(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 119785882 B
(45)授权公告日 2025.10.24

(21)申请号 202411978849.6

(22)申请日 2024.12.31

(65)同一申请的已公布的文献号
申请公布号 CN 119785882 A

(43)申请公布日 2025.04.08

(73)专利权人 合肥综合性国家科学中心大健康研究院
地址 230601 安徽省合肥市经济技术开发区宿松路4090号1号楼202

(72)发明人 吴增丁 费才溢 徐实

(74)专利代理机构 上海一平知识产权代理有限公司 31266
专利代理师 徐迅 崔佳佳

(51)Int.Cl.
G16B 20/30(2019.01)
G16B 40/00(2019.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)
G06N 3/045(2023.01)
G06N 3/048(2023.01)

(56)对比文件
WO 2024105409 A1,2024.05.23
CN 112382338 A,2021.02.19

审查员 郝永利

权利要求书2页 说明书15页 附图9页

(54)发明名称

基于深度学习的mRNA翻译蛋白质产物的预测方法与模型

(57)摘要

本发明开发了一种基于深度学习的mRNA翻译蛋白质产物的预测方法与模型。具体地,本发明设计了一种基于深度学习方法的转录组水平翻译产物精确预测的方法,用于对转录本序列的翻译位点和翻译产物进行预测,形成TRANSAID模型。本发明的TRANSAID模型采用了独特的序列编码策略和深度学习架构,能够同时捕捉局部序列特征和长距离依赖关系,大大提高了预测准确性;并能够对不同物种的转录本序列进行预测,具有本领域中首次出现的跨物种预测能力;还能够识别5'UTR和3'UTR区域对翻译调控的不同影响。

CN 119785882 B

1.一种构建转录本翻译蛋白质产物的预测模型的方法,其特征在于,所述方法包括步骤:

(S1)提供用于模型训练的数据集,对所述数据集进行数值化转化;

其中,所述数据集为转录本序列,所述转录本序列包括mRNA序列和非编码RNA序列;所述数值化为基于位置的嵌入;

(S2)将(S1)中所述的数据集用于模型架构的训练;

其中,所述模型架构包括基于多尺度卷积模块、残差网络模块和注意力机制模块的模型架构;

所述多尺度卷积模块卷积核对所述基于位置的嵌入进行卷积,获得卷积结果,卷积结果经过批标准化(batch normalization)和线性整流(rectified linear unit,ReLU)激活,输出到下游残差网络模块;

所述残差网络模块包括两个部分:

(Z1)残差网络模块第一部分:使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26的卷积核对上游输出进行卷积,所述卷积核的数量为64,获得卷积结果;对所述卷积结果进行1×1卷积,获得高阶卷积结果;对所述高阶卷积结果进行批标准化和线性整流激活,输出到下游残差网络模块第二部分;

(Z2)残差网络模块第二部分:使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26卷积核对上游输出进行卷积,所述卷积核的数量为128,经过批标准化和线性整流激活后,获得中间特征表示,将所述中间特征表示输出到下游注意力机制模块;

所述注意力机制模块包含3个1×1卷积,分别用于生成查询张量、键张量和值张量;根据所述查询张量、键张量和值张量生成聚焦后的序列表示;合并所述聚焦后的序列表示与上游输出,获得注意力机制模块的输出;

(S3)对(S2)中所述注意力机制模块的输出进行解码,得到预测结果;

(S4)当训练结果达到预定终止条件,则终止模型训练,并获得所述的转录本翻译蛋白质产物的预测模型,即TRANSAID模型。

2.如权利要求1所述的方法,其特征在于,所述基于位置的嵌入包括步骤:

(a1)对所述数据集中每个所述转录本序列中的每个核苷酸进行编码,获得编码的转录本序列;

(a2)对输出标签进行编码;

(a3)对所述编码的转录本序列进行嵌入,获得转录本序列嵌入;

(a4)对所述转录本序列嵌入进行位置编码;

其中,在(a1)中,对所述输入序列中的每个核苷酸进行base编码;在(a2)中,对所述输出标签进行one-hot编码。

3.如权利要求1所述的方法,其特征在于,所述多尺度卷积模块卷积核对所述基于位置的嵌入进行卷积包括步骤:

(B1)使用3种大小的卷积核对所述基于位置的嵌入进行卷积,所述卷积核大小分别为3、5、7,每种所述卷积核数量为32,卷积方式为“same”,获得卷积结果;

(B2)对所述卷积结果进行拼接,获得特征维度为96的卷积结果。

4.如权利要求1所述的方法,其特征在于,所述残差网络模块具体包括步骤:

(C1)使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26的卷积核对所述多尺度卷积模块的输出结果进行卷积,所述卷积核的数量为64,获得卷积结果;

(C2)对所述卷积结果进行1×1卷积,将所述卷积结果的特征维度从64变为128,获得高维卷积结果;

(C3)对所述高维卷积结果进行批标准化和线性整流激活;

(C4)使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26卷积核对(C3)的输出进行卷积,所述卷积核的数量为128;

(C5)对(C4)的输出进行批标准化和线性整流激活,获得中间特征表示。

5.如权利要求1所述的方法,其特征在于,所述注意力机制模块具体包括步骤:

(D1)利用3个1×1的卷积对所述中间特征表示进行卷积,将所述中间特征表示配置为长度为L、维度为C的张量,将所述张量转换为L×C/8的查询张量和键张量,以及L×C的值张量;其中,将所述查询和所述键的卷积核数量配置为所述中间特征表示的输入维度的1/8;将所述值的卷积核数量配置为与输入维度一致;

(D2)将所述查询张量与所述键张量的转置相乘,通过激活函数得到L×L的注意力矩阵,将所述注意力矩阵与所述值张量相乘,得到聚焦后的序列表示;

(D3)将所述聚焦后的序列表示与所述中间特征表示相加,乘以标量权重;

(D4)获得注意力机制模块的输出。

6.如权利要求1所述的方法,其特征在于,所述解码包括步骤:

(E1)对所述注意力机制模块的输出进行1×1卷积,将所述注意力机制模块的输出的特征维度从128降至64,并通过线性整流激活;

(E3)通过Dropout层,以0.2的概率随机遮蔽特征;

(E4)通过1×1卷积将所述特征维度从64降至3,并通过Permute函数将维度调整为(batch_size,seq_len,3),得到输出;

(E5)将所述输出与真实标签的交叉熵作为训练的损失函数;

(E6)通过Softmax函数提供的最大概率值决定转录本序列每个位置的输出标签。

7.一种转录本翻译蛋白质产物的预测系统,其特征在于,所述系统包括:

输入单元,所述输入单元被配置为输入数据,所述输入数据包括待预测的输入序列;

转录本翻译蛋白质产物的预测单元,所述预测单元被配置为执行一转录本翻译蛋白质产物的预测模型,获得所述待预测的输入序列的位置标签,从而获得所述待预测的输入序列的翻译位点和翻译产物;其中,所述预测模型为用权利要求1所述的方法构建的;

输出单元,所述输出单元被配置为输出所述转录本翻译蛋白质产物的预测单元的预测结果。

8.一种权利要求7所述的转录本翻译蛋白质产物的预测系统的用途,其特征在于,用于研究转录本序列中UTR区域和密码子完整性对结构标签的影响。

基于深度学习的mRNA翻译蛋白质产物的预测方法与模型

技术领域

[0001] 本发明属于生物信息学、机器学习及翻译位点预测领域,具体地涉及基于深度学习的mRNA翻译蛋白质产物的预测方法与模型。

背景技术

[0002] 在现代癌症研究与治疗中,准确识别和预测翻译产物具有非常重要的作用。癌症细胞中经常出现异常的翻译事件可能会产生独特的肿瘤特异性蛋白。这些蛋白是潜在的新抗原来源,对于肿瘤免疫治疗具有重要价值,是开发个性化免疫治疗方案的重要靶点。传统上,研究人员主要依赖质谱技术来鉴定和验证蛋白质翻译产物,但这种方法存在诸多局限性:首先,质谱分析需要大量样本且对样本质量要求严格;其次,实验操作复杂,耗时长且成本高;最重要的是,某些低丰度或瞬时表达的蛋白质往往难以被检测到。此外,传统的基因组注释方法也往往无法准确预测这些非典型翻译事件。以上局限限制了人们对肿瘤特异性抗原的发现和利用。

