序,不同样本的识别依赖于特定的样本标签,即条形码(barcode)。在测序过程中,用于区分样本的测序条形码也会发生插入与删节错误,导致索引跳变(index hopping),造成测序样本错误分配,影响下游分析[73-75]。文献[73]提出一种可以对抗插入删节错误的测序条形码方案,但是并未针对三代纳米孔测序进行设计。针对三代纳米孔测序,文献[74]和[75]对适合该场景的条形码进行了专门设计,可以有效对抗碱基的插入与删节错误。文献[74]结合插入与删节错误模型,提出将分组纠错码与预先确定的伪随机序列相结合,生成用于标记不同样本的碱基序列,并针对译码提出了一种由内译码和外译码两部分组成的软判决识别方法,以识别被破坏的测序条形码。内译码器利用伪随机序列建立隐马尔科夫模型(HMM)进行碱基插入删节估计,并采用前向-后向(FB)算法输出分组码中每个比特的软信息。外译码器利用软信息进行软判决译码,有效地纠正条形码中的多重错误。文献[75]将循环分组码与预先确定的伪随机序列逐位组合成比特对,然后将比特对转换成碱基生成 DNA 条形码。在此基础上,提出了一种基于循环移位与传统动态规划相结合的测序条形码识别方案,该方案通过标记插入和删除位置,对已损坏的码字进行擦除和纠错解码。仿真结果表明该方法能有效提高插入与删节错误的估计精度且识别插入与删节后的误码率大大降低。

除了在纳米孔测序中的应用,插入删节错误处理策略也广泛存在于各种通信与存储系统,成为信息理论中不可或缺的部分[76]。在纠正插入与删节的级联编码结构中,内码估计发生插入、删除的概率,外码纠正残留插入、删除及替代错误,结合内外译码算法的迭代译码方案被提出,进一步改善了性能[77-79]。针对内译码算法复杂度高的问题,提出了基于自适应删剪网格图的低复杂度译码算法[80]。在非二进制同步错误的纠正中,引入了水印码,可以识别错误的位置与边界[81]。文献[82]设计了一种反转级联水印码的硬判决迭代译码方案纠正插入删节错误,在复杂度较低的条件下获得了优越性能。无线光通信系统中的差分脉冲位置调制(differential pulse-position modulation,DPPM)系统中也存在类似的应用场景,码片因噪声发生跳表引起插入、删节错误。针对 DPPM 系统中出现的插入与删节错误,结合该错误模型提出了不同的纠错译码方法[83-86]。

3 纳米孔测序的仿真模型

随着纳米孔测序技术的发展,下游测序数据的分

析工具也陆续推出。利用精准纳米孔测序仿真模型来对这些新开发的数据分析工具进行性能测试具有重要价值。目前,纳米孔测序模型仿真方法可以分为纳米孔测序读段仿真和纳米孔测序信号仿真两类[87](表 2)。

表 2　纳米孔测序仿真软件比较

Table 2　The comparison of simulation software for nanopore sequencing

仿真软件	程序语言	是否生成仿真读段	是否生成仿真信号	是否模拟噪声特性	开发团队
ReadSim	Python	是	否	否	Lee 等[50]
SiLiCO	Python	是	否	否	Baker 等[51]
NanoSim	Python	是	否	是	Yang 等[52]
DeepSimulator	Python	是	是	是	Li 等[53-54]
NanosigSim	Python	是	是	是	Chen 等[55]

3.1　纳米孔测序读段仿真

测序器件的输出是测序读段,目前的序列比对和序列组装工具都是围绕测序读段设计的。直接生成可以用于后续操作的仿真测序读段是一项重要的工作,尤其在纳米孔测序的仿真研究中。目前,纳米孔测序的读段仿真程序主要有 ReadSim、SiLiCO 和 NanoSim 等。

ReadSim 是能够根据 PacBio 测序仪和纳米孔测序仪的读段长度分布生成仿真读段的仿真软件[56]。ReadSim 最初只被用于仿真 PacBio 测序数据,后将其用于纳米孔测序读段的仿真。然而,ReadSim 并不是专门为纳米孔测序技术而设计的,并且随着纳米孔测序技术的快速发展,该软件已经很难模拟最新纳米孔测序读段的错误分布。SiLiCO 是专门用于纳米孔测序技术的开源读段仿真软件,该研究工作通过分析实际测序读段发现纳米孔测序读段的长度分布遵循伽马分布[51]。SiLiCO 还可以扩展为一种 Monte-Carlo 仿真软件,SiLiCO 可以生成具有真实长度和高可扩展性覆盖率的仿真读段,但无法模拟纳米孔测序读段的错误分布且插入删节错误参数无法调节。

NanoSim 是能同时对纳米孔测序读段的长度分布和错误分布建模的仿真软件[52]。该软件的仿真流程可以分为两个主要步骤。第一个步骤为测序读段的特征提取,NanoSim 通过用户提供的真实测序读段来学习仿真读段的一系列特征,包括插入率、删节率、替代率、读段长度和质量分数等参数。第二个步骤为仿真读段的生成,NanoSim 首先根据学习到的长度分布从输入参考基因组序列中得到具有理想长度的仿真读段,随后建立隐马尔科夫模型来确定仿真读段的错误类型分布并将其引入到仿真读段之中。

