针对 DNA 语言模型的混合分词策略：使用字节对编码和 K-MER 方法

Ganesh Sapkota*, Md. Hasibur Rahman *
Missouri University of Science and Technology
Rolla, MO, USA
{ gsapkota, mrpk9 } @mst.edu

Abstract

本文提出了一种新颖的混合标记策略，通过结合 6-mer 标记法和字节对编码（BPE-600），提升了 DNA 语言模型（DLMs）的性能。传统的 k-mer 标记法在捕捉局部 DNA 序列结构方面效果显著，但常常面临诸如标记分布不均和对全局序列上下文理解有限等挑战。为解决这些局限性，我们建议将独特的 6-mer 标记与通过 600 个 BPE 循环生成的最佳选择的 BPE 标记合并。这种混合方法确保了一个平衡且具有上下文意识的词汇表，使模型能够同时捕捉 DNA 序列中的短期和长期模式。在此混合词汇基础上训练的 DLM 通过下一 k-mer 预测作为微调任务进行评估，表现出显著提升的性能。模型针对 3-mer、4-mer 和 5-mer 的预测准确率分别达到了 10.78 %、10.1 % 和 4.12 %，优于 NT、DNABERT2 和 GROVER 等最先进的模型。这些结果突出了混合标记策略在 DNA 建模中同时保留局部序列结构和全局上下文信息的能力。该工作强调了先进标记方法在基因组语言建模中的重要性，并为未来在下游 DNA 序列分析和生物研究中的应用奠定了坚实的基础。

1. 介绍

DNA 序列在某些方面可能类似于语言，包含可以比拟为语法、句法和语义的元素，但它们也面临独特的挑战。与人类语言不同，DNA 序列没有明显的“词”或特定的方向，除非是在研究转录或复制等生物过程时。这使得直接将语言模型应用于基因组数据变得困难。

研究人员开发了使用不同策略来解决这个问题的 DNA 语言模型（DLMs）。HyenaDNA [?] 提出了一个能够在单核苷酸层面处理一百万个标记的模型，推动了基因组学领域内上下文学习的边界。某些模型，如 Enformer [?] ，通过在特定任务（如预测基因表达）中使用卷积层与 transformer 块结合，避免了定义“单词”的需要。另一方面，基础模型如 DNABERT [?] 、Nucleotide Transformer（NT）[?] 和 LOGO [?] 采取了不同的方法。这些模型在基于双向编码器表示的 transformers（BERT）[?] 架构上进行预训练以预测被掩盖的标记，然后针对特定的基因组任务进行微调。这个过程需要为 DNA 创建一个“单词”的词汇表，每个模型使用其自己的方法来定义这些标记。

DNAGPT [?] 使用了经过修改的 GPT 架构，用于 DNA 序列分类和鸟嘌呤-胞嘧啶含量的回归任务。Caduceus [?] 使用了由 [?] 提出的长范围 Mamba 块，解决了如反向互补等变等挑战，而 VQDNA [?] 采用了带有矢量量化代码簿的卷积编码器以实现最佳的标记化。

传统方法，如 k-mer 分词，在像 NT 和 DNABERT 这样的模型中被广泛使用。通过将序列分割成固定长度 k 的重叠子串（k-mers），有效地保留了序列的局部顺序结构。然而，使用固定长度的 k-mers，模型本质上学习的是令牌序列，并难以理解更广泛的上下文。这往往会导致像不均匀的令牌分布这样的问题，从而引发稀有词问题和有偏训练。

为了解决这些问题，DNABERT2 [?] 引入了一种更灵活的方法，称为字节对编码（BPE）[?] ，通过将序列分解成单个字符，并迭代地合并最频繁的相邻字符对或子词为更大的单位，以更好地捕捉 DNA 的复杂性。此外，通过将稀有词拆分成子词，BPE 在整个数据集上具有良好的泛化性，并通过合并来自广泛分布模式的意义来帮助模型理解全局上下文。

基于 BPE-600 分词，像 GROVER [?] 这样的模型通过对长为 2-6 个核苷酸的 next-k-mers 预测进行准确性评估，以优化词汇作为基础模型训练的微调任务。然而，分词时仍然存在不平衡的 token 使用、预测准确性差异以及丢失重要 DNA token 的风险等挑战没有解决。

为了解决这些问题，我们提出了结合 k-mer 和 BPE 分词方法的优点，通过合并给定 DNA 序列的 k-mer 和 BPE-600 标记，确保更加平衡且具有上下文意识的表示。它能够捕捉到短期和长期模式，从而更好地理解 DNA 序列对局部序列结构和全局上下文的保留。这项研究工作的主要贡献如下所示：

1. 我们提出了一种混合分词策略，该策略结合了 6-mer 分词和 BPE 方法中的标记。然后，我们制定了一种新方式，将从 600 次 BPE 循环中获得的唯一 6-mer

*All authors have contributed equally. Names are listed in alphabetical order.