[0003] 翻译位点预测技术的发展经历了多个重要阶段。最早期的方法主要基于简单的序列特征识别,如搜索标准起始密码子(ATG)和终止密码子(TAG/TAA/TGA)。这类方法的代表是NCBI的ORFfinder和Swiss Institute of Bioinformatics开发的Expasy工具。这些工具虽然操作简单直观,但往往产生大量假阳性预测,因为它们无法区分真实的翻译位点和随机出现的类似序列模式。

[0004] 随着研究的深入,研究人员发现翻译起始和终止过程受到复杂的调控机制影响。2015年,Georgia Institute of Technology开发的GeneMarkS-T算法代表了这一领域的重要进展。该方法使用隐马尔可夫模型捕捉序列模式,并通过非监督学习实现了跨物种的预测能力。然而,这类统计学方法仍然难以处理长程依赖关系,也无法很好地识别非经典翻译事件。

[0005] 近年来,高通量测序技术的发展,特别是核糖体测序(Ribo-seq)技术的出现,为翻译位点预测提供了新的研究视角。这些实验方法能够在全转录组水平捕捉翻译起始位点和终止位点的信息。然而,这些实验方法成本高昂,且往往受限于特定的细胞类型和生理条件。

[0006] 在这一背景下,深度学习方法的引入为翻译位点预测带来了革命性的突破。2023年,Fan等人开发的TranslationAI模型展示了深度学习在这一领域的潜力。该模型使用扩张卷积网络处理RNA序列,实现了高精度的翻译位点预测。

[0007] 因此,本领域迫切需要开发一种基于深度学习方法,利用转录组数据准确预测翻译位点的方法,用于准确识别肿瘤中异常翻译产物等。

发明内容

[0008] 本发明的目的是提供一种根据转录组预测翻译位点和翻译产物的模型,该模型基于深度学习的方法利用转录本序列精确预测序列中的翻译位点,进而精确预测翻译产物,

可用于多种真核生物的转录组水平翻译产物的精确预测。

[0009] 本发明的第一方面,提供了一种构建转录本翻译蛋白质产物的预测模型的方法,所述方法包括步骤:

[0010] (S1)提供用于模型训练的数据集,对所述数据集进行数值化转化;

[0011] 其中,所述数据集为转录本序列,所述转录本序列包括mRNA序列和非编码RNA序列;所述数值化为基于位置的嵌入;

[0012] (S2)将(S1)中所述的数据集用于模型架构的训练;

[0013] 其中,所述模型架构包括基于多尺度卷积模块、残差网络模块和注意力机制模块的模型架构;

[0014] 所述多尺度卷积模块卷积核对所述基于位置的嵌入进行卷积,获得卷积结果,卷积结果经过批标准化(batch normalization)和线性整流(rectified linear unit,ReLU)激活,输出到下游残差网络模块;

[0015] 所述残差网络模块包括两个部分:

[0016] (Z1)残差网络模块第一部分:使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26的卷积核对上游输出进行卷积,所述卷积核的数量为64,获得卷积结果;对所述卷积结果进行1×1卷积,获得高阶卷积结果;对所述高阶卷积结果进行批标准化和线性整流激活,输出到下游残差网络模块第二部分;

[0017] (Z2)残差网络模块第二部分:使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26卷积核对上游输出进行卷积,所述卷积核的数量为128,经过批标准化和线性整流激活后,获得中间特征表示,将所述中间特征表示输出到下游注意力机制模块;

[0018] 所述注意力机制模块包含3个1×1卷积,分别用于生成查询/键/值(Query/Key/Value),计算注意力权重,获得注意力机制模块的输出;

[0019] (S3)对(S2)中所述注意力机制模块的输出进行解码,得到预测结果;

[0020] (S4)当训练结果达到预定终止条件,则终止模型训练,并获得所述的转录本翻译蛋白质产物的预测模型,即TRANSAID模型。

[0021] 在另一优选例中,所述数据集为人类转录本数据集。

[0022] 在另一优选例中,所述基于位置的嵌入包括步骤:

[0023] (a1)对输入序列中每个核苷酸进行编码;

[0024] (a2)对输出标签进行编码;

[0025] (a3)对所述编码的输入序列进行嵌入;

[0026] (a4)对所述嵌入进行位置编码;

[0027] 其中,在(a1)中,对所述输入序列中的每个核苷酸进行base编码;在(a2)中,对所述输出标签进行one-hot编码。

[0028] 在另一优选例中,在(a3)中,所述嵌入的维度为128。

[0029] 在另一优选例中,在(a4)中,所述位置编码的方法为正余弦位置编码。

[0030] 在另一优选例中,所述输出标签包括:翻译起始位点(TIS)、翻译终止位点(TTS)、其他标签。

[0031] 在另一优选例中,所述base编码将A配置为1、将C配置为2、将G配置为3、将U配置为4、将T配置为4和将未知核苷酸为0。

[0032] 在另一优选例中,所述one-hot编码将所述TIS配置为[1,0,0]、将所述TTS配置为[0,1,0]和将所述其他标签配置为[0,1,0]。

[0033] 在另一优选例中,所述卷积包括步骤:

[0034] (b1)使用3种大小的卷积核对所述基于位置的嵌入进行卷积,所述卷积核大小分别为3、5、7,每种所述卷积核的数量为32,卷积方式为“same”,获得卷积结果;

[0035] (b2)对所述卷积结果进行拼接,获得特征维度为96的卷积结果。

[0036] 在另一优选例中,所述4个残差块的空洞率分别为1、2、4、8。

[0037] 在另一优选例中,所述计算注意力权重包括步骤:

[0038] (c1)将所述查询和所述键的卷积核数量配置为输入维度的1/8;将所述值的卷积核数量配置为与输入维度一致;

[0039] (c2)将输入序列配置为长度为L、维度为C的张量,将所述张量转换为L×C/8的查询张量和键张量,以及L×C的值张量;

[0040] (c3)将所述查询张量与所述键张量相乘,通过激活函数得到L×L的注意力矩阵,将所述注意力矩阵与所述值张量相乘,得到聚焦后的序列表示;

[0041] (c4)将所述聚焦后的序列表示与所述中间特征表示相加,乘以标量权重。

[0042] 在另一优选例中,所述中间特征表示的形状为(batch_size,C,length),其中,所述batch_size为批次大小,所述C是特征通道数,所述length为序列长度。

[0043] 在另一优选例中,所述特征通道数为128。

[0044] 在另一优选例中,所述标量权重为一数值,所述数值用于平衡所述中间特征表示的信息和注意力聚焦后的信息。

[0045] 在另一优选例中,在(c3)中,所述激活函数为Softmax函数。

[0046] 在另一优选例中,所述解码包括步骤:

[0047] (d1)通过1×1卷积将所述注意力机制模块的输出的特征维度从128降至64,并通过激活函数激活;

[0048] (d2)通过Dropout层,以0.2的概率随机遮蔽特征;

[0049] (d3)通过1×1卷积将所述特征维度从64降至3,并通过转置函数将维度调整为(batch_size,seq_len,3),得到输出;

[0050] (d4)将所述输出与真实标签的交叉熵作为训练的损失函数,所述真实标签由所述数据集提供;

[0051] (d5)通过激活函数提供的最大概率值决定所述转录本序列每个位置的所述输出标签。

[0052] 在另一优选例中,在(d1)中,所述激活函数为Softmax函数。

[0053] 在另一优选例中,在(d3)中,所述转置函数为Permute函数。

[0054] 在另一优选例中,在(d5)中,所述激活函数为Softmax函数。

[0055] 本发明的第二方面,提供了一种构建转录本数值化转化模型的方法,所述方法包括步骤:

[0056] (A1)提供待数值化转化的转录本;

[0057] (A2)对所述转录本中每个核苷酸进行编码;