3.2　纳米孔测序信号仿真

纳米孔测序的输出是原始电流信号,然而已有的读段仿真软件无法生成仿真信号,限制了已有的利用电信号识别碱基的软件以及对电信号处理的方法与软件的应用。DeepSimulator(v1.0、v1.5)是可以同时生成仿真信号和仿真读段的纳米孔测序仿真软件[53-54]。其工作框架包括序列生成模块、信号生成模块和碱基识别模块。首先,序列生成模块随机选择输入参考基因组序列上的起始位置,生成满足实际测序读段长度分布的较短核苷酸序列。然后,信号生成模块根据已知的纳米孔测序孔模型生成序列生成与核苷酸序列所对应的仿真信号。最后,碱基识别模块使用碱基识别软件(如 Albacore、Guppy)将仿真信号转换为仿真测序读段。在 DeepSimulator 仿真软件主要工作框架中的三个模块中,信号生成模块是核心模块。该模块首先根据已知的纳米孔测序孔模型和实际测序信号的重复次数分布生成输入核苷酸序列所对应的真实背景(ground-truth)信号,其中真实背景信号是一种无噪声的背景测序信号。然后,使用低通滤波器滤掉嵌入在真实背景信号中与实际测序信号无关的高频分量。最后,在滤波信号上添加高斯噪声,输出最终的仿真测序信号。由于低通滤波器会衰减真实背景信号中所有高于截止频率的高频分量,无法保留实际测序信号可能包含的高频分量,对于某些输入核苷酸序列,可能会造成仿真信号与实际测序信号之间存在较大的差异,给信号敏感的一些应用带来挑战。

为进一步提高纳米孔测序信号的仿真精度,NanosigSim 采用了一种基于双向门控循环单元神经网络(bi-directional gated recurrent units,BiGRU)的信号仿真方法[55]。该方法建立基于 BiGRU 的信号处理模型代替传统低通滤波器对由输入核苷酸序列和孔模型计算得到的真实背景信号进行滤波处理,然后在滤波信号上添加高斯噪声生成最终的仿真信号。NanosigSim

软件可以根据实验测序数据来对真实背景(ground-truth)信号和实际测序(real sequencing)信号之间的关系进行准确的建模。仿真结果表明,NanosigSim 利用神经网络强大的学习能力,能生成在时域和频域上更接近于真实测序信号的仿真信号,有效提高了现有仿真方法的准确度。

3.3 纳米孔测序仿真平台搭建

现有的纳米孔测序仿真软件多是基于命令进行工作的,在实际应用中较为不便。本文采用 Python 提供的 Tkinter 模块完成了纳米孔测序仿真平台的搭建(图 4)。软件可以利用用户提供的参考基因组序列进行参数训练,拟合错误模型,分析读长、错误率等统计特征。然后,通过序列生成模块,按照设定的参数,输出为 FAST5 格式的仿真信号数据与 FASTA/FASTQ 格式的仿真测序数据。最后,分析模块对生成的仿真信号与仿真读段进行分析。纳米孔测序仿真平台的搭建可使用户能够高效地操作,避免繁琐的底层命令的调用,利于彼此之间的互动、沟通,具有良好的交互性与易用性。同时,仿真平台使用 Python 语言实现,具有较高的效率,可移植性强,可以直接扩展应用于未来实际的数据读取软件。

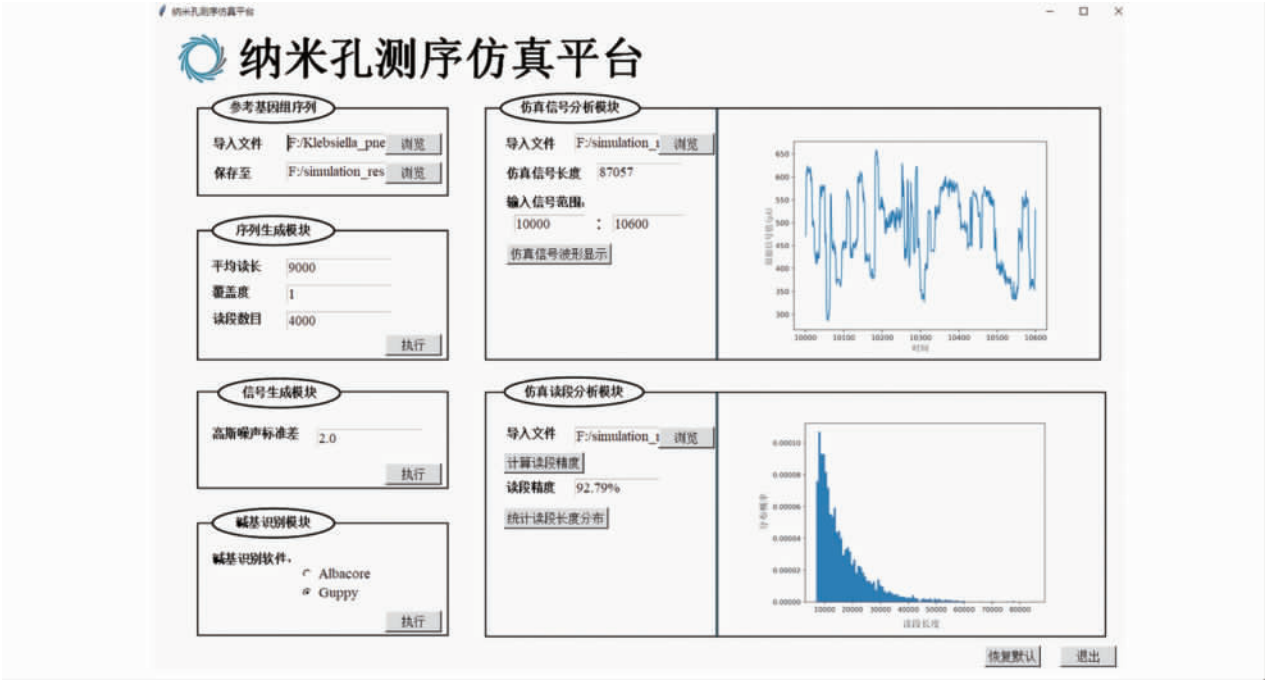

图 4 纳米孔测序仿真平台

Fig. 4 The simulation platform for nanopore sequencing

4 纳米孔测序信号建模与处理在 DNA 数据存储中的应用

DNA 数据存储是指通过编码技术将二进制信息转换成 DNA 序列信息,利用 DNA 合成技术实现信息写入,测序技术读取 DNA 序列中的信息,进而通过译码得到原始信息[63]。相比于传统的光盘、磁盘以及固态硬盘存储方式,DNA 数据存储读写速度有限[59]。纳米孔测序的读段长、测序速度快,与基于二代测序的读出方式相比,具有一定的优势[59,88]。随着纳米孔测序技术的发展,尤其是通量的提高,其快速读出的优势与轻巧便携性,使其成为未来 DNA 数据存储极具潜力的读出方式[62]。

