***Overview of SeqFormer and its performance***

【卷积局部相关+Transformer全局相关, Bi-LISM高效并行】【嵌入bsalign做对齐】【训练集和过程。分别针对illumina和ONT针对性训练两个模型】

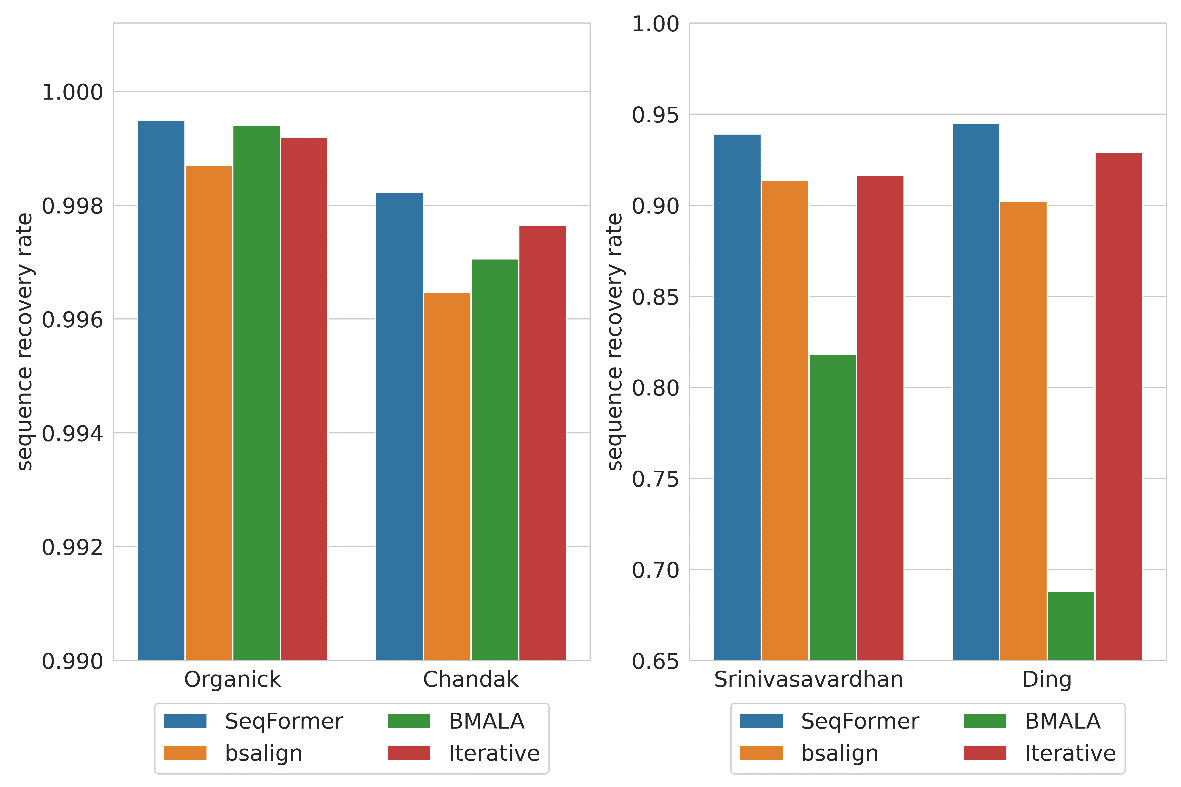
【测试集和效果】【与线性方法的比较优势】

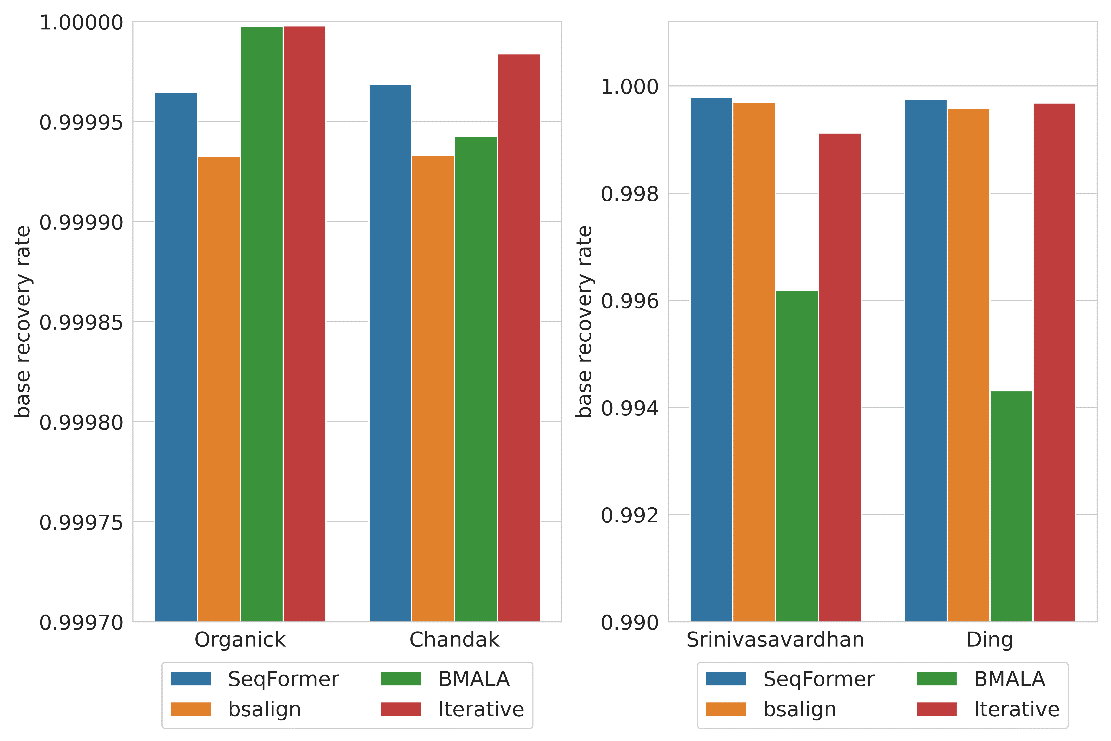
SeqFormer的整体架构是建立在编码器-解码器结构上的，由三个模块组成：特征提取模块，融合卷积及多头注意力的编码模块和Bi-LSTM高效并行的解码模块。融合卷积及多头注意力层能够有效捕捉测序序列中的局部特征及全局依赖关系，Bi-LSTM提高了时序信息的学习和捕捉能力。

表1 数据集概况

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | Organick[14] | Chandak[28] | Srinivasavaradhan[45] | Ding[27] |
| 序列长度 | 150 | 150 | 110 | 260 |
| 序列数量 | 9637 | 11700 | 4146 | 14196 |
| 测序技术 | Ilumina | Ilumina | Nanopore | Nanopore |
| 聚类大小 | 15 | 15 | 20 | 20 |

我们使用序列索引进行聚类，并将聚类结果作为完美簇，使用聚类后的簇序列训练并测试Illumina和Nanopore公开数据集来验证模型的效果，具体数据集信息如表1所示。在实验中，对于每个数据集，训练数据与测试数据的比例设置为4:1。这四个数据集测试的序列重建恢复率结果如图1所示。我们使用SeqFormer与三种传统序列重建方法进行了对比实验。第一种是多序列比对方法bsalign，第二种是按符号计算的多数投票方法BMALA，其中序列的每个位置出现频率最高的符号被视为算法的预测，第三种是迭代算法Iterative，该算法使用动态编程方法来检测在序列中观察到的顺序模式（序列的一小段子序列），然后使用最多的字序列连接作为算法的输出。