[0058] (A3)对所述转录本的标签进行编码;

[0059] (A4)对所述编码的输入序列进行嵌入,所述嵌入的维度为128;

[0060] (A5)对所述嵌入进行位置编码,所述位置编码的方法为正余弦位置编码;

[0061] (A6)获得转录本数值化转化模型,以及转录本数值化输出,所述转录本数值化输出为基于位置的嵌入。

[0062] 在另一优选例中,在(A2)中,对所述转录本中的每个核苷酸进行base编码。

[0063] 在另一优选例中,在(A3)中,对所述标签进行one-hot编码。

[0064] 在另一优选例中,所述标签包括:TIS、TTS、其他标签。

[0065] 本发明的第三方面,提供了一种构建转录本局部特征提取模型的方法,所述方法包括步骤:

[0066] (B1)提供基于位置的嵌入,所述基于位置的嵌入是通过本发明第二方面所述方法获得的;

[0067] (B2)使用3种大小的卷积核对所述基于位置的嵌入进行卷积,所述卷积核大小分别为3、5、7,每种所述卷积核数量为32,卷积方式为“same”,获得卷积结果;

[0068] (B3)对所述卷积结果进行拼接,获得特征维度为96的卷积结果;

[0069] (B4)对所述特征维度为96的卷积结果进行批标准化和线性整流激活;

[0070] (B5)获得转录本局部特征提取模型,以及转录本局部特征数据。

[0071] 本发明的第四方面,提供了一种构建转录本全局交互模型的方法,所述方法包括步骤:

[0072] (C1)提供转录本局部特征数据,所述转录本局部特征数据是通过本发明第三方面所述方法获得的;

[0073] (C2)使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26的卷积核对所述转录本局部特征数据进行卷积,所述卷积核的数量为64,获得卷积结果;

[0074] (C3)对所述卷积结果进行1×1卷积,将所述卷积结果的特征维度从64变为128,获得高维卷积结果;

[0075] (C4)对所述高维卷积结果进行批标准化和线性整流激活;

[0076] (C5)使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26卷积核对(C4)的输出进行卷积,所述卷积核的数量为128;

[0077] (C6)获得转录本全局交互模型,以及转录本全局交互特征数据。

[0078] 本发明的第五方面,提供了一种构建转录本全局交互特征的注意力聚焦模型的方法,所述方法包括步骤:

[0079] (D1)提供转录本全局交互特征数据,所述转录本全局交互特征数据是通过本发明第四方面所述方法获得的;

[0080] (D2)利用3个1×1的卷积,分别生成查询/键/值;

[0081] (D3)将所述查询和所述键的卷积核数量配置为输入维度的1/8;将所述值的卷积核数量配置为与输入维度一致;

[0082] (D4)将所述转录本全局交互特征数据配置为长度为L、维度为C的张量,将所述张量转换为L×C/8的查询张量和键张量,以及L×C的值张量;

[0083] (D5)将所述查询张量与所述键张量相乘,通过激活函数得到L×L的注意力矩阵,将所述注意力矩阵与所述值张量相乘,得到聚焦后的序列表示;

[0084] (D6)将所述聚焦后的序列表示与所述转录本全局交互特征数据相加,乘以标量权重;

[0085] (D7)获得转录本全局交互特征的注意力聚焦模型,以及注意力聚焦输出。

[0086] 在另一优选例中,所述标量权重为一数值,所述数值用于平衡转录本全局交互特征数据的信息和注意力聚焦后的信息。

[0087] 本发明的第六方面,提供了一种构建转录本高阶特征解码模型的方法,所述方法包括步骤:

[0088] (E1)提供转录本高阶特征数据,所述转录本高级特征数据是通过本发明第五方面所述方法获得的;

[0089] (E2)对所述转录本高阶特征数据进行1×1卷积,将所述转录本高阶特征数据的特征维度从128降至64,并通过Softmax函数激活;

[0090] (E3)通过Dropout层,以0.2的概率随机遮蔽特征;

[0091] (E4)通过1×1卷积将所述特征维度从64降至3,并通过Permute函数将维度调整为(batch_size,seq_len,3),得到输出;

[0092] (E5)将所述输出与真实标签的交叉熵作为训练的损失函数;

[0093] (E6)通过Softmax函数提供的最大概率值决定转录本序列每个位置的输出标签;

[0094] (E7)获得转录本高阶特征解码模型,以及所述转录本序列每个位置的输出标签。

[0095] 在另一优选例中,所述真实标签来自参考序列数据库。

[0096] 在另一优选例中,所述参考序列数据库包括Refseq。

[0097] 在另一优选例中,所述参考序列数据库包括人源转录本序列信息和人源转录本序列标签信息。

[0098] 本发明的第七方面,提供了一种转录本翻译蛋白质产物的预测系统,所述系统包括:

[0099] 输入单元,所述输入单元被配置为输入数据,所述输入数据包括待预测的输入序列;

[0100] 转录本翻译蛋白质产物的预测单元,所述预测单元被配置为执行一转录本翻译蛋白质产物的预测模型,获得所述待预测的输入序列的位置标签,从而获得所述待预测的输入序列的翻译位点和翻译产物;其中,所述预测模型为用本发明第一方面所述的方法构建的;

[0101] 输出单元,所述输出单元被配置为输出所述转录本翻译蛋白质产物的预测单元的预测结果。

[0102] 在另一优选例中,所述待预测的输入序列包括真核生物的转录组。

[0103] 在另一优选例中,所述真核生物的转录组包括人类转录组、小鼠转录组、斑马鱼转录组、酿酒酵母转录组。

[0104] 在另一优选例中,所述转录本翻译蛋白质产物的预测系统对真核生物转录组的预测准确率保持稳定,所述“保持稳定”是指,与所述预测系统对人类转录组的预测准确率相比M0,所述预测系统对其他真核生物转录组的预测准确率M1,满足M1/M0≥98%,较佳地≥98.5%,更佳地≥99%,所述其他真核生物转录组包括小鼠转录组、斑马鱼转录组、酿酒酵母转录组。

[0105] 本发明的第八方面,提供了一种对转录本进行数值化转化的系统,所述系统包括:

[0106] 输入单元,所述输入单元被配置为输入数据,所述输入数据包括待数值化转化的输入序列;

[0107] 数值化转化单元,所述数值化转化单元被配置为执行一数值化转化模型,获得所述待数值化转化的输入序列的基于位置的嵌入,其中,所述数值化转化模型是用本发明第二方面所述方法构建的;

[0108] 输出单元,所述输出单元被配置为输出所述转录本数值化转化的结果。

[0109] 本发明的第九方面,提供了一种转录本翻译蛋白质产物的预测系统的用途,用于研究转录本序列中UTR区域和密码子完整性对结构标签的影响。

[0110] 在另一优选例中,对UTR区域进行扰动,所述TRANSAID模型对结构标签的预测准确性下降。

[0111] 在另一优选例中,所述UTR区域包括5'UTR区域和3'UTR区域。

[0112] 在另一优选例中,所述结构标签选自下组:TIS、TTS、开放阅读框(ORF)、或其组合。

[0113] 在另一优选例中,对所述5'UTR区域进行扰动,所述TRANSAID模型对TIS的预测准确性显著下降,所述"对TIS的预测准确性显著下降"是指,与不进行扰动时所述TRANSAID模型对TIS的预测准确性P0相比,对5'UTR区域进行扰动时所述TRANSAID模型对TIS的预测准确性P1,满足P1/P0≤82%。

[0114] 在另一优选例中,对所述3'UTR区域进行扰动,所述TRANSAID模型对TIS的预测准确性没有显著变化。

[0115] 在另一优选例中,对所述5'UTR区域和3'UTR区域进行扰动,所述TRANSAID模型对TIS的预测准确性显著下降,所述"对TIS的预测准确性显著下降"是指,与不进行扰动时所述TRANSAID模型对TIS的预测准确性P0相比,对5'UTR区域和3'UTR区域进行扰动时所述TRANSAID模型对TIS的预测准确性P2,满足P2/P0≤81%。