标记和最佳 BPE 标记合并，以确保代表两种分词方法的标记的词汇平衡。
2. 我们使用从混合分词方法中获得的最优词汇训练了一个 Foundation DLM 模型。
3. 我们使用下一个 k-mer 预测作为微调任务来评估基础模型的性能，并获得了 3-mer 的准确率为 10.78 %，4-mer 为 10.1 %，5-mer 为 4.12 %，这表明与 SOTA 基础 DLMs 相比性能有所提高。

本文分为五个部分。部分 1 介绍了研究问题和动机，并强调了我们在解决该问题上的贡献。部分 2 讨论背景和相关工作。部分 ?? 描述了我们提出的方法，并提供了详细的系统概述。部分 ?? 讨论了实验设置、数据集、评估指标、性能结果以及不同方法的比较。部分 ?? 讨论了本研究工作中遇到的限制和挑战。部分 3 通过强调贡献和研究结果来总结全文。部分 ?? 讨论了所提出的 DLM 在下游任务中的应用和未来的研究方向。

2. 背景和相关工作

2.1. 分词方法

HyenaDNA [?] 使用一种直接在单核苷酸水平处理 DNA 序列的分词方法，避免了基于频率的聚合分词器的需求。这种方法为短序列和长序列提供了高分辨率分析，使其对检测单核苷酸多态性（SNPs）和理解基因表达中的长程依赖特别有效。该模型的标记集合包括标准的 DNA 碱基（A、G、C、T 和代表未知碱基的 N），以及用于填充、分隔和未知字符的特殊标记，这些标记被映射到一个嵌入空间以进行进一步分析。

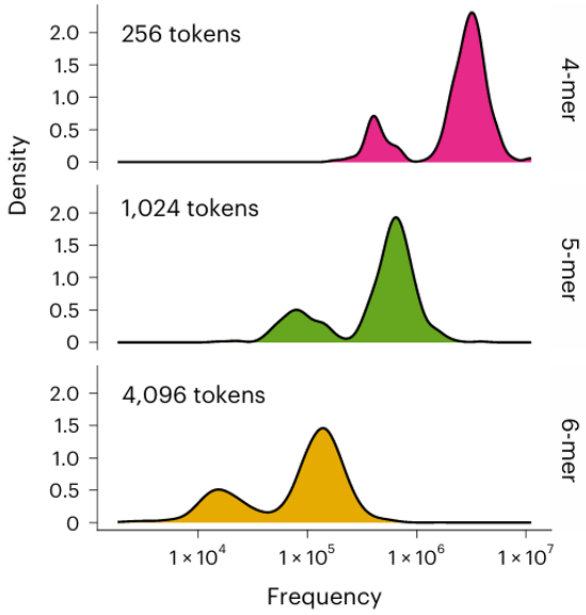

Figure 1. 展示由于 k-mer 标记化导致的不平衡标记分布

DNABERT [?] 使用 k-mer 分词法，其中 DNA 序列被分割成长度为 k 的重叠子串，称为 k-mer，以捕捉更丰富的上下文信息。例如，序列"ATGGCT" 可以被分词成 3-mers 为 ["ATG," "TGG," "GGC," "GCT"] 或分词成 5-mers 为 ["ATGGC," "TGGCT"]。DNABERT 为不同的 k 值（3, 4, 5, 和 6）训练了不同的模型，每个模型都有一个词汇表，包括所有可能的 k-mer 排列以及五个特殊标记：[CLS], [PAD], [UNK], [SEP], 和 [MASK]。K-mer 可以作为重叠和非重叠类型来应用。重叠的 K-mer 旨在通过包含相邻标记之间的共享部分来捕获更多的上下文，但它们往往无法提供比单核苷酸方法更显著的优势，还面临信息泄漏的风险。另一方面，非重叠的 K-mer 减少了标记化序列的长度，使其在计算上更有效。然而，它们可能将有意义的 DNA 模式分割成独立的 K-mer，可能会丢失重要的上下文信息，如 2 和 1 所示。虽然这种方法在核苷酸水平上增强了上下文理解，但重叠的 k-mer 导致词汇量大且标记分布不平衡，如图 1 所示，并且模型难以捕捉更广泛的序列上下文，主要集中于标记级别的模式而非整体基因组结构。此外，由于标记分布不均，模型面临罕见词问题。DNABERT-2 [?] 用 Byte Pair Encoding (BPE) [?] 替代了传统的 k-mer 标记化，这是一种广泛使用的数据压缩和词缀标记化算法，常用于自然语言处理（NLP）。这种方法有效地解决了 k-mer 标记化的局限性，如效率低下和有限的上下文理解，同时保留了非重叠标记化的计算效率，使其成为基因语言建模的强大解决方案。图 2 展示了重叠和非重叠 k-mer 方法的缺点。