文献[56]中初步验证了纳米孔测序用于 DNA 数据存储读出的可能性。该设计方法采用纠错能力非常强的里德-所罗门(Reed-Solomon,RS)码,并且译码之前采用聚类、一致性序列合并算法对测序读段进行合并纠错,消除了插入删节错误,实现数据可靠读出。在此基础上,该研究组对上述方法进行了深入研究,将寡核苷酸池的短序列进行了组装,支持三代纳米孔测序仪,设计并验证了一种采用寡核苷酸池存储的先组装后三代纳米孔测序的策略,成功解码了 1.67 Mb 的存储在合成 DNA 短片段中的信息[57]。文献[58]采用现代通

信领域广泛应用的低密度奇偶校验(low-density parity-check,LDPC)码叠加伪随机序列,设计了可纠正严重插入删节错误的高效编码方案,从头编码设计合成了一条长度为 254 886 bp 专用于数据存储的酵母人工染色体。该设计借助无线通信中的 LDPC 码,并将其稀疏化。然后,对两张图片和一段视频进行编码,得到用于数据存储的 DNA 序列。进一步,实际构建了高效组装的人工染色体,实验展示了该人工染色体可以利用酵母增殖实现数据的稳定复制。在读出方面,利用三代纳米孔测序器件实现了碱基的快速读出与无错恢复。设计的存储方法直接基于大片段 DNA,可方便地提取核酸建库。碱基识别后的错误率高于 10%,包含严重的插入删节错误。为处理这些插入删节错误,设计了一个融合生物信息处理中的组装与纠错的方案,进一步结合设计的可纠正插入与删节错误的纠错码,实现了数据无错恢复。在该设计中,设计了两种不同码率的编码方法,一种采用码率为 1/2 的多进制 LDPC 码,另一种采用码率为 5/6 的二进制 LDPC 码,验证了提出的数据读出方案的可靠性。在测序实验中,为方便展示设计方案的快速读出功能,开发了专用的基于三代纳米孔测序的全流程恢复软件系统,该软件支持从纳米孔测序数据到原始文件的无错恢复(图 5)。

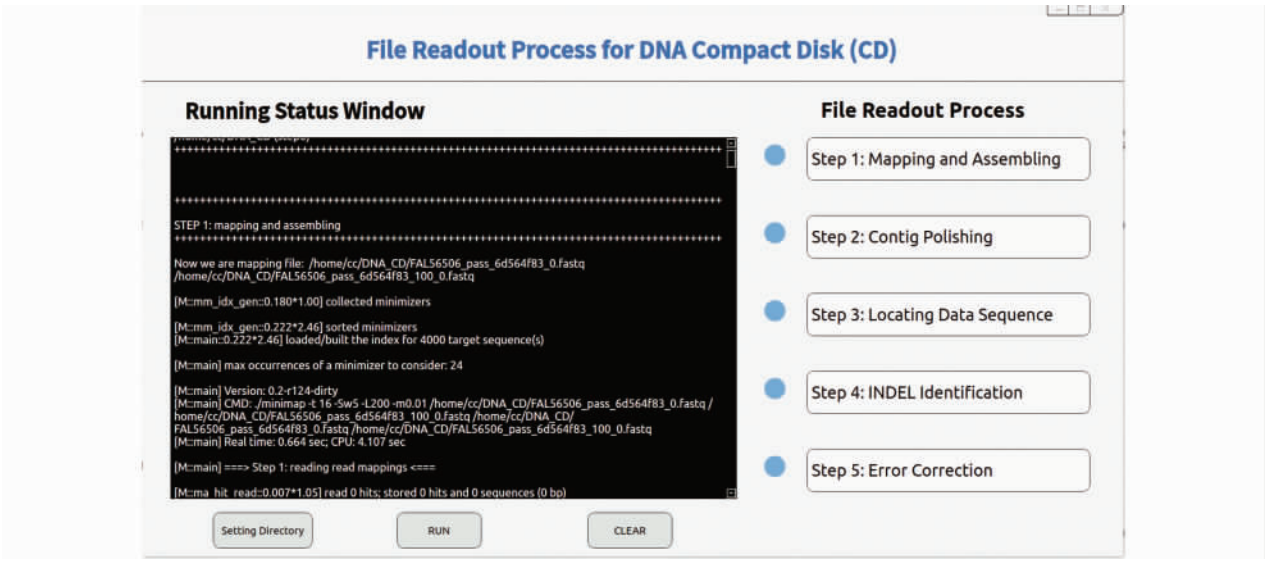

图 5　数据存储专用人工染色体的数据读出验证平台[58]

Fig. 5　The verification platform for data readout from the encoded artificial chromosome specific for data storage

目前 DNA 合成的成本高、耗时长、实验复杂,在寻求 DNA 信息存储编译码方案的最优参数以及研究纳米孔测序的错误特性时,需要进行多次实验。由于实际"湿"实验较为复杂,频繁进行实验测试性价比较低,而通过计算机仿真技术可以进行信道建模,同时也可以利用仿真结果的可重复性,分析影响存储系统性能的主要参数。因此,纳米孔仿真模型在 DNA 数据存储研究中可以用于存储系统优化,提高了存储系统设计的效率。在文献[58]中,基于读段仿真软件 NanoSim,对酵母人造染色体数据存储的仿真读段进行分析,开展实际测序结合纳米孔测序仿真在 DNA 数据存储中的可行性验证,以人工合成的 254 886 bp 的人工染色体测序得到的实际测序数据为训练模板,设计了 DNA 数据存储的全流程仿真验证(图 6)。首先,一组真实测序数据送入三代测序仿真软件 NanoSim[52]进行参数训练,模拟近似真实环境的错误特性。然后,随机生成的比特序列映射为碱基序列,组装成一条长度为 254 kb 的虚拟碱基序列。进一步,将设计的长序列输入测序数据仿真单元,生成测序数据,利用测序读段组装与译码恢复策略进行测序数据的恢复验证[58]。最后,对仿真的数据进行端对端的性能评估,并与真实的测序数据错误性能进行对比。