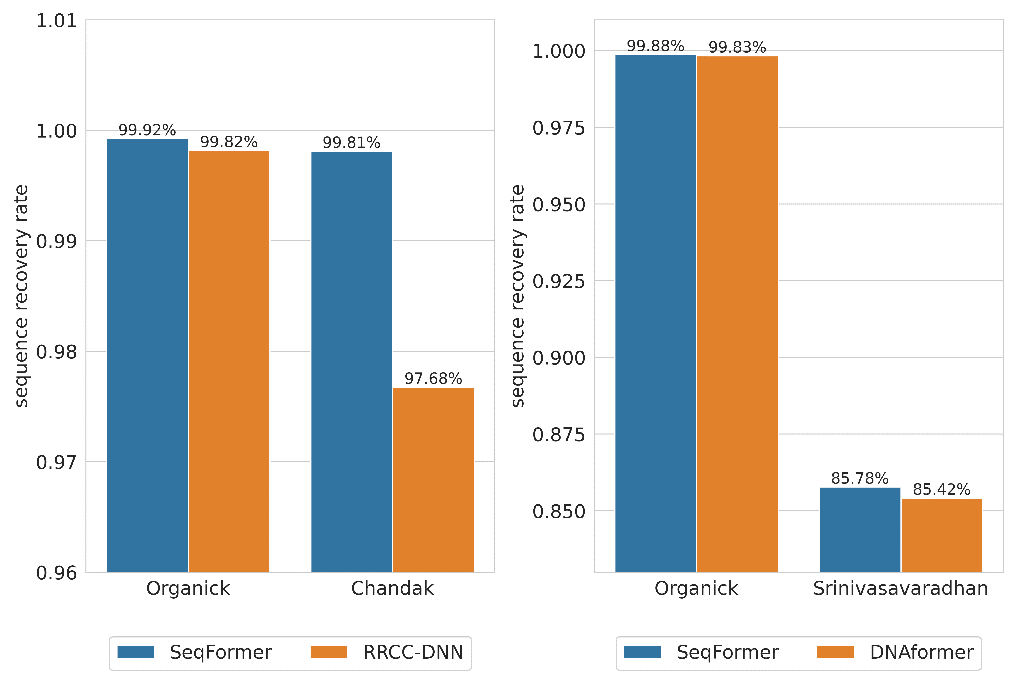




**图1 恢复率对比**

分析SeqFormer在这几个数据集中与其他方法的序列恢复率对比数据可知，SeqFormer的序列恢复率都比其他方法更高，而且在错误率更高的数据集中，重建恢复率提高的效果更明显。这是因为SeqFormer首先使用多序列比对，已经使大多数序列成功对齐，再使用基于深度学习的模型能够再提取其中一些对齐错误的位置，从而达到更高的序列恢复率。分析SeqFormer的碱基恢复率对比数据可知，在错误率更小的数据集里面，SeqFormer与其他方法相差不大，都达到了99.99%以上。但在错误率更大的数据集里面，SeqFormer的碱基恢复率会比其他方法更高，说明SeqFormer具有更强的纠错能力。

相比于基于深度学习的序列重建模型DNAformer和RRCC-DNN，SeqFormer取得了更高的重建精度。如图2展示了SeqFormer与其他方法的序列恢复率对比。RRCC-DNN通过BWA工具对数据集聚类，使用了5-30的聚类大小进行预测，SeqFormer对数据采用相同的处理方法，实验结果如图2左所示。DNAformer使用了不大于16的聚类大小进行预测，SeqFormer在相同的数据处理下，比对结果如图2右所示。从结果可以看出，SeqFormer在多个数据集上的恢复率都高于其他模型，在Organick数据集上几乎达到了接近100% 的准确率。而在Chandak数据集上相比RRCC-DNN模型，恢复率提高了2.13%。尽管在Srinivasavaradhan数据集上的表现有所下降，SeqFormer的准确率仍然高于DNAformer模型。总体而言，无论是在Organick、Chandak，还是Srinivasavaradhan数据集上，SeqFormer均展现出了更强的重建性能，相较于DNAformer和RRCC-DNN，具有更高的恢复率和更好的表现。



**图2 SeqFormer与其他深度学习方法恢复率对比**

【信道量化任务的输入和输出】

为了评估SeqFormer模型得到的碱基置信度的可靠性，以便在解码时利用软信息来增强解码恢复率，我们使用灵敏度及识别率来评估它的可靠程度。灵敏度及识别率定义如下：假如总共有x个碱基，其中e个碱基发生了错误，有个碱基的置信度落在到的区间内，在这个碱基中，发生错误的碱基数量为个，那么灵敏度= ，这个区间的识别率= 。为了确保实验结果的准确性，选择了错误率较高的Ding数据集进行实验，从每个簇中随机选择5条进行评估，实验结果如表2所示。

表2 碱基错误预测

|  |  |  |  |
| --- | --- | --- | --- |
| 碱基可信度 | 错误预测识别率 | 碱基可信度 | 错误预测灵敏度 |
| 0~0.4 | 65.85% | <0.4 | 9.92% |
| 0.4~0.5 | 58.11% | <0.5 | 17.98% |
| 0.5~0.6 | 53.18% | <0.6 | 25.92% |
| 0.6~0.7 | 45.84% | <0.7 | 33.33% |
| 0.7~0.8 | 39.51% | <0.8 | 41.45% |

从数据中可以看出，碱基的可信度越高，模型的错误预测识别率越低，这表明碱基可信度在一定程度上反映了模型预测的可靠性。灵敏度通常用于衡量模型对错误预测的反应能力，即在所有真实的错误预测中，模型能够正确识别出多少。实验结果表明，碱基可信度与错误预测的灵敏度之间存在正相关关系。随着可信度的提高和范围的增大，模型的灵敏度也随之增加，能够识别更多的错误预测碱基。