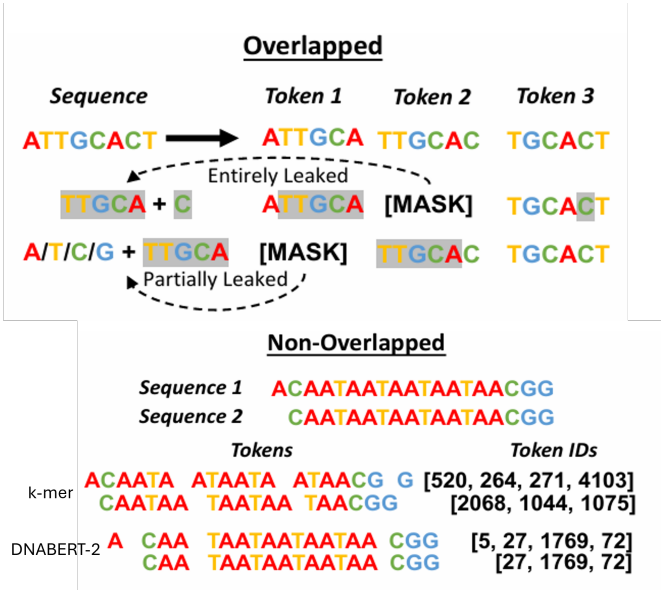

Figure 2. 显示重叠和非重叠 k-mer 分词的缺点。在重叠 k-mer 的情况下，关于被掩盖的标记的信息会被其相邻标记泄露，而在非重叠设置中，添加或删除一个核苷酸碱基会导致标记序列发生剧烈变化。

GROVER [?] 利用 BPE 作为微调任务，通过下一个 k-mer 预测的准确性来获得最优词汇表。使用 BPE 创建频率平衡的词汇表，GROVER 可以捕获既有的词

级特征又有更大序列的上下文。字节对编码（BPE）通过将稀有或不认识的单词拆分为较小、有意义的子词单元来解决稀有词问题，而不是将它们视为单一、不识别的标记。

基于 BERT 的 transformer 架构，其通过掩码标记预测进行训练，并使用具有下一 k-mer 预测的内在验证进行优化，以确保无偏性能。GROVER 在启动子识别、蛋白质-DNA 结合预测和下一 k-mer 预测等任务中表现出色，超过了现有模型，同时提供了对基因组注释、序列方向性和复制时机的见解。然而，它面临着不均匀的标记使用、预测准确性差距以及在标记化过程中丢失必要的 DNA 标记的风险。此外，尽管 GROVER 减少了频率不平衡，但仍存在一些残留效应，并且在某些任务中可能优先考虑标记频率而非更深层次的生物学背景。

2.2. DNA 基础模型

最近在 DNA 建模方面的进展利用基础模型来解开基因组序列的复杂性。DNABERT [?] 引入了一种基于 BERT 的方法，通过注意力机制分析核苷酸之间的关系，而核苷酸 Transformer 通过在 1000G 基因组和 850 个物种的数据上训练大规模模型（高达 25 亿参数）对此进行了扩展。DNABERT-2 [?] 通过引入更高效的字节对编码（BPE）分词器改进了基因组建模，减少了内存和计算需求同时提高了性能。HyenaDNA [?] 挑战了极限，推出了能够以单核苷酸分辨率处理 100 万标记的模型，为基因组学的上下文学习开辟了先河。Caduceus [?] 通过其长距离 Mamba 模块解决了如反向互补性等特点，而 VQDNA [?] 则采用带有向量量化码本的卷积编码器来增强分词。DNAGPT [?] 扩展了传统的 GPT 架构，用于 DNA 序列分类和鸟嘌呤-胞嘧啶含量回归等任务。这些模型中的每一个都引入了独特的策略来解码基因组数据。我们的研究特别关注改进分词方法，以进一步推动该领域的发展。

在之前的研究中，BPE-600 和 6-mer 分词策略在编码基因组序列以训练 DNA 语言模型（DLMs，如 GROVER 和 DNABERT）方面显示出更好的性能。特别是 GROVER 使用下一个 k-mer 预测任务评估了各种词汇配置，以识别基因组序列的最佳分词方法。结果表明，为 600 次迭代（BPE-600）训练一个字节对编码（BPE）模型，可以提高预测下一个 k-mer 的性能，从而增强预训练 DLM 在下游任务中的效能。然而，随着 k 值的增加，使用 BPE-600 训练的 DLMs 的性能开始下降。

为了解决这一限制，我们提出了一种混合分词方法，将 6-mer 和 BPE-600 分词方法的优势结合起来，为 DLM 模型生成一个最佳词汇表。在我们的方法中，首先将 DNA 序列分割成长度为 305 的片段。然后每个片段分别使用 6-mer 和 BPE-600 方法进行分词。结果得到的词汇单元被连接在一起，形成了一种复合表示，利用了两种分词技术的优势。这种合并策略旨在通过结合 6-mer 的细粒度固定长度分割和 BPE-600 的自适应子词建模来提高下一个 k-mer 预测的性能。通过使用 k-mer 预测作为微调任务，我们获得了训练 DLM 模型的最佳词汇表。我们的方法学详细概述在图 3 中。

2.3. BPE-600 分词

这个任务的字节对编码（BPE），如算法 ?? 所示，过程从四个基本核苷酸（A, C, G 和 T）开始，每个核苷酸都被表示为单独的标记。在第一次迭代中，算法识别出序列中最常出现的相邻标记对。例如，如果 “CG” 对是最常见的，它将合并为一个新的标记 “CG”，然后被添加到词汇表中，并在序列中替换每个 “CG” 的出现。这个合并过程以迭代的方式继续，每一个新的对（二元组）被识别并组合成一个标记，随着词汇量的增长，在序列中捕获越来越复杂的模式。随着 BPE 算法的进展，新的标记代表更大和更复杂的子结构，使得模型能够捕获 DNA 序列中的高级依赖性。对于这个任务，我们采用了在图 Fig. 4 中概述的 GROVER 配置，并运行了总共 600 次的 BPE 过程，这被实验证明能达到最佳性能（见图 Fig ?? ）。经过这 600 次迭代，从训练数据集中生成了一个包含 601 个独特标记的词汇表。这些标记随后被一致地应用于 BPE 标记化的测试序列的编码，确保训练和测试数据的表示保持统一和连贯。