仿真数据的恢复单元主要分为序列比对与组装、序列纠错、序列定位、纠正插入删节和译码纠错等步骤。序列组装是指根据测序读段之间的重叠区域对测序读段片段进行拼接组装。本节使用长序列比对工具 Minimap 和组装工具 Miniasm 对未处理的长读段进行序列比对和快速组装[89],得到较长的重叠群(contig)。

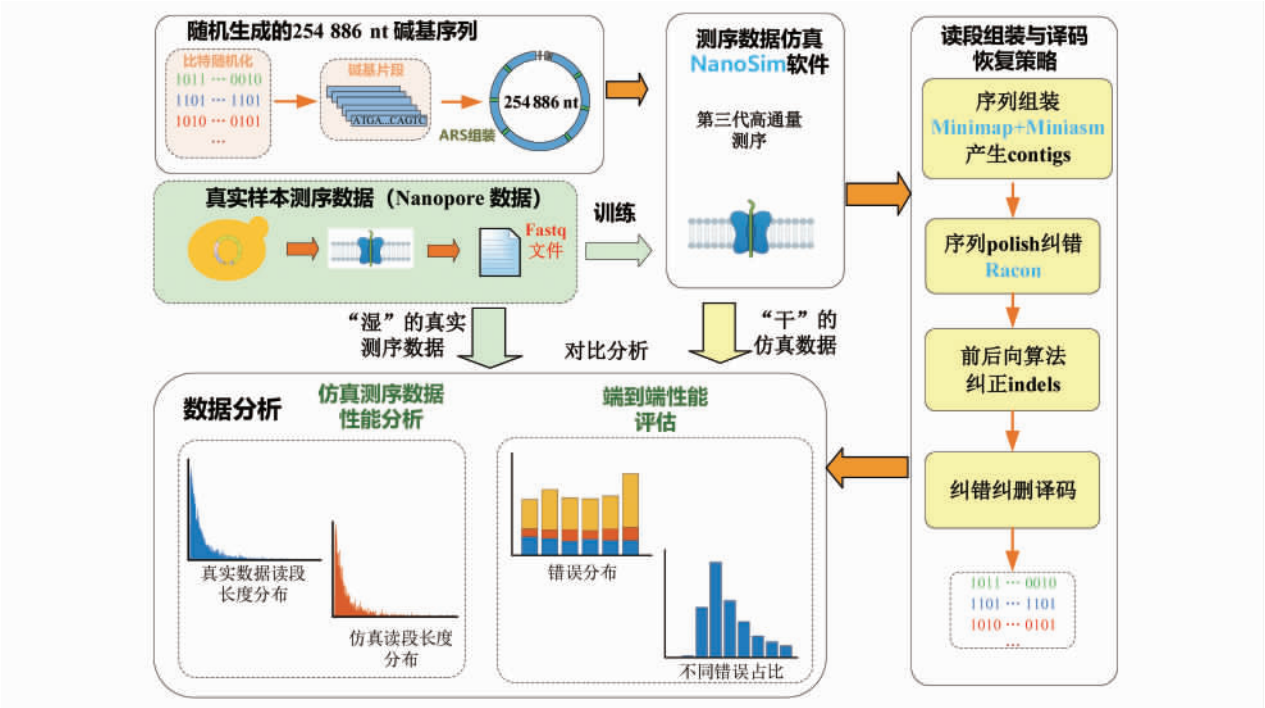

图 6　利用实际测序数据训练的基于三代纳米孔测序的 DNA 数据存储仿真流程

Fig. 6　The simulation of nanopore sequencing based DNA data storage using real sequencing data for training

得到的重叠群序列中存在大量的插入与删节错误，利用 RACON 软件[35]进行纠错处理。进一步，利用文献[58]提出的序列定位，前后向算法纠正插入删节错误，译码算法恢复数据。

进一步，为验证纳米孔仿真软件与读段组装软件的可靠性，进行了仿真数据与真实数据混合组装的仿真研究（图 7）。首先利用设计的编码方案，按照文献中的人工染色体的结构[58]，从头设计多条等长的长序列，并利用 NanoSim 仿真产生测序读段文件。然后，将其与真实上机产出的 4 000 条测序数据文件混合在一起，组合成拥有多种测序读段的文件。进一步，利用组装软件对该测序文件进行组装，组装过程中，Minimap 和 Miniasm 软件可以根据不同测序读段之间的区别，对读段进行区分与拼接，实验证明经过 RACON 的抛光纠错过程后，成功地组装成长序列并将其区分。最后，利用定位、纠错、译码算法完成文件的恢复，成功还原为原始图片文件。

5　总结与展望

纳米孔测序作为新一代测序技术，可以用于普通核酸样本的测序，同时在碱基修饰检测、实时测序监控、结构变异检测、RNA 表达分析以及 DNA 数据存储信息读出等领域均得到了广泛应用。本文对近年来基于分段事件的碱基识别方法与基于原始电流信号的碱基识别方法的最新进展进行了综述，并对纳米孔测序的错误模型以及基于该模型的信号处理方法进行了总结。然后，对纳米孔测序信号仿真与读段仿真两种仿真方法进行了综述，介绍了在 DNA 数据存储中的应用进展，开展了基于实际测序数据训练的 DNA 数据存储读段仿真研究。

碱基识别算法性能是决定纳米孔测序质量和可用性的关键因素。为进一步提高碱基识别准确度和识别速度，一方面，碱基识别神经网络模型必须使用真实测序训练数据来完成模型参数的训练。另一方面，目前大多碱基识别方法都使用了基于循环神经网络的预测模型将原始电流信号直接转换为碱基序列。在循环神经网络的结构中，一个时间点的计算必须等待前一个时间点的结果，这使其在处理长信号序列时的识别速度较慢，未来可通过对信号序列分段并行处理以及简化循环神经网络结构来加快识别速度。纳米孔仿真模型及其软件在测序数据分析处理方法的开发测试中有重要作用。未来可结合更准确的纳米孔测序反应体系的参数与机制，建立更准确、更合理的纳米孔仿真模型，推动纳米孔测序技术的发展，也为下游分析提供更