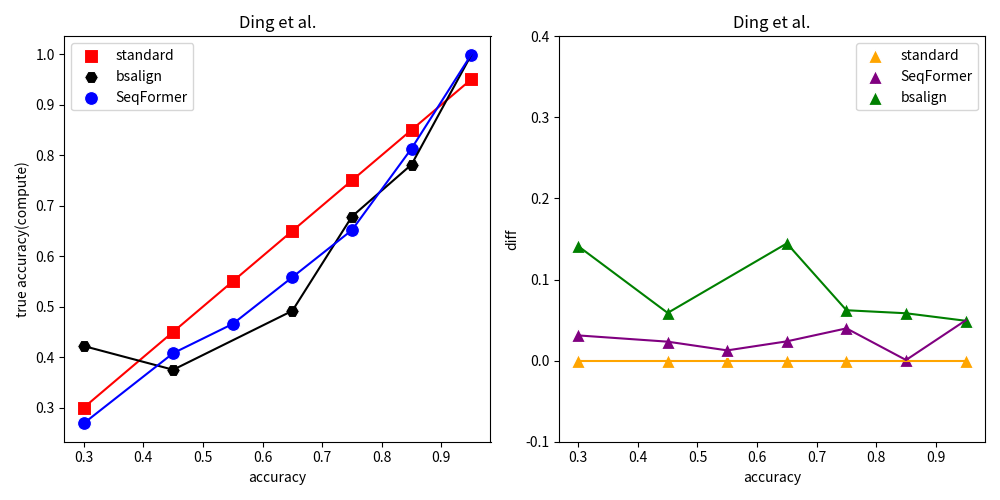
为了评估碱基置信度的可靠性，本文对比了由bsalign的碱基置信度与SeqFormer模型生成的碱基的置信度，这里通过差异值来衡量。首先需要把bsalign生成的碱基质量值转换为与碱基置信度相同的形式，每个碱基的可信度可定义为：



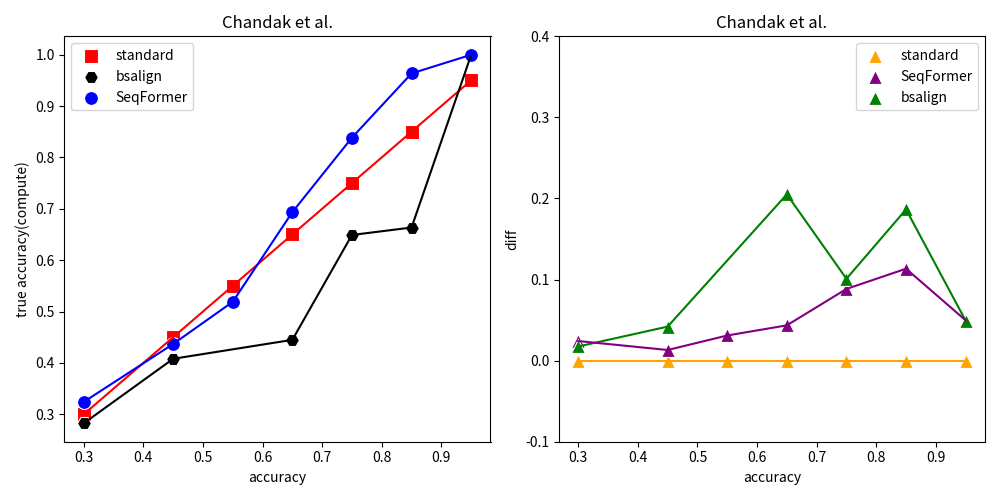
假如有个碱基的置信度落在到的区间内，在这个碱基中，实际发生错误的碱基数量为个，那么该区间的准确度可定义为，这里希望这些碱基的准确度接近于该区间的中间值，故通过以下公式定义：



这个差异值越小说明置信度越准确。



(a) Ding et al.