2.4. k-mer 标记化

这 k-mer 算法旨在从一组 DNA 序列中提取所有重叠的指定长度的子串。在这种特定情况下，我们专注于使用长度为 $k = 6$ 的子串，称为 6-mers。算法的目的是通过遍历每一个可能的起始位置系统地处理每一个 DNA 序列，从中提取一个 6-mer 子串。给定长度为 n 的 DNA 序列（在此情况下为 $n = 305$ ）和指定的子串长度 $k = 6$ ，算法生成从位置 0 到位置 $n - k$ 的重叠子串。由于 $n = 305$ 和 $k = 6$ ，这意味着对于每个 DNA 序列的片段，算法将生成总计 $n - k + 1 = 300$ 个重叠的 6-mer 标记。换句话说，每个 DNA 片段被分解为 300 个不同的 6-mers，确保序列子串的全面覆盖。为了创建一个强大的训练词汇，我们在人体基因组数据集中使用 6-mer 标记化方法，特别选择了 GROVER 推荐的同一数据集中的片段。该基因组数据集提供了一个大的、有代表性的各种 DNA 序列中重叠 6-mers 的来源，使我们能够构建一个全面的 6-mers 词汇。我们选择了 6-mers，因为它在 DNABERT 中表现更好 [?] 。这些标记然后在随后的建模和标记化过程中起到了关键作用，确保我们的方法能有效捕获人体基因组序列中的必要模式和依赖关系。

2.5. 混合分词（BPE+6-mer）

我们的研究采用了两种分词策略——BPE 和 6-mer ——来预处理 DNA 序列以进行不同的建模任务。最初，在训练过程中，为每种方法分别生成了词汇列表。这些列表被合并后，仅保留了独特的标记，生成了一个包含 4461 个独特标记的合并词汇表，其中包括五个特殊标记。混合分词方法使用的独特标记如下：[CLS] 标记序列的起始，[SEP] 用于分隔 DNA 序列段，[MASK] 用于掩码语言建模，[PAD] 填充较短的序列，[UNK] 表