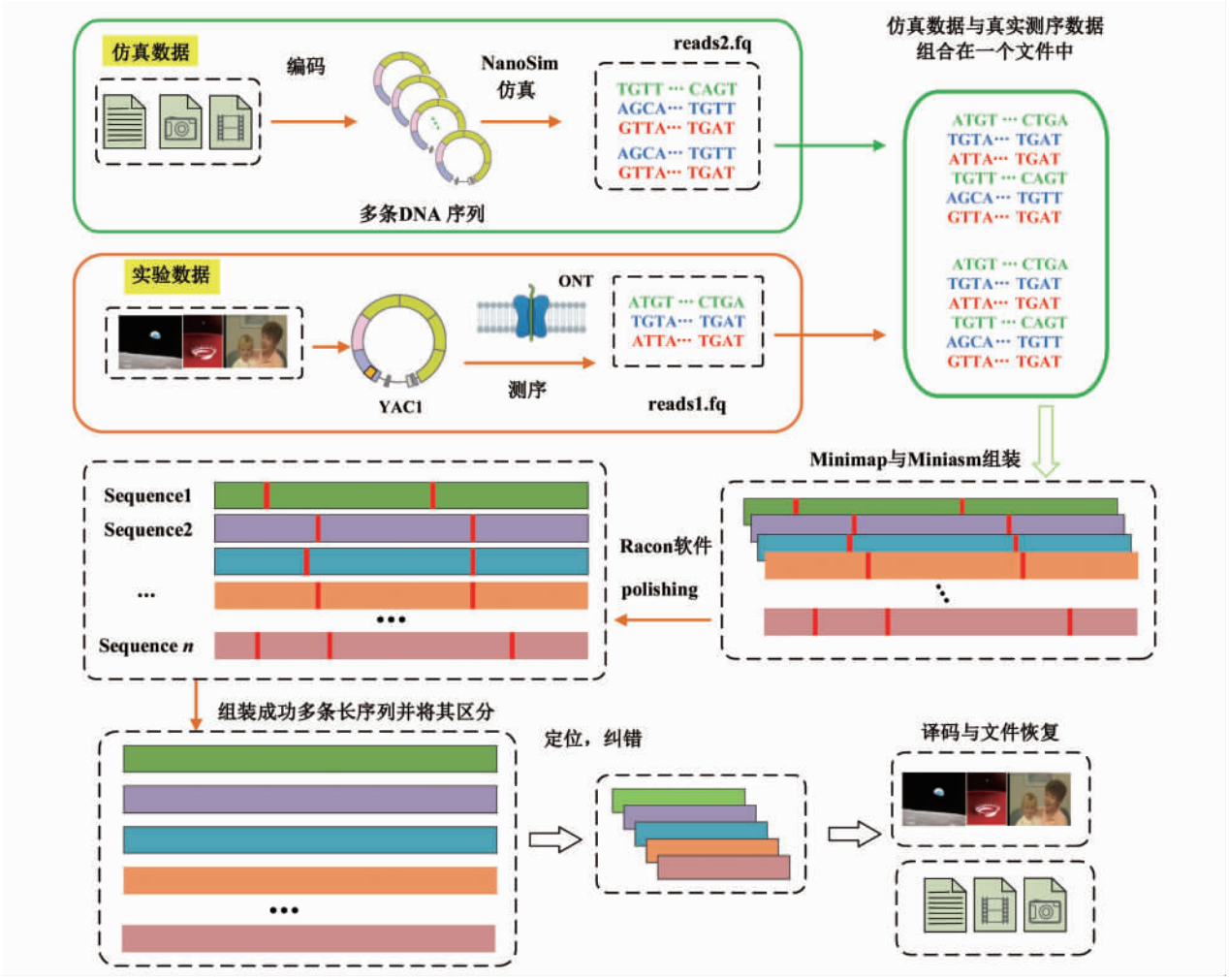

图 7　基于纳米孔测序的仿真数据的寻址方式研究

Fig. 7　The addressing method of simulation data based on nanopore sequencing

为准确的仿真数据。在 DNA 数据存储领域,集成碱基识别方法、纳米孔测序器件仿真可搭建 DNA 数据存储的全流程读出仿真系统,降低 DNA 数据存储研发的成本,提高效率。同时,也为研究 DNA 数据存储专用纳米孔器件打下基础。

参考文献

[1] Kasianowicz J J, Bezrukov S M. On 'three decades of nanopore sequencing'. Nature Biotechnology, 2016, 34(5): 481-482.

[2] National Institute of Standards and Technology (NIST), Semiconductor Research Corporation (SRC). 2018 Semiconductor synthetic biology roadmap. [2021-03-15]. https://www.src.org/library/publication/p095387/p095387.pdf.

[3] Semiconductor Industry Association (SIA), Semiconductor Research Corporation (SRC). Decadal plan for semiconductors. [2021-03-15]. https://www.src.org/about/decadal-plan/.

[4] 伊克巴尔 S M, 巴希尔 R, 德克 C, 等. 纳米孔: 生物分子相互作用传感基础. 刘全俊, 陆祖宏, 谢骁, 等. 译. 北京: 科学出版社, 2013: 1-7.
Iqbal S M, Bashir R, et al. Nanopores: sensing and fundamental biological interactions. Liu Q J, Lu Z H, Xie X, et al. Beijing: Science Press, 2013: 1-7.

[5] 鞠熀先, 张学记, 约瑟夫 W. 纳米生物传感: 原理, 发展与应用. 雷建平, 吴洁, 鞠熀先. 译. 北京: 科学出版社, 2012: 1-8.
Ju H X, Zhang X J, Joseph W. NanoBiosensing: principles, development and application. Lei J P, Wu J, Ju H X. Beijing: Science Press, 2012: 1-8.