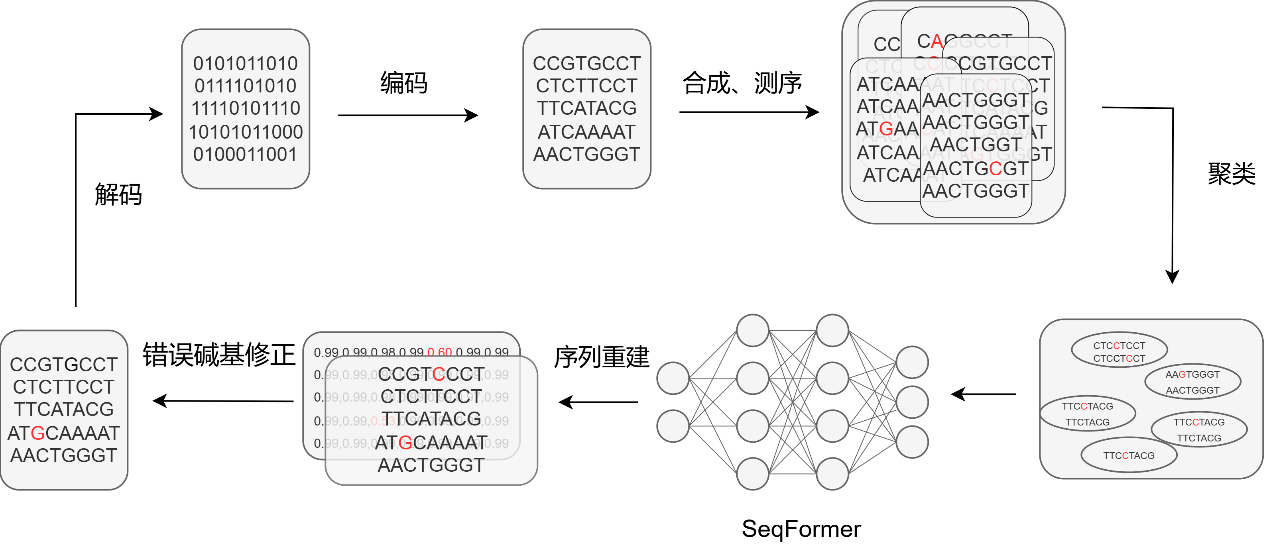
(b) Chandak et al.

**图3 碱基准确度对比**

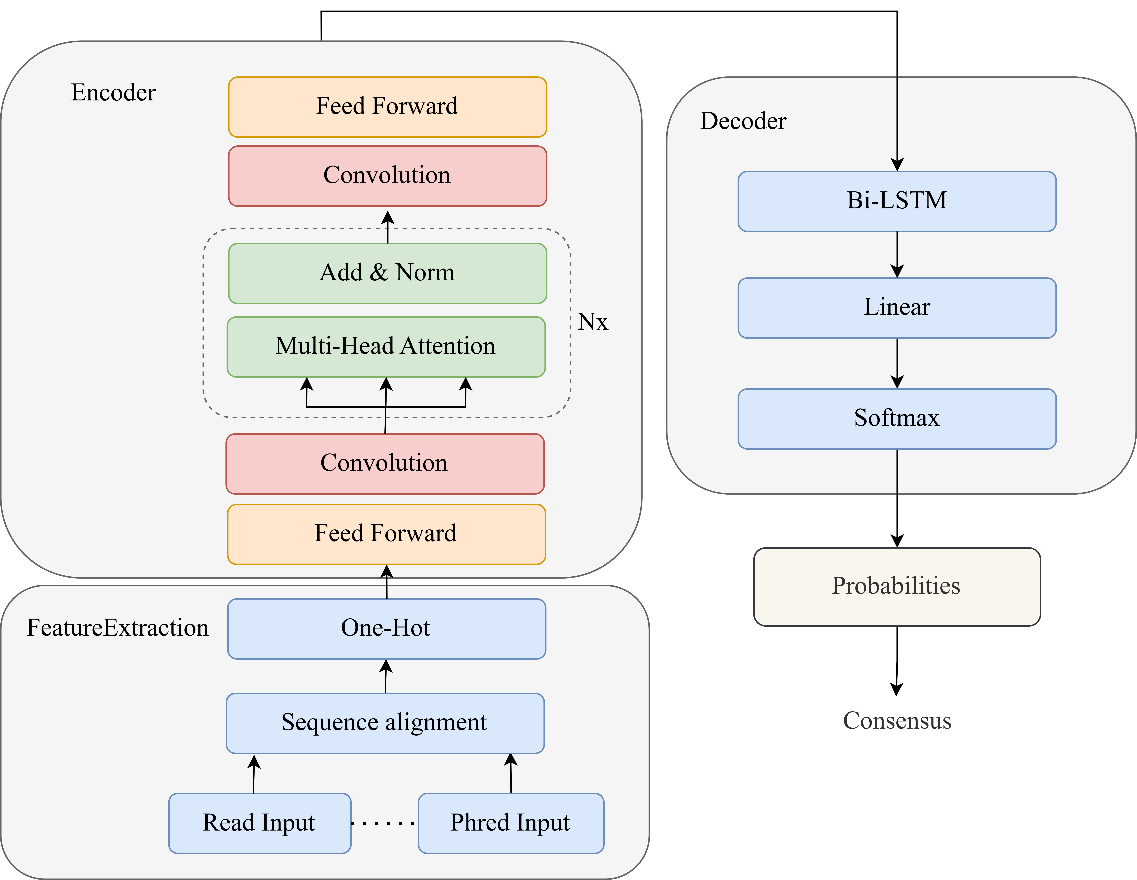
随着置信度的增加，SeqFormer 在大部分情况下表现出较低的差异值，尤其在较高置信度区间（如 0.85 和 0.95）时，差异值较小，表明 SeqFormer 模型的准确度更接近预期的目标准确度。这说明 SeqFormer 在处理更高置信度的碱基时，能够更好地匹配目标准确度。相比之下，bsalign 模型在部分区间（如 0.65 和 0.75）上出现较大差异，尤其是在Ding数据集上，差异值较大。这表明 bsalign 在某些置信度区间的表现较不稳定，模型的预测结果与目标准确度的差异较为明显。总体而言，差异值的分析揭示了 SeqFormer 相较于 bsalign 模型在大多数置信度区间内具有更高的准确性和稳定性。