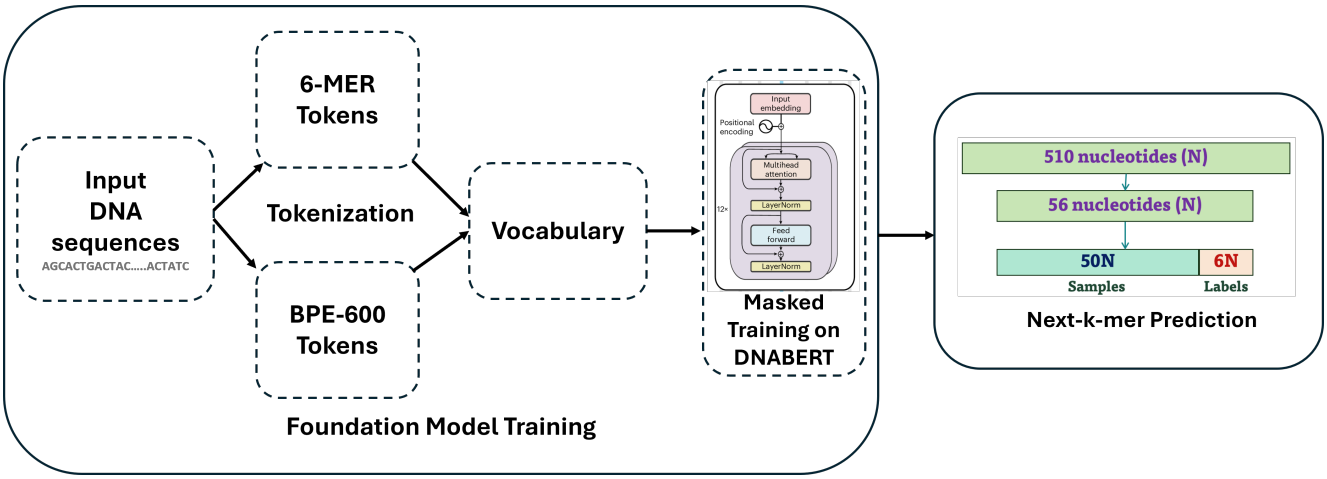

Figure 3. 拟议方法概述

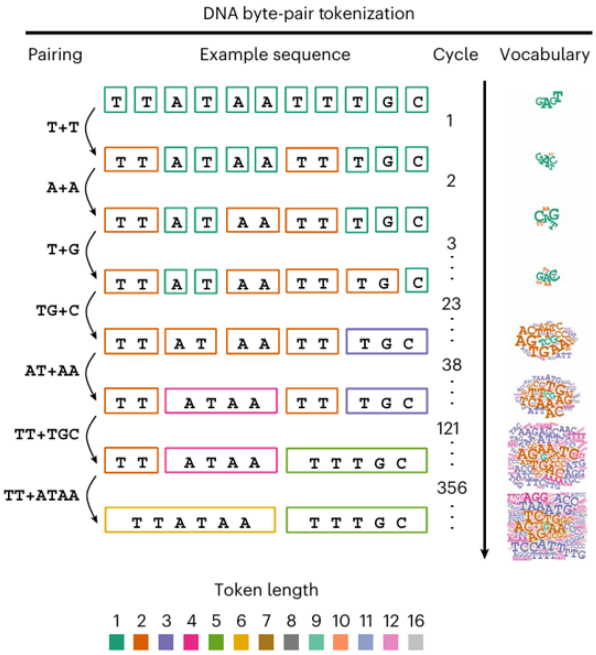

Figure 4. BPE 方法在一个样本序列中进行了突出显示，展示了与该序列相关的分词步骤。生成的词汇按标记长度进行着色，并以词云形式显示，其中单词的相对权重由其频率决定。

示词汇表外的单词。这个词汇表成为了 DNA 序列分词的基础。对于长度为 n 的 DNA 序列，应用 BPE-600 分词产生的标记序列长度严格小于 n，反映了 BPE 算法的压缩能力。在我们的实验中，对于长度 n=305 的序列，BPE 分词的最大观测标记长度为 99。

相比之下，采用 6-mer 分词法，由于其预定义的分段逻辑，对于序列长度为 305 的情况，始终生成固定长度为 300 的输出。当合并 BPE 和 6-mer 分词产生的标记时，生成的序列在输入深度学习模型（DLM）时，最大观察到的标记长度为 399。在这项工作中，我们采用了一种特定的掩蔽策略来训练基础深度学习模型（DLM）。对于初始 300 个标记（对应于 k-mer 标记），我们应用了局部掩蔽方法。对于每个被掩蔽的标记，我们隐藏左边的 3 个标记和右边的 2 个标记，从而防止被掩蔽区域的重叠。该方法确保每个 k-mer 的上下文被充分遮蔽而不引入冗余。对于剩余的标记，我们采用随机掩蔽方案，其中随机选择并掩蔽 15% 的标记。

在这一架构中，我们通过加入一个额外的长短期记忆（LSTM）层来扩展传统的 BERT 模型，以改善序列建模，类似于 GROVER 采用的方法。基础模型是使用 BERT 构建的，该模型包含 12 层 Transformer，每层包含 12 个注意力头，能够通过双向自注意力捕捉丰富的上下文依赖关系，如图中所示。在 BERT 编码器之后，引入一个单独的 LSTM 层以进一步处理标记表示。这个具有 768 个隐藏单元的 LSTM 层允许模型结合序列依赖关系，提供记忆机制以捕捉 Transformer 架构单独可能未充分利用的长期上下文信息。LSTM 层的隐藏大小被设置为 768，以使其与 BERT 的隐藏维度对齐，确保在 Transformer 和 LSTM 组件之间拥有一致的表示空间。通过结合 Transformer 用于全局上下文和 LSTM 用于序列依赖性，该模型在捕捉局部模式和更广泛关系之间达到了平衡，使其成为需要上下文和序列保留任务的理想选择。

为了确保混合标记化的序列处理的一致性，每个 DNA 序列在应用标记化之前首先被分成固定长度为 305 个核苷酸的片段。这一步的预分割处理解决了在分词后可能出现的词元数量变化问题，特别是对于像字节对编码（BPE）这样的方法，它不会产生固定数量的词元。图 ?? 显示，在之前的方法中，选择固定数量词元（例如，12）可能仅导致选择了序列的 k-mer 部分，遗漏了其他重要的词元。为了解决这个问题，我们预先确定序列长度，以一种确保给定序列的所有词元始终包含的方式来计算词元长度。对于 6-mer 分词，分割会