[6] 余静文, 陈云飞. 基于微纳制造的下一代基因测序系统研究现状与展望. 中国科学: 技术科学, 2017, 47(4): 345-354.
Yu J W, Chen Y F. Research status and prospects of next generation sequencing system based on micro-nano manufacturing.

Scientia Sinica (Technologica) , 2017, 47(4) : 345-354.

[7] 陈文辉, 罗军, 赵超. 固态纳米孔: 下一代 DNA 测序技术: 原理、工艺与挑战. 中国科学: 生命科学, 2014, 44(7) : 649-662.
Chen W H, Luo J, Zhao C. Solid-state nanopore: the next-generation sequencing technology-principles, fabrication and challenges. Scientia Sinica (Vitae) , 2014, 44(7) : 649-662.

[8] 张宇, 魏胜, 李民权, 等. 用于单个纳米颗粒检测的固态纳米孔器件的仿真与优化. 传感技术学报, 2015, 28(10) : 1425-1431.
Zhang Y, Wei S, Li M Q, et al. Simulation and optimization of solid-state nanopore for single-nanoparticle detection. Chinese Journal of Sensors and Actuators, 2015, 28(10) : 1425-1431.

[9] 张庞, 唐鹏, 闫汉, 等. 基于 LiCl 盐浓度梯度的固态纳米孔 DNA 分子检测. 微纳电子技术, 2021, 58(1) : 72-79.
Zhang P, Tang P, Yan H, et al. Detection of DNA molecule with solid-state nanopores based on LiCl salt concentration gradient. Micronanoelectronic Technology, 2021, 58(1) : 72-79.

[10] Guo B Y, Zeng T, Wu H C. Recent advances of DNA sequencing via nanopore-based technologies. Science Bulletin, 2015, 60 (3) : 287-295.

[11] 丁克俭, 张海燕, 胡红刚, 等. 生物大分子纳米孔分析技术研究进展. 分析化学, 2010, 38(2) : 280-285.
Ding K J, Zhang H Y, Hu H G, et al. Progress of research on nanopore-macromolecule detection. Chinese Journal of Analytical Chemistry, 2010, 38(2) : 280-285.

[12] Yuan Z S, Wang C Y, Yi X, et al. Solid-state nanopore. Nanoscale Research Letters, 2018, 13(1) : 1-10.

[13] Deng T, Li M W, Wang Y F, et al. Development of solid-state nanopore fabrication technologies. Science Bulletin, 2015, 60 (3) : 304-319.

[14] Chen Q, Liu Z W. Fabrication and applications of solid-state nanopores. Sensors, 2019, 19(8) : 1886.

[15] Luan B Q, Bai J W, Stolovitzky G. Fabricatable nanopore sensors with an atomic thickness. Applied Physics Letters, 2013, 103 (18) : 183501.

[16] 陈剑, 邓涛, 吴次南, 等. 面向新型 DNA 检测方法的固态纳米孔研究进展. 微纳电子技术, 2013, 50(3) : 143-150.
Chen J, Deng T, Wu C N, et al. Research progress of solid-state nanopores for the new DNA detection method. Micronanoelectronic Technology, 2013, 50(3) : 143-150.

[17] Garalde D R, Snell E A, Jachimowicz D, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nature Methods, 2018, 15(3) : 201-206.

[18] Faria N R, Sabino E C, Nunes M R T, et al. Mobile real-time surveillance of Zika virus in Brazil. Genome Medicine, 2016, 8 (1) : 1-4.

[19] Stancu M C, van Roosmalen M J, Renkens I, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nature Communications, 2017, 8: 1326.

[20] Stoloff D H, Wanunu M. Recent trends in nanopores for biotechnology. Current Opinion in Biotechnology, 2013, 24(4) : 699-704.

[21] Wanunu M. Nanopores: a journey towards DNA sequencing. Physics of Life Reviews, 2012, 9(2) : 125-158.

[22] Wescoe Z L, Schreiber J, Akeson M. Nanopores discriminate among five C5-cytosine variants in DNA. Journal of the American Chemical Society, 2014, 136(47) : 16582-16587.

[23] Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nature Methods, 2016, 13(9) : 751-754.

[24] Norris A L, Workman R E, Fan Y F, et al. Nanopore sequencing detects structural variants in cancer. Cancer Biology & Therapy, 2016, 17(3) : 246-253.

[25] Quick J, Loman N J, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature, 2016, 530 (7589) : 228-232.

[26] Wang M, Fu A S, Hu B, et al. Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. medRxiv, 2020. DOI: 10. 1101/2020. 03. 04. 20029538.

[27] Chan J F W, Yuan S F, Kok K H, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet, 2020, 395(10223) : 514-523.

[28] Prazsák I, Moldován N, Balázs Z, et al. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. BMC Genomics, 2018, 19(1) : 873.

[29] Wee Y, Bhyan S B, Liu Y N, et al. The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. Briefings in Functional Genomics, 2019, 18(1) : 1-12.

[30] Jain M, Fiddes I T, Miga K H, et al. Improved data analysis for the MinION nanopore sequencer. Nature Methods, 2015, 12 (4) : 351-356.

[31] Shabardina V, Kischka T, Manske F, et al. NanoPipe: a web server for nanopore MinION sequencing data analysis. GigaScience, 2019, 8(2) : giy169.

[32] Loman N J, Quick J, Simpson J T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nature Methods, 2015, 12(8) : 733-735.

[33] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 2018, 34(18) : 3094-3100.

[34] Koren S, Walenz B P, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and

repeat separation. Genome Research, 2017, 27(5): 722-736.

[35] Vaser R, Sović I, Nagarajan N, et al. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research, 2017, 27(5): 737-746.

[36] Ferguson J M, Smith M A. SquiggleKit: a toolkit for manipulating nanopore signal data. Bioinformatics, 2019, 35(24): 5372-5373.

[37] Wick R R, Judd L M, Holt K E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. bioRxiv, 2019, DOI: 10.1101/543439.

[38] Leggett R M, Clark M D. A world of opportunities with nanopore sequencing. Journal of Experimental Botany, 2017, 68(20): 5419-5429.