[0116] 在另一优选例中,对所述5'UTR区域进行扰动,所述TRANSAID模型对TTS的预测准确性没有显著变化。

[0117] 在另一优选例中,对所述3'UTR区域进行扰动,所述TRANSAID模型对TTS的预测准确性没有显著变化。

[0118] 在另一优选例中,对所述5'UTR区域和3'UTR区域进行扰动,所述TRANSAID模型对TTS的预测准确性没有显著变化。

[0119] 在另一优选例中,对所述转录本序列进行删除或插入3个碱基,所述TRANSAID模型的预测准确性没有显著变化。

[0120] 在另一优选例中,对所述转录本序列进行删除或插入1或2个碱基,所述TRANSAID模型的预测准确性显著下降,所述"预测准确性显著下降"是指,与删除或插入3个碱基时所述TRANSAID模型的预测准确性R0相比,进行删除或插入1或2个碱基时所述TRANSAID模型的预测准确性R1,满足R1/R0≤7%。

[0121] 应理解,在本发明范围内中,本发明的上述各技术特征和在下文(如实施例)中具体描述的各技术特征之间都可以互相组合,从而构成新的或优选的技术方案。限于篇幅,在此不再一一累述。

附图说明

[0122] 图1为TRANSAID模型介绍。

[0123] 图2为TRANSAID模型在碱基水平的预测混淆矩阵。

[0124] 图3为TRANSAID模型在碱基水平的性能统计。

[0125] 图4为TRANSAID模型在三联密码子水平对开放阅读框预测性能。

[0126] 图5显示了UTR的扰动翻译起始位点预测的影响。

[0127] 图6显示了UTR的扰动对翻译终止位点预测的影响。

[0128] 图7显示了UTR的扰动对整体开放阅读框预测的影响。

[0129] 图8显示了CDS区域的碱基插入与缺失对整体开放阅读框预测的影响。

[0130] 图9显示了TRANSAID的预测结果。

[0131] 图10显示了GeneMarks-T的预测结果。

具体实施方式

[0132] 本发明人经过广泛而深入的研究,首次开发了一种基于深度学习的mRNA翻译蛋白质产物的预测方法与模型。具体地,本发明设计了一种基于深度学习方法的转录组水平翻译产物精确预测的方法,用于对转录本序列的翻译位点和翻译产物进行预测,形成TRANSAID模型。第一,本发明的TRANSAID模型采用了独特的序列编码策略和深度学习架构,能够同时捕捉局部序列特征和长距离依赖关系,大大提高了预测准确性;第二,本发明的TRANSAID模型能够对不同物种的转录本序列进行预测,具有本领域中首次出现的跨物种预测能力;第三,本发明的TRANSAID模型能够识别5'UTR和3'UTR区域对翻译调控的不同影响。在此基础上完成了本发明。

[0133] 应当理解,以下以各种详细程度描述本发明的具体方法和实验条件、是用于提供对本发明的实质理解。下面提供了本说明书中使用的某些术语的定义。除非另外定义,否则本文中所用的全部技术与科学术语均具有如本发明所属领域的普通技术人员通常理解的相同含义。

[0134] 术语

[0135] 在提供数值范围的情况下,除非上下文另外清楚地指出,否则应当理解,该值20的每个中间整数、该值的每个中间整数的每十分之一、在该范围的上限与下限之间和在该规定范围中的任何其他中间值都包括在本发明内。这些较小范围的上限和下限可以独立地包括在较小范围内,并且也涵盖在本发明内,但须遵守规定范围内的任何明确排除的限制。例如,“1至50”包括“2至25”、“5至20”、“25至50”、“1至10”等。

[0136] 如本文所用,术语“含有”或“包括(包含)”可以是开放式、半封闭式和封闭式的。换言之,所述术语也包括“基本上由……构成”、或“由……构成”。

[0137] 如本文所用,术语“转录本”、“转录组”、“转录本序列”可以互换使用,均是指由几个基因通过转录形成的一种或多种成熟的RNA序列,包含编码RNA和非编码RNA。

[0138] 如本文所用,术语“TRANSAID(TRANSlation Analysis through Intelligent Deep learning)模型”是指在本发明中利用深度学习方法搭建的用于预测转录本翻译位点和翻译产物的模型。

[0139] 本发明的构建转录本翻译蛋白质产物的预测模型的方法

[0140] 本发明提供了一种构建转录本翻译蛋白质产物的预测模型的方法,利用该方法构建的模型即为TRANSAID。

[0141] 具体地,所述方法包括步骤:

[0142] (S1)提供用于模型训练的数据集,对所述数据集进行数值化转化;其中,所述数据集为转录本序列,所述转录本序列包括mRNA序列和非编码RNA序列;所述数值化为基于位置的嵌入,包括对输入序列中每个核苷酸和输出标签进行编码,优选地,用base编码对输入序列中每个核苷酸进行编码,用one-hot编码对输出标签进行编码。

[0143] (S2)将(S1)中所述的数据集用于模型架构的训练;

[0144] 其中,所述模型架构包括基于多尺度卷积模块、残差网络模块和注意力机制模块的模型架构;

[0145] 所述多尺度卷积模块卷积核对所述基于位置的嵌入进行卷积;

[0146] 使用3种大小的卷积核对所述基于位置的嵌入进行卷积,所述卷积核大小分别为3、5、7,每种所述卷积核的数量为32,卷积方式为“same”,并对结果进行拼接,获得96维的卷积结果。

[0147] 卷积结果经过批标准化(batch normalization)和线性整流(rectified linear unit,ReLU)激活,输出到下游残差网络模块;所述残差网络模块包括两个部分:

[0148] (Z1)残差网络模块第一部分:使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26的卷积核对上游输出进行卷积,所述卷积核的数量为64,获得卷积结果;对所述卷积结果进行1×1卷积,获得高阶卷积结果;对所述高阶卷积结果进行批标准化和线性整流激活,输出到下游残差网络模块第二部分;

[0149] (Z2)残差网络模块第二部分:使用4个残差块,每个所述残差块有2个空洞卷积,使用大小为26卷积核对上游输出进行卷积,所述卷积核的数量为128,经过批标准化和线性整流激活后,获得中间特征表示,将所述中间特征表示输出到下游注意力机制模块。

[0150] 优选地,上述4个残差块的空洞率分别为1、2、4、8。

[0151] 所述注意力机制模块包含3个1×1卷积,分别用于生成查询/键/值(Query/Key/Value),通过以下步骤计算注意力权重:

[0152] (c1)将所述查询和所述键的卷积核数量配置为输入维度的1/8;将所述值的卷积核数量配置为与输入维度一致;

[0153] (c2)将输入序列配置为长度为L、维度为C的张量,将所述张量转换为L×C/8的查询张量和键张量,以及L×C的值张量;

[0154] (c3)将所述查询张量与所述键张量相乘,通过激活函数得到L×L的注意力矩阵,将所述注意力矩阵与所述值张量相乘,得到聚焦后的序列表示;

[0155] (c4)将所述聚焦后的序列表示与所述中间特征表示相加,乘以标量权重;获得注意力机制模块的输出。

[0156] (S3)对(S2)中所述注意力机制模块的输出进行解码;通过以下步骤进行解码:

[0157] (d1)通过1×1卷积将所述注意力机制模块的输出的特征维度从128降至64,并通过激活函数激活;

[0158] (d2)通过Dropout层,以0.2的概率随机遮蔽特征;

[0159] (d3)通过1×1卷积将所述特征维度从64降至3,并通过转置函数将维度调整为

(batch_size,seq_len,3),得到输出;

[0160] (d4)将所述输出与真实标签的交叉熵作为训练的损失函数,所述真实标签由所述数据集提供;

[0161] (d5)通过激活函数提供的最大概率值决定所述转录本序列每个位置的所述输出标签得到预测结果。

[0162] 优选地,所述激活函数为Softmax函数。

[0163] 优选地,所述转置函数为Permute函数。

[0164] (S4)当训练结果达到预定终止条件,则终止模型训练,并获得所述的转录本翻译蛋白质产物的预测模型,即TRANSAID模型。

[0165] RefSeq数据库

[0166] RefSeq数据库是一个权威的参考序列数据库,其数据经过了严格的质量控制和人工审核,每条序列都具有详细的功能注释信息。与其他数据库相比,RefSeq的一个显著优势在于它提供了非冗余的、经过验证的参考序列,这对于训练机器学习模型特别重要,因为它可以降低数据噪声,提高模型的学习效率。