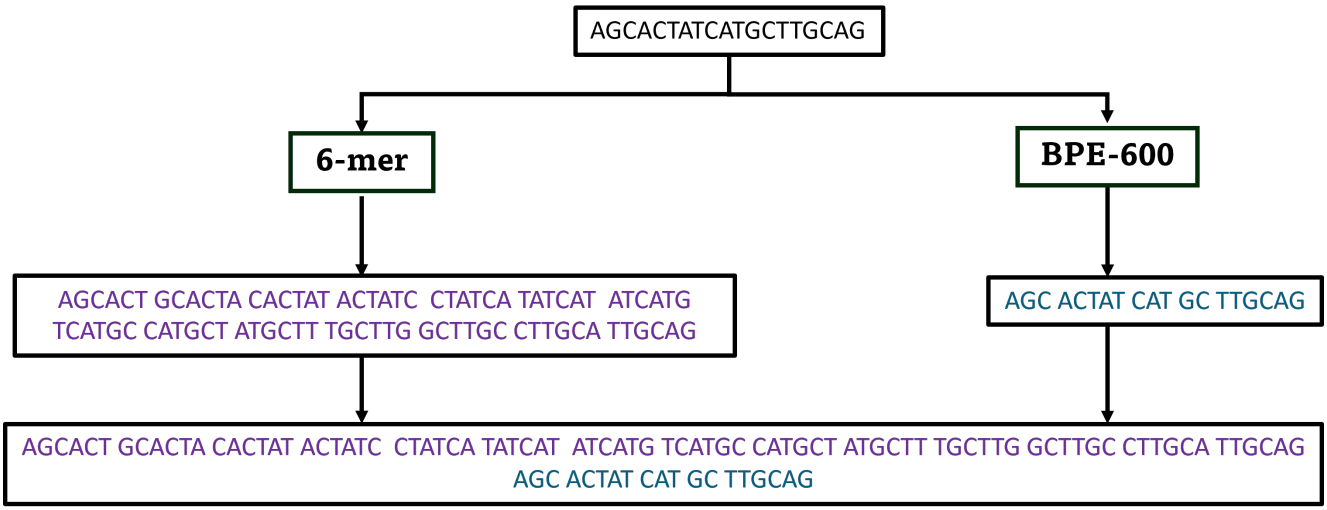

Figure 5. 6-mer BPE 合并概述

在每个段中产生确定数量的词元，计算为 $n-k+1$ ，其中段长度为 $n=305$ 个核苷酸，k-mer 大小为 $k=6$ 。为了在掩码语言建模（MLM）中增强模型学习，我们以 0.15 的概率掩盖词元。当词元被选择用于掩盖（标记为位置 0）时，掩盖将扩展到其周围位置 $-2,-1,0,1,2,3$ ，以捕获 k-mer 的重叠性质。这确保了模型专注于从被选词元周围的上下文中学习稳健的表示，鉴于 6-mer 固有的重叠性。对于 BPE 分词，由于其灵活的子词编码方法，每段生成的词元数量可变，掩盖以 0.15 的概率独立应用于随机选择的词元。这确保了在两种分词方案中 MLM 目标的一致性。通过在分词前先对序列进行分段，保持固定长度的 k-mer 分词，并对重叠的 k-mer 应用上下文感知掩盖，这种混合方法结合了确定性 k-mer 表示的优势和 BPE 的适应性，从而促进了本地和全局序列模式的有效学习。

为了评估我们提出的混合分词方法的效率，我们采用了 Sanabria 等人提出的评估策略，并在 GROVER 框架中进一步优化，该策略通过精细调整任务来识别最有效的基础模型。此方法专门调整为优化 k-mer 和 BPE 分词方法的组合，以训练基础模型。在这一评估方法中，我们首先从每个给定序列中提取前 56 个核苷酸，以符合 GROVER 中使用的方法。然而，代替仅通过一种方法对序列进行分词，我们应用了在第 2.5 节中描述的混合分词方法，结合了 k-mer 和 BPE 分词。序列中剩余的六个核苷酸随后用于生成下一个 k-mer 类别标签。例如，在预测 2-mer 时，位置 51 和 52 的核苷酸作为类别标签，其目标是从 $4k$ 个可能标签的池中预测下一个 k-mer 标签。在分词过程中，我们分别使用 k-mer 和 BPE 方法对输入序列进行独立分词，然后将所得的标记串联以形成统一的标记化表示。在训练期间，我们保持固定的序列长度为 80 个标记。这一长度的选择基于分词方法的特征：对于 6-mer 分词（$k=6$ ），表示 50 个核苷酸的序列由于 k-mer 的重叠结构，所需的标记不超过 45 个。同样，我们的实验观察表明，对于这种长度的序列，BPE 分词不会超过 30 个标记。通过设置 80 个标记的限制，我们提供了一个足够的缓冲区，以容纳这两种分词方法而不截断有意义的信息，同时确保所有序列具有一致的输入长度。这个设置使我们能够有效地评估和比较不同分词组合的性能，同时保持基础模型训练中输入数据的完整性。

我们使用 NVIDIA A100 80GB PCIe GPU 进行了实验。训练设置利用了配置（用于基础和下一个 k-mer 预测模型训练），学习率为 4e-4，Adam 优化器参数设置为 epsilon=1e-6，beta1=0.9 和 beta2=0.98，并使用 0.01 的权重衰减来防止过拟合。模型最多训练 20,000 个 epoch，采用梯度累积策略，每次累积 25 步，以实现每个设备 16 的批量大小进行训练和 32 的批量大小进行评估的高效训练。每 2,500 步保存检查点，总共保留最多 20 个，在训练结束时加载最佳模型。此外，包括 1,000 个预热步骤以稳定训练，同时评估和日志记录分别在 2,500 和 500 步的间隔进行。

2.6. 数据集

我们使用了人类基因组组装 GRCh37（hg19）[?] 。我们仅考虑包含核苷酸 A、C、G 和 T 的序列。我们使用了 40 万条序列进行训练和 10 万条序列进行测试。

2.7. 基础模型训练结果

图 6 显示了基础模型训练结果，突出了训练和评估损失指标。最佳评估损失在步骤 12,500（周期 1.8436）处达到 0.1072，而最后的评估损失在步骤 20,000（周期 2.9497）处为 0.2241。最佳训练损失记录在步骤 11,500（周期 1.6961）为 0.1990，最后的训练损失在步骤 20,000（周期 2.9497）为 0.3883。