[39] Jain M, Koren S, Miga K H, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology, 2018, 36(4): 338-345.

[40] Wang L T, Qu L, Yang L S, et al. NanoReviser: an error-correction tool for nanopore sequencing based on a deep learning algorithm. Frontiers in Genetics, 2020, 11: 900. DOI: 10.3389/fgene.2020.00900.

[41] David M, Dursi L J, Yao D L, et al. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. Bioinformatics, 2017, 33(1): 49-55.

[42] Boža V, Brejová B, Vinař T. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLoS One, 2017, 12(6): e0178751. DOI: 10.1371/journal.pone.0178751.

[43] Stoiber M, Brown J. BasecRAWller: streaming nanopore basecalling directly from raw signal. bioRxiv, 2017, DOI: 10.1101/133058.

[44] Teng H T, Cao M D, Hall M B, et al. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience, 2018, 7(5): giy037.

[45] Rang F J, Kloosterman W P, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biology, 2018, 19(1): 90.

[46] Goodwin S, McPherson J D, Richard McCombie W. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 2016, 17(6): 333-351.

[47] Li Y, Huang C, Ding L Z, et al. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods, 2019, 166: 4-21.

[48] Makałowski W, Shabardina V. Bioinformatics of nanopore sequencing. Journal of Human Genetics, 2020, 65(1): 61-67.

[49] Yue J X, Liti G N. SimuG: a general-purpose genome simulator. Bioinformatics, 2019, 35(21): 4442-4444.

[50] Lee H, Gurtowski J, Yoo S, et al. Error correction and assembly complexity of single molecule sequencing reads. bioRxiv, 2014. DOI: 10.1101/006395.

[51] Baker E A G, Goodwin S, Richard McCombie W, et al. SiLiCO: a simulator of long read sequencing in PacBio and Oxford nanopore. bioRxiv, 2016. DOI: 10.1101/076901.

[52] Yang C, Chu J, Warren R L, et al. NanoSim: nanopore sequence read simulator based on statistical characterization. GigaScience, 2017, 6(4): gix010.

[53] Li Y, Han R M, Bi C W, et al. DeepSimulator: a deep simulator for nanopore sequencing. Bioinformatics, 2018, 34(17): 2899-2908.

[54] Li Y, Wang S, Bi C W, et al. DeepSimulator1.5: a more powerful, quicker and lighter simulator for nanopore sequencing. Bioinformatics, 2020, 36(8): 2578-2580.

[55] Chen W G, Zhang P, Song L F, et al. Simulation of nanopore sequencing signals based on BiGRU. Sensors, 2020, 20(24): 7244.

[56] Organick L, Ang S D, Chen Y J, et al. Random access in large-scale DNA data storage. Nature Biotechnology, 2018, 36(3): 242-248.

[57] Lopez R, Chen Y J, Ang S D, et al. DNA assembly for nanopore data storage readout. Nature Communications, 2019, 10: 2933.

[58] Chen W G, Han M Z, Zhou J T, et al. An artificial chromosome for data storage. National Science Review, 2021, 8(5). DOI: 10.1093/nsr/nwab028.

[59] Ceze L, Nivala J, Strauss K. Molecular digital data storage using DNA. Nature Reviews Genetics, 2019, 20(8): 456-466.

[60] Dong Y M, Sun F J, Ping Z, et al. DNA storage: research landscape and future prospects. National Science Review, 2020, 7(6): 1092-1107.

[61] 丁明珠, 李炳志, 王颖, 等. 合成生物学重要研究方向进展. 合成生物学, 2020, 1(1): 7-28.
Ding M Z, Li B Z, Wang Y, et al. Significant research progress in synthetic biology. Synthetic Biology Journal, 2020, 1(1): 7-28.

[62] 韩明哲, 陈为刚, 宋理富, 等. DNA 信息存储: 生命系统与信息系统的桥梁. 合成生物学, 2021, 2(3): 309-322.
Han M Z, Chen W G, Song L F, et al. DNA information storage: bridging biological and digital world. Synthetic Biology Journal, 2021, 2(3): 309-322.

[63] 钱珑, 沈玥, 元英进, 等. DNA 数字信息存储: 造梦, 追梦与圆梦. 合成生物学, 2021, 2(3): 303-304.
Qian L, Shen Y, Yuan Y J, et al. DNA digital information storage: dreaming, chasing and realizing. Synthetic Biology Journal, 2021, 2(3): 303-304.

[64] 陈为刚, 葛奇, 王盼盼, 等. 细胞内大片段 DNA 数据存储的多 RS 码交织编码. 合成生物学, 2021, 2(3): 428-443.

Chen W G, Ge Q, Wang P P, et al. Multiple interleaved RS codes for data storage using up to Mb-scale synthetic DNA in living cells. Synthetic Biology Journal, 2021, 2(3): 428-443.

[65] 陈为刚, 黄刚, 李炳志, 等. 音视频文件的 DNA 信息存储. 中国科学: 生命科学, 2020, 50(1): 81-85.
Chen W G, Huang G, Li B Z, et al. DNA information storage for audio and video files. Scientia Sinica (Vitae), 2020, 50(1): 81-85.

[66] Varongchayakul N, Song J X, Meller A, et al. Single-molecule protein sensing in a nanopore: a tutorial. Chemical Society Reviews, 2018, 47(23): 8512-8524.

[67] Jain M, Olsen H E, Paten B, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology, 2016, 17(1): 1-11.

[68] Magi A, Giusti B, Tattini L. Characterization of MinION nanopore data for resequencing analyses. Briefings in Bioinformatics, 2017, 18(6): 940-953.

[69] Goodwin S, Gurtowski J, Ethe-Sayers S, et al. Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. Genome Research, 2015, 25(11): 1750-1756.

[70] Smith M, Chan R, Gordon P. Evaluation of simulation models to mimic the distortions introduced into squiggles by nanopore sequencers and segmentation algorithms. PLoS One, 2019, 14(7): e0219495. DOI: 10.1371/journal.pone.0219495.