***Methods***

***SeqFormer architecture***



**图4 SeqFormer重建流程示意图**

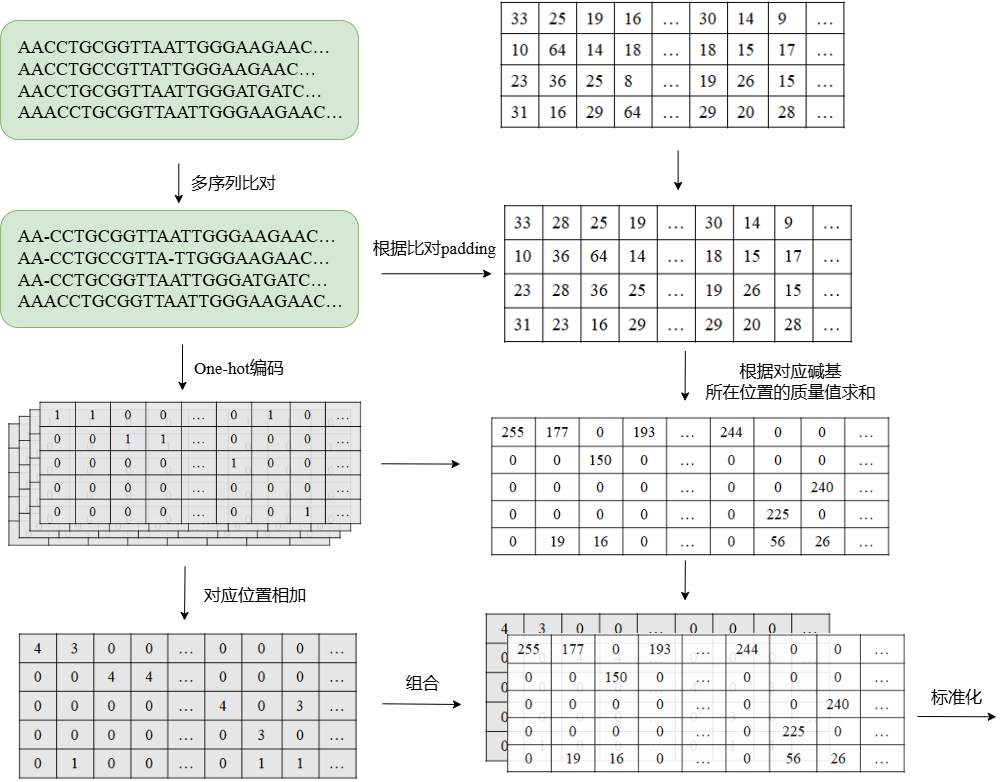


**图5 SeqFormer模型整体架构**

在DNA存储中使用SeqFormer的进行序列重建的完整流程如图4所示。首先需要使用结合了ECC的编码方法把二进制数据文件编码为DNA序列，然后经过合成、测序步骤引入了一些随机错误，得到了多条噪声拷贝序列。一般来说，经过碱基调用测序后得到的fastq文件会包含了序列及质量分数，可用于解码。得到噪声拷贝序列后，需要根据序列的相似度把这些序列进行聚类，本文使用的聚类方法是根据索引聚类。聚类后使用SeqFormer模型对每个类的序列进行重建，得到一致性序列，所有一致性序列都会带有它所对应的碱基置信度，可根据置信度再对序列进行错误碱基修正。最后使用对应的解码方法进行解码即可恢复二进制文件。