[0167] TRANSAID的训练数据

[0168] 如本文所用,术语“信使RNA”、“mRNA”和“NM类转录本”可以互换使用;术语“非编码RNA”和“NR类转录本”可以互换使用。

[0169] NM和NR为RefSeq数据集的编号开头,其中,NM表示与mRNA有关的数据集条目,NR表示与非编码RNA有关的数据集条目。

[0170] 本发明重点关注信使RNA(NM类转录本)和非编码RNA(NR类转录本)的数据。NM类转录本代表了经典的蛋白质编码基因,而NR类转录本则包含了可能存在非经典翻译事件的序列。通过同时考虑这两类转录本,模型能够学习到更广泛的翻译调控模式。在数据集划分上,采用了80:20的比例进行训练集和验证集的划分,这是深度学习领域普遍采用的比例,可以在保证充足训练样本的同时,提供足够的验证数据来评估模型性能。

[0171] Base编码

[0172] 本发明采用Base编码方式对输入RNA数据进行编码,将A、C、G、T/U分别映射为1、2、3、4的数值,同时使用0作为填充符号。这种编码方式相比传统的one-hot编码具有多个优势:首先,它大大减少了数据的存储空间和计算开销,因为每个位置只需要一个整数而不是一个四维向量;其次,这种编码方式为后续的嵌入(Embedding)层提供了理想的输入格式,使得模型能够在训练过程中学习到更丰富的序列特征表示。

[0173] One-hot编码

[0174] 本发明采用one-hot编码方式对序列标签进行编码,将每个位置的标签编码为三维向量:[1,0,0]表示翻译起始位点(TIS),[0,1,0]表示翻译终止位点(TTS),[0,0,1]表示非特殊位点。采用One-hot编码可以保证各类标签之间的等价性,避免了数值编码可能带来的偏差;并且这种编码方式便于计算损失函数和设计模型的输出层。此外,对标签进行非连续的编码更符合生物学认知,因为TIS、TTS和非特殊位点在功能上是互斥的离散状态,而不是连续变化的数值关系。

[0175] 标量权重

[0176] 如本文所用,术语“标量权重(gamma)”是注意力机制中的一个可学习参数,该

gamma参数的作用是动态调节注意力机制的影响程度:

[0177] (1)初始化时,gamma被设为0,此时注意力机制不起作用,输出是原序列表示。在本发明的模型中,所述原序列表示为卷积层和残差块处理后得到的中间特征表示。

[0178] (2)在训练过程中,gamma会逐渐学习到一个合适的值,用于平衡原序列信息和注意力加权后的信息。

[0179] 本发明的主要优点包括:

[0180] (1)本发明的TRANSAID模型采用了独特的序列编码策略和深度学习架构,能够同时捕捉局部序列特征和长距离依赖关系,使得模型能够理解复杂的调控模式,大大提高了预测准确性。

[0181] (2)本发明的TRANSAID模型具有优异的泛化能力,能够准确预测标准翻译事件,还能识别非典型翻译位点,并且能够在人类、小鼠、斑马鱼和酵母等不同物种中保持高准确率,具有本领域中首次出现的跨物种预测能力。

[0182] (3)本发明的TRANSAID模型的预测结果具有很强的生物学解释性,研究发现模型成功学习了遗传密码子的三联体规则,并能够识别5'UTR和3'UTR区域对翻译调控的不同影响,为理解翻译调控机制提供了新的见解。

[0183] (4)本发明的TRANSAID模型在人类转录组预测中,其准确匹配率高达97.5%,远超GeneMarks-T(69.6%)。在标准翻译事件和非典型翻译事件方面,都具有极高的预测准确率。

[0184] 下面结合具体实施例,进一步阐述本发明。应理解,这些实施例仅用于说明本发明而不用于限制本发明的范围。下列实施例中未注明具体条件的实验方法,通常按照常规条件,例如Sambrook等人,分子克隆:实验室手册(New York:Cold Spring Harbor Laboratory Press,1989)中所述的条件,或按照制造厂商所建议的条件。除非另外说明,否则百分比和份数是重量百分比和重量份数。

[0185] 实施例1:TRANSAID训练数据的准备与编码

[0186] 本实施例涉及TRANSAID的训练数据集以及序列编码策略的选择。

[0187] 训练数据集:本实施例选择使用NCBI RefSeq数据库中的GRCh38最新版本RNA序列数据(GRCh38_latest_rna.fna)和对应的基因组注释文件(GRCh38_latest_rna.gbff)。所述RNA序列数据包括信使RNA(NM类转录本)和非编码RNA(NR类转录本)。其中,NM类转录本总计67,072条,NR类转录本23,078条。按照80:20的比例划分训练集和验证集。

[0188] 序列编码策略:本实施例选择采用双轨编码方式,分别针对输入序列(碱基序列)和输出标签(翻译位点)采用不同的编码策略。对于输入的RNA序列,采用了base编码方式,将A、C、G、T/U分别映射为1、2、3、4的数值,同时使用0作为填充符号。对于输出标签,项目采用了one-hot编码方式,将每个位置的标签编码为三维向量:[1,0,0]表示翻译起始位点(TIS),[0,1,0]表示翻译终止位点(TTS),[0,0,1]表示非特殊位点。

[0189] 结果:

[0190] 对于输入序列,所使用的base编码可以配合嵌入层的设计,使得模型能够自动学习碱基之间的协同关系。例如,在翻译起始位点周围,常见的Kozak序列(GCCACC)就体现了碱基之间的特定组合模式;在密码子识别过程中,三个相邻碱基的组合决定了对应的氨基酸,这种三联体模式也能够被模型通过嵌入空间中的距离关系很好地捕捉。

[0191] 对于输出标签,所使用的one-hot编码方式的合理性体现在以下几个方面:首先,它保证了各类标签之间的等价性,避免了数值编码可能带来的偏差;其次,这种编码方式便于计算损失函数和设计模型的输出层;最重要的是,它符合生物学认知,因为TIS、TTS和非特殊位点在功能上是互斥的离散状态,而不是连续变化的数值关系。

[0192] 在实际训练过程中,上述编码策略展现出了显著的优势。从计算效率来看,base编码显著减少了模型的参数量和内存占用;从学习效果来看,通过嵌入层,模型成功捕捉到了复杂的序列模式,这一点从删除/插入实验结果中得到了验证,当删除或插入3个碱基时,模型的预测准确率保持稳定,这说明模型确实学习到了遗传密码子的三联体特性。

[0193] 实施例2:TRANSAID模型的构建

[0194] 本实施例涉及TRANSAID模型的构建,包括输入数据处理、嵌入、多尺度卷积、两步残差网络、注意力机制、输出数据处理等。

[0195] TRANSAID模型是一个用于预测RNA翻译起始位点(TIS)和终止位点(TTS)的深度学习模型。下面将从数据流的角度详细阐述该模型的结构与特点。该模型的流程如图1所示。

[0196] 步骤01:输入RefSeq转录本序列。对序列进行数值化表示:

[0197] (1)序列中的每个核苷酸(A/C/G/T/U)用一个整数编码表示,分别对应1/2/3/4/4,未知核苷酸编码为0;

[0198] (2)序列中的标签用三维向量表示,其中翻译起始位点对应[1,0,0],翻译终止位点对应[0,1,0],非特殊位点对应[0,0,1]。

[0199] 步骤02:采用基于位置的嵌入方法,对编码后的序列进行嵌入,将离散的编码映射为连续的向量表示:

[0200] (1)嵌入层共有5种输入(对应编码0-4),嵌入维度为128。这种嵌入能捕捉到核苷酸之间的相似性信息;