2.8. 微调下一 **kmer** 预测

我们使用预训练的基础 DNA 模型并进行微调，以预测下一个 k-mer，其中 k 为 3、4、5。

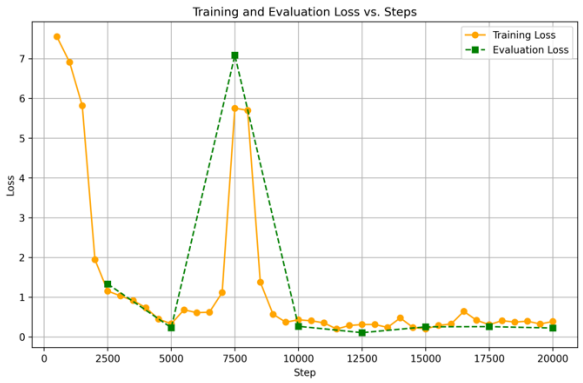

Figure 6. 图表展示了训练和评估损失在训练步骤中的趋势，显示了损失值如何波动并最终稳定下来。

2.8.1 数据准备

在 Ch21 数据集上，我们通过将基因组划分为 510 个核苷酸的重叠序列来进行实验。从每个 510 核苷酸序列中，仅保留前 56 个核苷酸。随机选择了共计 500,000 个序列，其中 80 % 用于训练，20 % 用于测试。对于每个样本，我们使用前 50 个核苷酸作为输入，而标签对应于 4^k 类。在这里，k 表示跟随 50 个输入核苷酸的核苷酸数量，每个类代表这些后续核苷酸的唯一排列。该设置旨在通过预测输入子序列后续的核苷酸模式组合来捕捉基因组序列中的依赖性和模式。3-mer、4-mer 和 5-mer 预测模型的训练和验证性能以评估准确性、评估损失和训练损失的形式显示在图 7 中。

图 8 展示了不同深度学习模型在 3-mer 预测准确性上的比较，突出了显著的性能差异，从 HyenaDNA 的 5 % 到 BPE-6-MER 的 Gr-600 的 10.78 % 。基于 BPE-600 和 6-mer 分词方式的模型（10.78 % ）表现优于其他模型，显示出结合 BPE-600 和 6-mer 分词可以提高准确性。DNA-BERT2 达到了 8 % ，反映出其对 DNA 序列任务的强适应性，而基于 k-mer 的模型（4-mer、5-mer 和 6-mer）表现一般，其中 4-mer 模型略领先，为 7 % 。相比之下，NT-Human 和 HyenaDNA 分别以 6 % 和 5 % 落后，表明这些模型不太适合 3-mer 预测等细粒度任务。

图 ?? 中 4-mer 预测的准确性比较显示，使用 BPE-6-MER 的 Gr-600 以 10.1 % 领先，其次是 Gr-600 (BPE) 达到 7 % 。DNA-BERT2 表现适中，达到 4 % ，而包括 HyenaDNA 和 kk-mer 模型（4-mer、5-mer 和 6-mer）在内的其他模型约为 3 % 。NT-Human 落后为 2 % 。

图 ?? 中 5-mer 预测准确度的比较显示，Gr-600 使用 BPE-6-MER 时达到最高准确度 4.12 % ，其次是 Gr-600 （BPE）为 3.5 % 和 DNA-BERT2 为 2 % 。其他模型，包括 HyenaDNA 和 k-mer 模型（4-mer、5-mer 和 6-mer），表现类似于 1 % ，而 NT-Human 则落后于 0.5 % 。

这些结果强调了在提高预测准确性方面，分词策略和架构的重要性，其中像 Gr-600 与 BPE-6-MER 这样的混合方法提供了最佳性能。

我们在训练基础模型时面临了几个挑战，主要是由于基因组数据的庞大复杂性和规模。高昂的计算成本和时间需求使得很难获得训练所需的 GPU 资源。此外，DNA 序列通常对于标准模型来说过长，当被拆分成较小的片段时导致上下文丢失，而需要将 k-mer 和 BPE 用于同一序列时则导致序列变得更长，进一步复杂化了训练过程。

3. 结论

这项研究本质上深入探讨了传统分词方法的问题，考察了 NT、DNABERT、GROVER 和基于 k-mer 模型的优缺点，并突出说明了应用所提出的混合分词（BPE+K-MER）方法训练基础模型的实验结果。我们通过将 BPE-600 和 6-mer 分词获取的最佳词汇应用于基础模型训练，并通过 3、4 和 5-mer 作为微调任务的后续预测准确性来评估其性能。实验结果表明，在训练基础 DLM 模型时应用混合分词策略可获得最佳后续 k-mer 预测性能。

我们计划在未来的工作中扩展训练和评估，以包括 2-mer 和 6-mer 的分词。随后，我们将探索 k-mer（2 和 6）与 BPE-600 的额外组合，以确定基础模型训练的最佳词汇配置。我们的基础模型将通过多个下游任务进行严格评估，包括启动子识别（Prom-300）、启动子扫描（Prom Scan）、转录因子结合位点预测（CTCF 基序结合）和剪接位点预测，以评估其在多种基因组数据集上的表现。此外，我们将对最新的 DNA 语言模型（DLMs），如 DNABERT2、NT、HyenaDNA 和其他基于 k-mer 的模型进行微调实验。将进行对比分析，以在指定的下游任务中基准化我们的模型与这些现有方法的表现。鉴于由两种标记化技术集成导致的序列长度增加，我们还将研究能够处理长序列的 DLM 上的训练标记化策略。这将解决较大的标记序列带来的挑战，确保在维持模型在复杂基因组数据集上的表现的同时，进行高效的训练和评估。