如图5所示，SeqFormer由三个模块组成：特征提取模块，融合了卷积层及多头注意力的编码模块和Bi-LSTM解码模块。特征提取模块将测序后得到的FASTQ文件中的序列及其质量值信息经过序列比对并采用独热编码方式输入到模型中。编码器部分结合了卷积层和多头注意力层，能够有效捕捉测序序列中的局部特征及全局依赖关系，解码器部分采用双向LSTM（Bi-LSTM），旨在进一步提高时序信息的学习和捕捉能力，最后通过线性处理转换为原来的维度，采用Softmax函数输出每个碱基的置信度概率分布，根据此概率分布得到一致性序列Consensus。此概率分布及一致性序列可用于解码恢复。

SeqFormer的输入为FASTQ文件中的序列及质量值信息，对于质量值信息，将每个ASCII码值减去一个偏移量得到对应的整数。Sanger格式的偏移量通常是33（因为最小的phred质量分数0对应ASCII码33）。因此，转换公式为：质量值= ASCII码值-33。



**图6 SeqFormer特征提取模块**

特征提取模块会将测序后得到的FASTQ文件中的序列及其处理后的质量值信息经过序列比对并采用独热编码方式输入到模型中，具体流程如图6所示。首先根据index对FASTQ文件进行聚类后，得到多组聚类序列。由于每组聚类序列包含数量不固定且长度各异的DNA序列，因此采用bsalign工具进行多序列比对。比对后的序列包含五种字符：A、C、G、T以及用于填充对齐的“-”。接下来，通过独热编码（也称One-Hot编码）技术把比对后的序列转换为高维向量表示，得到一个的矩阵，其中L表示序列的长度。独热编码后每个符号都能以稀疏向量的形式输入到神经网络中。同时，质量值作为反映每个测序序列碱基可靠程度的重要指标，也被视为特征数据与DNA序列一同输入。为了确保所有序列的质量值长度一致，根据序列比对结果进行填充（padding）。其中“-”位置的值是其周围n个质量值的平均数。随后，每个质量值根据其对应的符号进行独热编码，并将其质量值作为权重，赋给独热编码矩阵中的相应位置。最终，这样的操作也会得到一个质量值加权的矩阵。最后，将处理后的序列特征矩阵和质量值特征矩阵组合在一起，并进行标准化处理，以确保数据的稳定性和一致性。经过这些预处理步骤后，数据被输入到SeqFormer模型中。

编码模块结合了卷积层和多头自注意力层，卷积层主要用于学习由于删除或插入错误造成的基于相对偏移的局部交互特征，而自注意力层则专注于捕捉序列内部的全局依赖关系。自注意力机制能够动态聚焦于不同的位置信息，从而更精准地模拟序列中的长期依赖特性。如图5的Encoder部分所示，编码模块的最外层分别在首部和尾部采用了前馈神经网络（FFN），首部用于扩展特征维度，尾部则恢复原始特征维度。这两个前馈网络均为线性层，内部维度为1024。编码模块的卷积层使用的核大小为5、padding大小为2、步长为1。中间还包含了 N=2 层的多头自注意力机制，两层之间使用了残差连接和层归一化，以提升深度模型的稳定性和训练效果。

多头注意力模块（MHA）由缩放点积计算，其中



其中是特征的线性变换。我们采用了个并行的注意力头，每个注意力头隐藏维度为512。令这h个缩放点击注意力结果的串联得到



其中，。前馈神经网络对输入的每个位置都会执行两个线性变换操作，然后通过ReLU激活函数实现非线性变换，这个过程表示如下：



全连接前馈网络作用是在保留输入序列中位置信息的基础上，对输入的每个位置进行非线性变换，以便更好地提取特征和增强模型的表达能力。

解码模块使用了双向LSTM（Bi-LSTM），它能够在时序维度上补充编码器可能遗漏的序列信息，增强模型的时序特征学习能力。其中层数为2，隐藏层维度为64。接着，使用全连接层将双向LSTM 的输出映射到解码器的输出维度64。最后通过线性变化恢复原始特征维度，得到一个的向量，采用Softmax函数输出每个碱基的置信度概率分布，根据此概率分布得到一致性序列Consensus。此概率分布及一致性序列可用于解码。

***Training and evaluation***

模型训练使用了8个GPU核进行并行计算，共进行30个epoch的训练。每个GPU核的批处理大小为64。神经网络的初始学习率设置为1e-3，采用Adam优化器进行优化，该优化器能够自适应地调整每个参数的学习率。整个训练过程在PyTorch 2.1和Python 3.10环境下实现。模型的时间复杂度为:



其中L代表序列长度，N为层数，D为自注意力机制隐藏单元数，Df为前馈神经网络隐藏层大小，Ks为卷积核大小，H为BiLSTM单元的隐藏层大小。

***loss and optimization***

SeqFormer模型采用了基于交叉熵的加权组合损失函数，用于优化每条序列的预测结果。假设有一个样本序列，其原始标签为，模型输出的概率分布为，则交叉熵损失函数可以表示为：



其中是序列的长度，是第个位置的真实标签，是模型预测的第个位置的碱基概率分布。为了增强模型对序列重建的能力，损失函数引入了原序列与重建序列的编辑距离，并将其乘以一个权重系数，作为附加的正则化项。引入编辑距离损失可以鼓励模型生成与目标序列在字符级别上更加相似的输出。



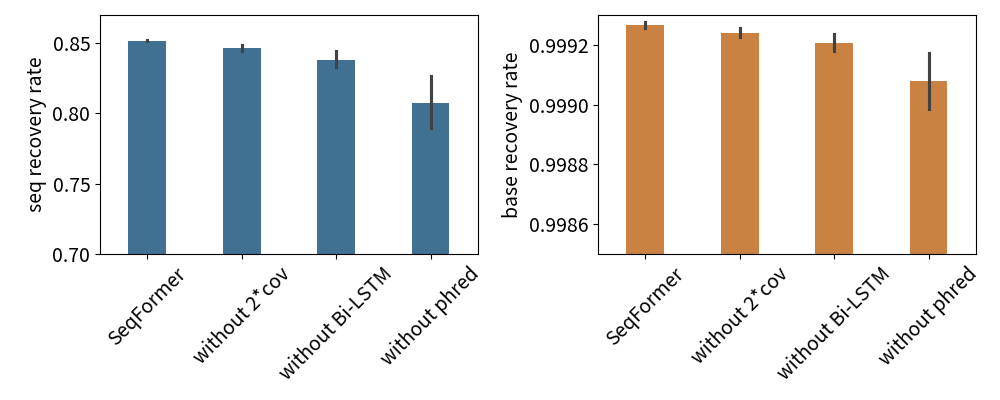
其中，分别表示真实序列和模型预测的重建序列。最终，SeqFormer 模型的总损失函数是交叉熵损失和编辑距离损失的加权和，具体形式如下：



***Ablation study***

表3 不同模型组件恢复率对比

|  |  |  |
| --- | --- | --- |
| 模型组件 | 平均序列恢复率 | 平均碱基恢复率 |
| SeqFormer | **85.238%** | **99.927%** |
| 去除Bi-LSTM | 83.825% | 99.921% |
| 去除2\*cov | 84.643% | 99.924% |
| 去除序列质量值 | 80.775% | 99.908% |



**图7 不同模型组件恢复率对比**

为了评估SeqFormer模型中不同组件对序列重建性能的具体影响，我们进行了消融实验，使用Ding数据集进行训练和测试，其中训练数据包含10000条序列，测试数据包含8000条序列，且每条参考序列的噪声拷贝数量为10条。表3展示了在不同模型组件配置下的平均序列恢复率及平均碱基恢复率。如图7是它的柱形展示图。

首先，我们对比了SeqFormer模型在解码部分是否包含Bi-LSTM解码器时的效果。实验结果表明，添加Bi-LSTM解码器有效弥补了基于深度学习的编码器在识别序列长程时序依赖关系上的不足，从而提高了DNA序列的恢复准确性。当去除Bi-LSTM解码器后，模型的序列恢复率下降了约1.4%。这是因为虽然编码器在序列碱基恢复中起到了基础性作用，但对于那些编辑距离错误较小的序列，Bi-LSTM解码器能够进一步捕捉复杂的时序模式，从而准确恢复整条序列。去除Transformer编码器中的卷积层后，模型的恢复率下降了0.6%。这主要是由于模型在序列局部错误识别上的性能受损。值得注意的是，去除卷积层后的效果与去除Bi-LSTM解码器的情况相似，原因在于它们都能够捕捉到那些剩余错误较少的序列特征。此外，还进行了仅使用序列而不考虑序列质量值的消融实验，去除质量后，模型序列恢复率下降了约4.5%，说明质量值对SeqFormer模型的序列恢复性能起着至关重要的作用。质量值能够很好地评估每个碱基的可靠程度，模型可以依据当前位置每个碱基的具体值及其可靠程度来预测真实值。

通过一系列消融实验，进一步验证了SeqFormer各关键组件对性能的贡献。实验结果表明，卷积层和双向LSTM解码器对序列恢复至关重要，而质量值则在提升模型性能上发挥了显著作用。去除这些模块后，模型的性能均出现了明显下降。