[0201] (2)在嵌入之后,引入了位置编码(Positional Encoding),来表示序列中每个位置的先后顺序关系。具体采用的是正余弦位置编码,将绝对位置映射为嵌入空间的方向性信息。

[0202] 输入数据经过步骤01和步骤02的输入数据表示方法的处理,能为后续模型提供丰富的特征,包括碱基的类型,以及碱基在序列中的位置关系。

[0203] 步骤03:局部特征提取。采用多尺度卷积,从序列的分布式中提取局部的序列特征。该步骤可以视为模型的“感受野”,从不同大小的窗口来“观察”序列:

[0204] (1)使用3种不同核大小(3/5/7)的一维卷积对嵌入层输出进行卷积,每种卷积核数量为32,卷积方式为“same”,即输出序列长度与输入一致;

[0205] (2)在特征维度上对3种卷积的输出进行拼接,实现了多尺度的特征融合,融合后的特征维度为96;

[0206] (3)卷积结果经过批标准化与ReLU函数激活,引入了非线性因素,提高了模型的表达能力。

[0207] 经过步骤03,模型提取了序列中的潜在相关性与规律性。在批标准化过程中,在python v3.12.3版本环境中采用pytorch v2.4.0软件的BatchNorm1d函数进行批标准化处理。

[0208] 步骤04:全局特征交互。通过残差网络建模全局范围内的特征交互,即整合局部特

征,探索序列的长程联系:

[0209] (1)第一部分采用4个残差块,每个残差块中有2个一维卷积,卷积核数量为64,卷积核大小为26,空洞率(dilation)分别为1/2/4/8,这种空洞卷积能扩大感受野,探索更长程的特征关联;

[0210] (2)残差块的输出先经过一个1×1卷积,将特征维度从64变为128,再通过批标准化和ReLU激活,作为下一部分的输入;

[0211] (3)第二部分采用4个与第一部分类似的残差块,卷积核数量为128,逐层加深了模型的理解深度。

[0212] 经过步骤04,两部分残差块的设计引入了跨层的信息流,使得浅层特征能与深层特征充分交互,形成了对序列全局的理解。这种残差结构使得模型能探索序列中蕴含的层次化语义信息,全局地建模TIS/TTS这一序列标注任务的内在规律。同时跨层连接也缓解了梯度消失,使训练更加稳定。

[0213] 步骤05:注意力汇聚。引入注意力机制,通过聚焦任务相关特征,进一步提炼与翻译紧密相关的信息:

[0214] (1)注意力模块包含三个1×1卷积,分别用于生成Query/Key/Value。其中Query和Key的卷积核数量为输入维度的1/8,Value的卷积核数量与输入一致;

[0215] (2)将输入序列看作是一个长度为L(序列长度)、维度为C(通道数)的张量,注意力机制先将其变换为L×C/8的Query张量和Key张量,以及L×C的Value张量;

[0216] (3)Query与Key的转置相乘,并通过softmax函数得到L×L的注意力矩阵,代表了序列中每个位置对其他位置的关注程度。注意力矩阵与Value相乘,得到聚焦后的序列表示;

[0217] (4)聚焦后的序列表示与原序列表示相加,并乘以一个可学习的标量权重,作为注意力模块的输出。这种残差连接方式可以调控注意力机制的影响力。其中,所述原序列表示为经过步骤03和步骤04处理后得到的中间特征表示。

[0218] 所述标量权重(gamma)是注意力机制中的一个可学习参数,该gamma参数的作用是动态调节注意力机制的影响程度:

[0219] (a)初始化时gamma被设为0,此时注意力机制不起作用,输出就是原序列表示;

[0220] (b)在训练过程中,gamma会逐渐学习到一个合适的值,用于平衡原序列信息和注意力加权后的信息;

[0221] (c)最终的注意力模块输出计算公式为:output＝(x+self.gamma)*attention_weighted_sequence

[0222] 经过步骤05,注意力机制使得模型能根据任务目标自适应地调整对输入信息的关注,突出强调对预测翻译更重要的特定区域,提高了模型的泛化和鲁棒性。

[0223] 步骤06:输出解码。模型将提炼后的高阶特征解码,得到每个位置的TIS/TTS预测结果:

[0224] (1)先通过一个1×1卷积将特征维度从128降为64,并过ReLU激活;

[0225] (2)再通过一个Dropout层,以0.2的概率随机遮蔽特征。这种正则方式能提高模型的泛化能力;

[0226] (3)最后通过一个1×1卷积将特征维度从64变为3(代表TIS/TTS/Other三种状

态),再通过Permute操作将维度调整为(batch_size,seq_len,3),得到最终输出;

[0227] (4)模型输出与真实标签的交叉熵作为训练的损失函数。预测时,每个位置的状态可通过softmax后的最大概率值来决定。

[0228] 至此,模型完成了从原始RNA序列到每个位置TIS/TTS状态的端到端预测。输出形式与输入序列一一对应,方便直观地解释每个碱基的预测结果。

[0229] 实施例3:TRANSAID基础预测性能的评估

[0230] 本实施例涉及采用RefSeq数据集,从基础预测角度对TRANSAID的性能进行评估,所述数据集包括GRCh38_latest_rna.fna和GRCh38_latest_rna.fna.gzGRCh38_latest_rna.gbf f。

[0231] 在碱基级别的评估中,混淆矩阵分析显示TRANSAID模型展现出极高的预测准确性(图2)。对于起始密码子(TIS)的预测,模型实现了39,371个正确预测,仅有4个假阴性和777个假阳性预测;对于终止密码子(TTS)的预测,模型成功识别了39,790个位点,仅有457个假阴性和534个假阳性预测。这些数据表明模型在处理转录本的关键功能位点时具有极高的精确度。

[0232] 性能矩阵进一步量化了模型的预测能力。通过统计计算,发现模型达到了98.91%的精确率、98.89%的召回率和98.90%的F1分数(图3)。具体来看,对于TIS预测,模型获得了98.05%的精确率和97.81%的召回率;对于TTS预测,则达到了98.68%的精确率和98.85%的召回率。这种均衡的高性能指标证明了模型在不同预测任务上的稳定性。

[0233] 实施例4:TRANSAID完整性预测性能的评估

[0234] 本实施例涉及采用RefSeq数据集,从完整性预测角度对TRANSAID的性能进行评估,所述数据集包括GRCh38_latest_rna.fna和GRCh38_latest_rna.fna.gzGRCh38_latest_rna.gbf f。

[0235] 除了在单碱基水平上进行性能评估外,还在三联密码子水平上进行评估,因为只有所有翻译起始位点和翻译终止位点同时正确,才能得到准确的翻译阅读框ORF。在这方面,本发明模型具有非常出色的性能(图4),统计发现95.51%的转录本能够被完全正确预测,即模型能够准确识别出完整的开放阅读框(ORF)。对于预测错误的情况,仅有1.60%的转录本出现TIS预测错误,2.24%的转录本出现TTS预测错误,而其他类型的错误仅占0.66%。这些数据表明本发明模型不仅能够准确预测单个位点,更能把握转录本的整体翻译模式。

[0236] 实施例5:序列特征对TRANSAID性能的影响分析

[0237] 本实施例涉及采用RefSeq数据集,探究序列特征对TRANSAID性能的影响分析所述,数据集包括GRCh38_latest_rna.fna和GRCh38_latest_rna.fna.gzGRCh38_latest_rna.gbff。

[0238] 为了深入理解模型的预测机制,研究团队进行了一系列序列扰动实验。实验结果如下:

[0239] 1.UTR序列的影响:5'UTR区域的扰动对TIS预测的影响显著,当打乱5'UTR序列时,TIS预测的准确率从98.05%降至80.49%(图5)。相比之下,3'UTR的扰动对预测的影响相对较小,表明模型成功捕捉到了5'UTR在翻译起始调控中的关键作用(图6)。在整个翻译阅读框ORF预测准确度水平,也同样反应5'UTR序列对模型有显著影响(图7)。这说明模型学习到