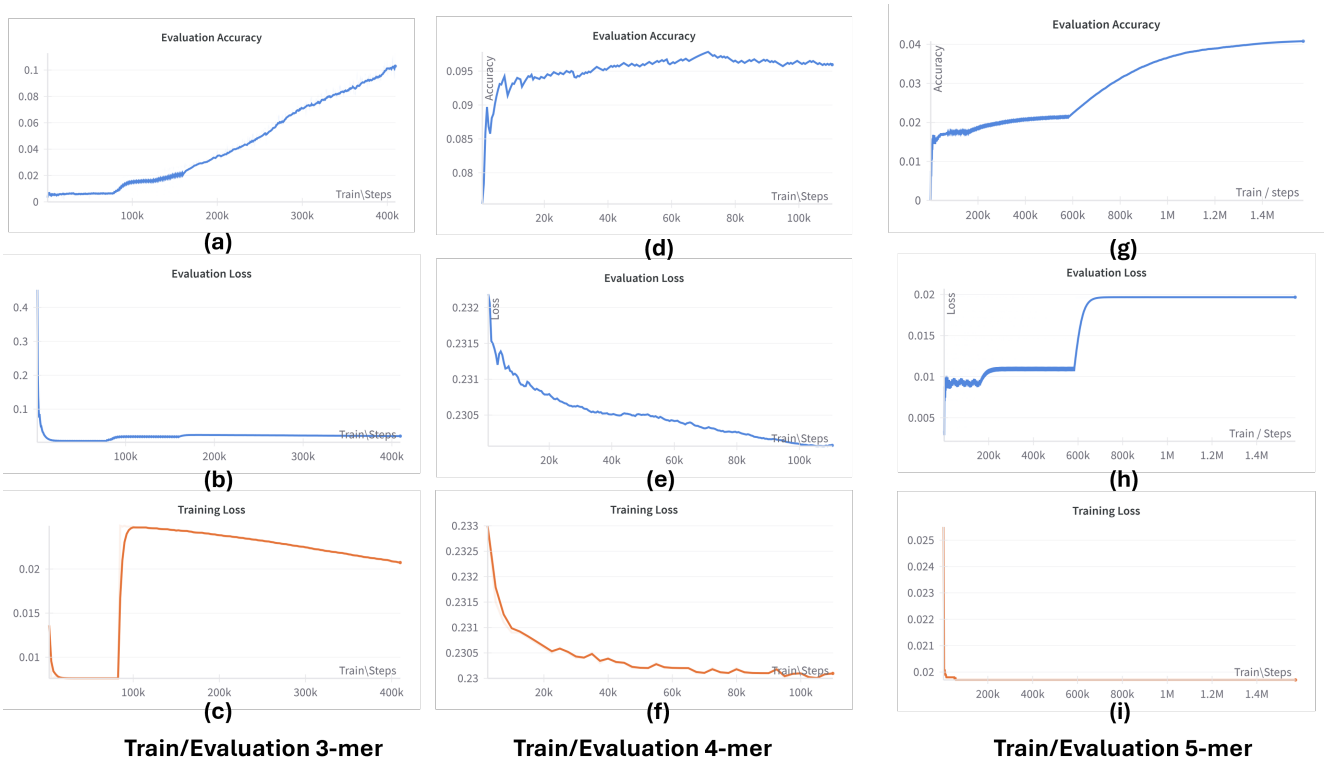

Figure 7. 显示下一个 3、4 和 5-mer 预测的训练和评价性能。图 (a)、(b) 和 (c) 分别表示 3-mer 预测任务的评价精度、评价损失和训练损失。图 (d)、(e) 和 (f) 分别表示 4-mer 预测任务的评价精度、评价损失和训练损失。图 (g)、(h) 和 (i) 分别表示 5-mer 预测任务的评价精度、评价损失和训练损失。

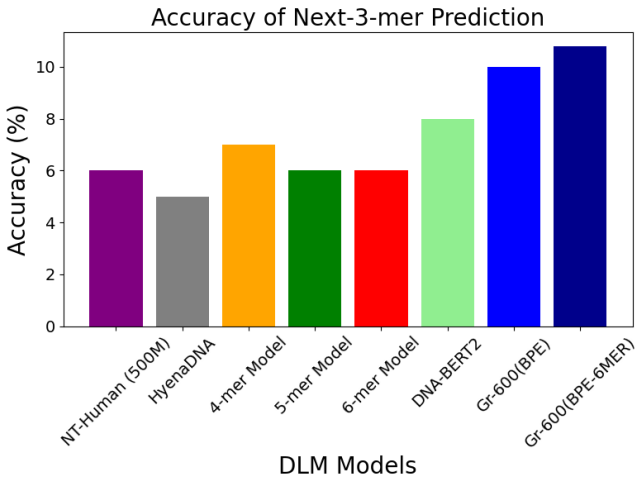

Figure 8. 我们模型的下一个 3-mer 预测准确性与 NT-Human(500H)、HyenaDNA、kmer 模型、DNA-BERT2 Grover(BPE-600) 的比较