[71] Schreiber J, Karplus K. Analysis of nanopore data using hidden Markov models. Bioinformatics, 2015, 31(12): 1897-1903.

[72] Davey M C, MacKay D J C. Reliable communication over channels with insertions, deletions, and substitutions. IEEE Transactions on Information Theory, 2001, 47(2): 687-698.

[73] Hawkins J A, Jones S K Jr, Finkelstein I J, et al. Indel-correcting DNA barcodes for high-throughput sequencing. PNAS, 2018, 115(27): E6217-E6226. DOI: 10.1073/pnas.1802640115.

[74] Chen W G, Wang L X, Han M Z, et al. Sequencing barcode construction and identification methods based on block error-correction codes. Science China Life Sciences, 2020, 63(10): 1580-1592.

[75] Chen W G, Wang P P, Wang L X, et al. Low-complexity and highly robust barcodes for error-rich single molecular sequencing. 3 Biotech, 2021, 11(2): 1-11.

[76] Mercier H, Bhargava V K, Tarokh V. A survey of error-correcting codes for channels with symbol synchronization errors. IEEE Communications Surveys & Tutorials, 2010, 12(1): 87-96.

[77] Liu Y, Chen W G. Iterative decoding for the concatenated code to correct nonbinary insertions/deletions. 2017 IEEE 85th Vehicular Technology Conference (VTC Spring). Sydney, NSW, Australia. IEEE, 2017: 1-5.

[78] Liu Y, Chen W G. An iterative decoding scheme for Davey-MacKay construction. China Communications, 2018, 15(6): 187-195.

[79] Liu Y, Chen W G. Hard-decision iterative decoder for the Davey-MacKay construction with symbol-level inner decoder. Electronics Letters, 2016, 52(12): 1026-1028.

[80] Liu Y, Chen W G. Decoding on adaptively pruned trellis for correcting synchronization errors. China Communications, 2017, 14(7): 1-9.

[81] 柳元, 陈为刚, 杨晋生. 针对非二进制同步错误的高效水印调制方案. 信号处理, 2017, 33(8): 1034-1039.
Liu Y, Chen W G, Yang J S. Efficient watermark modulation schemes for correcting non-binary synchronization errors. Journal of Signal Processing, 2017, 33(8): 1034-1039.

[82] 张林林, 陈为刚, 刘敬浩, 等. 纠正同步错误的反转级联水印码的迭代译码. 信号处理, 2017, 33(2): 144-151.
Zhang L L, Chen W G, Liu J H, et al. Iterative decoding of the reverse concatenated watermark code for correcting synchronization errors. Journal of Signal Processing, 2017, 33(2): 144-151.

[83] 柳元. 插入/删节错误纠错码的研究. 天津: 天津大学, 2017.
Liu Y. Research on insertions/deletions correcting codes. Tianjin: Tianjin University, 2017.

[84] 张译方, 陈为刚. 纠正 DPPM 中插入删节错误的纠错码方案. 信息技术, 2014, 38(8): 29-33.
Zhang Y F, Chen W G. Coding for correcting insertion/deletion errors in differential pulse-position modulation. Information Technology, 2014, 38(8): 29-33.

[85] Chen W G, Liu Y. Efficient transmission schemes for correcting insertions/deletions in DPPM. 2016 IEEE International Conference on Communications (ICC). Kuala Lumpur, Malaysia. IEEE, 2016: 1-6.

[86] Chen W G, Wang L X, Han C C. Correcting insertions/deletions in DPPM using hidden Markov model. IEEE Access, 2020, 8: 46417-46426.

[87] Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data. Nature Reviews Genetics, 2016, 17(8): 459-469.

[88] Press W H, Hawkins J A, Jones S K, et al. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. PNAS, 2020, 117(31): 18489-18496.

[89] Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics, 2016, 32(14): 2103-2110.

Signal Processing for Nanopore Sequencing and Its Application in DNA Data Storage

GE Qi[1]　ZHANG Peng[1]　HAN Ming-zhe[2,3]　YANG Jin-sheng[1]　ZHANG Da-lu[4]　CHEN Wei-gang[1,3]

(1 School of Microelectronics, Tianjin University, Tianjin　300072, China)

(2 School of Chemical Engineering and Technology, Tianjin University, Tianjin　300072, China)

(3 Frontiers Science Center for Synthetic Biology (MOE), Tianjin University, Tianjin　300072, China)

(4 China National Center for Biotechnology Development, Beijing　100039, China)

Abstract　With the continuous update of high-throughput sequencing technologies, the third-generation sequencing technology that can read nucleotide sequences at the single-molecule level has developed rapidly. Nanopore sequencing technology is its representative single-molecule sequencing technology, which realizes base calling by detecting the characteristic changes of electrical current when the DNA single-stranded molecule is passing through a nanopore channel. Compared with the traditional first-generation and the next-generation sequencing (NGS) technologies, the nanopore sequencing of DNA has great advantages in device portability, base acquisition speed and read length, which has attracted much attention. With the continuous development of nanopore sequencing technologies, various signal processing schemes and biological information processing tools for nanopore sequencing have been developed, and base calling and model simulation are two of the key research directions. The fundamental principle and signal processing flow of nanopore sequencing are surveyed, the current challenges are discussed, then the development trend of base calling and nanopore model simulation in recent years are summarized, and the performance of different base calling methods are compared by using real sequencing reads. Then, an integrated simulation platform for the evaluation of signal processing algorithms of nanopore sequencing is developed. Furthermore, with the explosive growth of global data volume, DNA data storage is becoming a promising medium for future massive data storage, and the use of nanopore for sequencing and reading is a very effective method. The application progress of the nanopore sequencing technology for DNA data storage is summarized, and its feasibility is analyzed. The rapid readout method of artificial chromosome data storage based on nanopore sequencing is analyzed, and the application of the simulation of nanopore sequencing reads combined with actual sequencing data in DNA data storage is discussed, which provides a reference for the development of a suitable DNA data storage program.

Keywords　Nanopore sequencing　Base calling　Nanopore signal processing　DNA data storage