了5'UTR序列特征。

[0240] 2.密码子完整性验证:通过碱基删除/插入实验,模型展现出对遗传密码子框架的深刻理解。当删除或插入3个碱基时,预测准确率保持在90%以上;而当删除或插入1-2个碱基时,准确率显著下降至约5%,这与生物学中密码子三联体的概念完全吻合(图8)。

[0241] 实施例6:TRANSAID的跨物种预测能力的评估

[0242] 本实施例涉及利用多个物种的转录组数据,对TRANSAID的跨物种预测能力进行评估。

[0243] 结果显示,TRANSAID模型展现出优秀的跨物种预测能力:

[0244] (1)人类(H.sapiens,参考版本为GRCh38,GRCh38_latest_rna.fna和GRCh38_latest_rna.fna.gzGRCh38_latest_rna.gbff:作为训练物种,模型达到了95.52%的ORF预测准确率,TIS和TTS的F1分数分别为97.93%和98.77%。

[0245] (2)小鼠(M.musculus,数据集来源分别为GCF_000001635.27_GRCm39_rna.fna和GCF_000001635.27_GRCm39_rna.gbff):模型保持了95.46%的预测准确率,F1分数分别为97.50%和99.09%。

[0246] (3)斑马鱼(D.rerio数据集来源分别为/GCF_000002035.6_GRCz11_rna.fna和GCF_000002035.6_GRCz11_rna.gbff):尽管进化距离较远,模型仍达到94.11%的准确率,F1分数分别为96.22%和99.06%。

[0247] (4)酿酒酵母(S.ludwigii数据集来源分别为fungi.14.rna.fna和fungi.14.rna.gbff):即便在单细胞真核生物中,模型仍保持94.18%的准确率,F1分数分别为97.26%和98.88%。

[0248] 这一特性在实验中得到了充分验证,结果见表1。这种跨物种的稳定表现证明了模型捕捉到了真核生物翻译机制的共同特征,同时也说明真核生物的翻译机制是充分保守的。

[0249] 表1TRANSAID模型展现出优秀的跨物种预测性能统计

[0250]

物种	所有碱基正确且 ORF 正确	TIS 的 F1 分数	TTS 的 F1 分数
人类	95.52%	97.93%	98.77%
小鼠	95.46%	97.50%	99.09%
斑马鱼	94.11%	96.22%	99.06%
酿酒酵母	94.18%	97.26%	98.88%

[0251] 实施例6:TRANSAID与现有模型性能比较

[0252] 本实施例涉及对TRANSAID与GeneMarks-T的预测性能进行比较。

[0253] 在与现有方法的对比中,TRANSAID展现出显著优势。与GeneMarks-T进行比较的方法如下:

[0254] 用H.sapiens/M.musculus数据评估方法:

[0255] (A)分别用TRANSAID和GeneMarks-T对RefSeq的NM转录本序列(GRCh38_latest_rna.fna)进行翻译分析得到蛋白序列,然后将蛋白序列与RefSeq中注释的蛋白序列(GRCh38_latest_protein.faa)进行对比;

[0256] (B)根据预测蛋白序列和RefSeq注释蛋白序列对比情况,可分为:

[0257] a)精确匹配(Exact Match);

[0258] b)RefSeq是预测蛋白的子集(RefSeq is Predicted subset);

[0259] c)预测蛋白是RefSeq的子集(Predicted is RefSeq subset);

[0260] d)部分重叠(Partial overlap);

[0261] (C)统计以上对比类型中的序列数、比例;

[0262] (D)把非精确匹配的蛋白质序列,将预测序列和RefSeq注释序列比对分析相似度,并按照相似度从高到底排序。

[0263] 结果:

[0264] 1.精确匹配率:TRANSAID达到97.5%的精确匹配比例,远超GeneMarks-T的69.6%。

[0265] 2.预测覆盖度:在预测结果的分布上,TRANSAID表现更加集中和准确,仅有0.7%的预测属于“RefSeq is Predicted subset”类型,1.5%属于“部分重叠”类型,而GeneMarks-T在这些类别上的比例分别为21.5%和7.4%。

[0266] 这些结果清楚地表明TRANSAID在准确性和可靠性方面都优于现有方法。与其他工具(如Orfinder和EXpasy)相比,TRANSAID不仅避免了假阳性预测泛滥的问题,还提供了更精确的位点定位。

[0267] 在本发明提及的所有文献都在本申请中引用作为参考,就如同每一篇文献被单独引用作为参考那样。此外应理解,在阅读了本发明的上述讲授内容之后,本领域技术人员可以对本发明作各种改动或修改,这些等价形式同样落于本申请所附权利要求书所限定的范围。

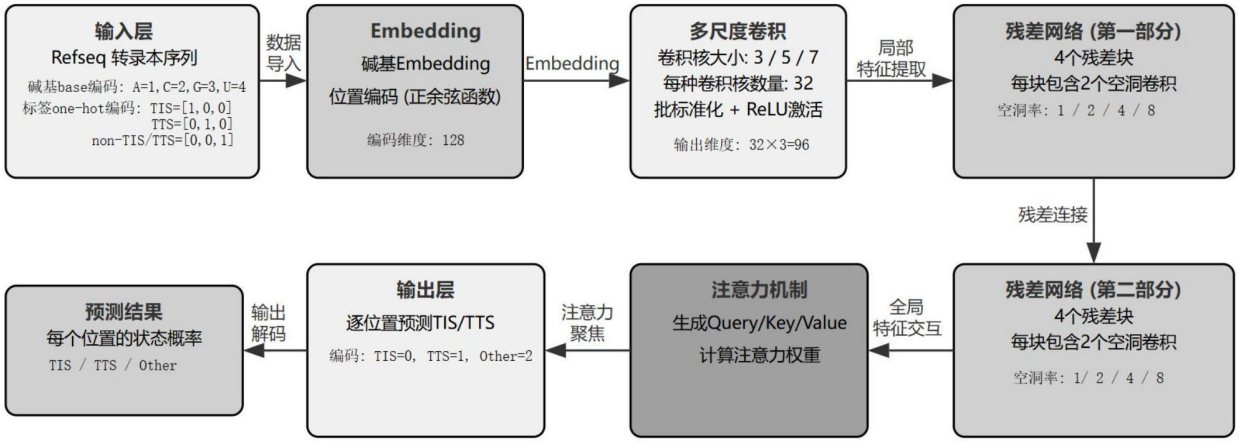
输入层
Refseq 转录本序列
碱基base编码：A=1, C=2, G=3, U=4
标签one-hot编码：TIS=[1, 0, 0]
TTS=[0, 1, 0]
non-TIS/TTS=[0, 0, 1]
数据导入
Embedding
碱基Embedding
位置编码 (正余弦函数)
编码维度：128
Embedding
多尺度卷积
卷积核大小: 3 / 5 / 7
每种卷积核数量: 32
批标准化 + ReLU激活
输出维度：32×3=96
局部特征提取
残差网络 (第一部分)
4个残差块
每块包含2个空洞卷积
空洞率：1 / 2 / 4 / 8
残差连接
残差网络 (第二部分)
4个残差块
每块包含2个空洞卷积
空洞率：1/ 2 / 4 / 8
全局特征交互
注意力机制
生成Query/Key/Value
计算注意力权重
注意力聚焦
输出层
逐位置预测TIS/TTS
编码：TIS=0, TTS=1, Other=2
输出解码
预测结果
每个位置的状态概率
TIS / TTS / Other

图1

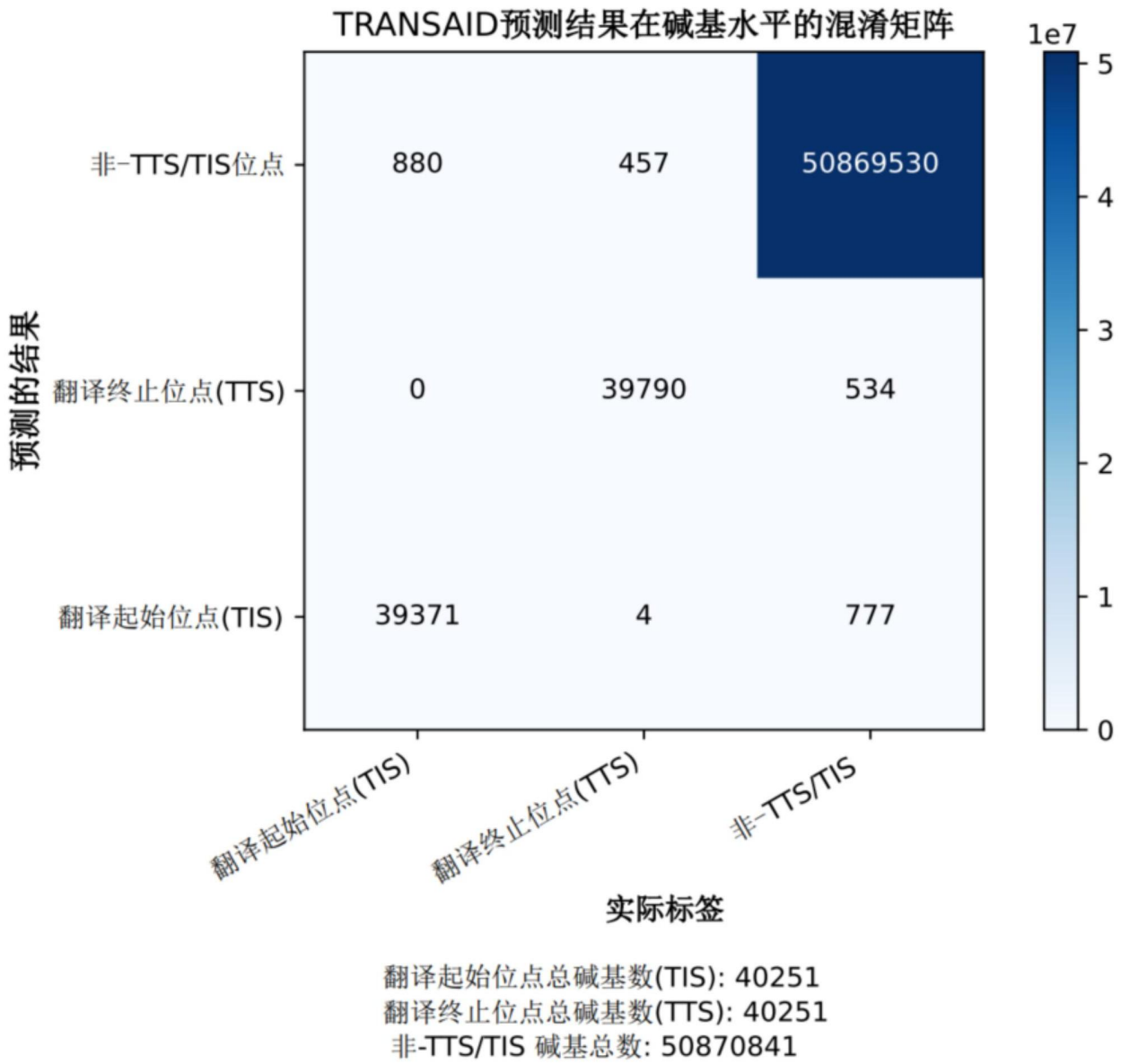
TRANSAID预测结果在碱基水平的混淆矩阵
1e7
5
4
3
2
1
0
非-TTS/TIS位点
翻译终止位点(TTS)
翻译起始位点(TIS)
880
457
50869530
0
39790
534
39371
4
777
预测的结果
翻译起始位点(TIS)
翻译终止位点(TTS)
非-TTS/TIS
实际标签
翻译起始位点总碱基数(TIS): 40251
翻译终止位点总碱基数(TTS): 40251
非-TTS/TIS 碱基总数: 50870841

图2

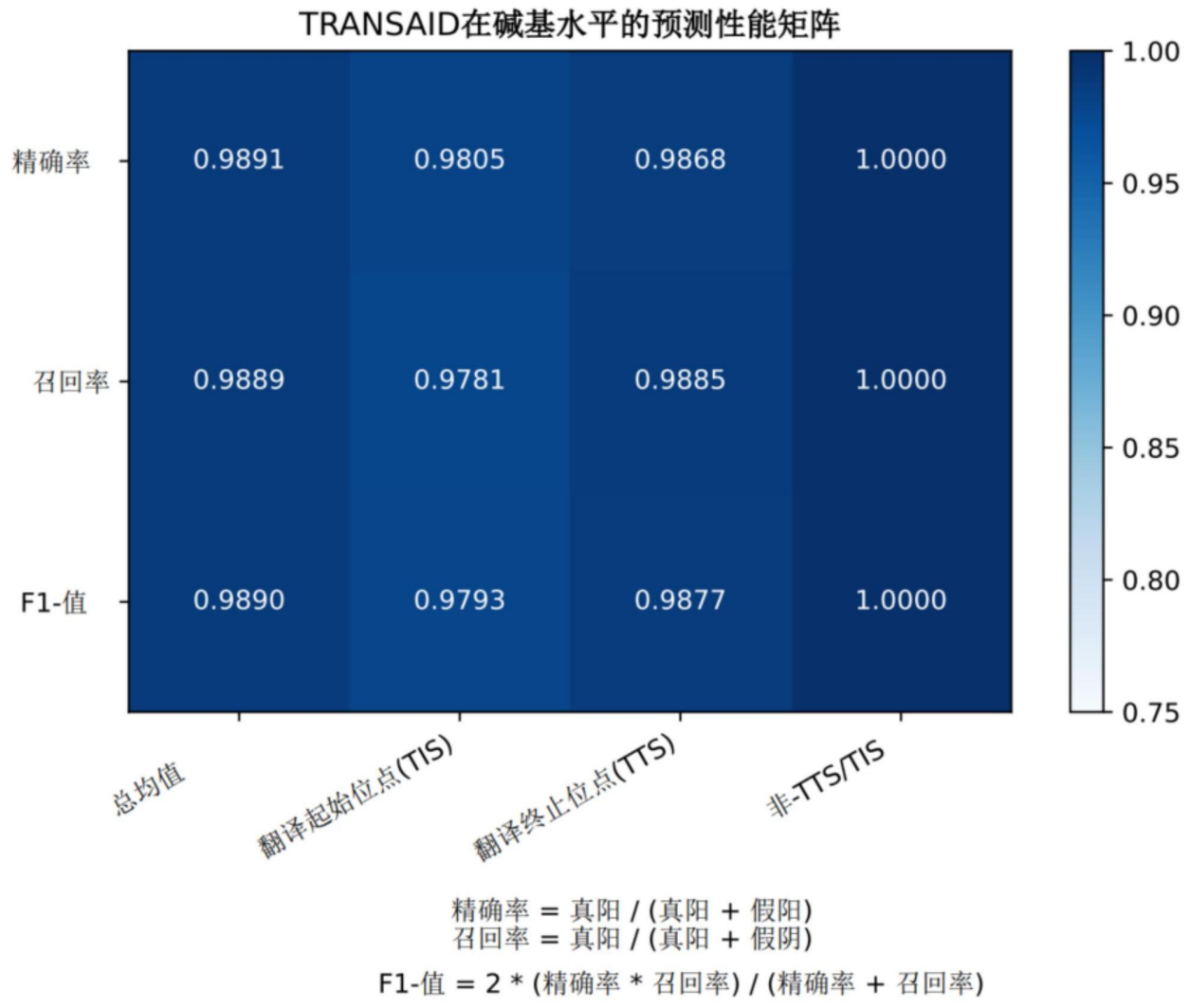
TRANSAID在碱基水平的预测性能矩阵
精确率
召回率
F1-值
0.9891
0.9805
0.9868
1.0000
0.9889
0.9781
0.9885
1.0000
0.9890
0.9793
0.9877
1.0000
1.00
0.95
0.90
0.85
0.80
0.75
总均值
翻译起始位点(TIS)
翻译终止位点(TTS)
非-TTS/TIS
精确率 = 真阳 / (真阳 + 假阳)
召回率 = 真阳 / (真阳 + 假阴)
F1-值 = 2 * (精确率 * 召回率) / (精确率 + 召回率)

图3

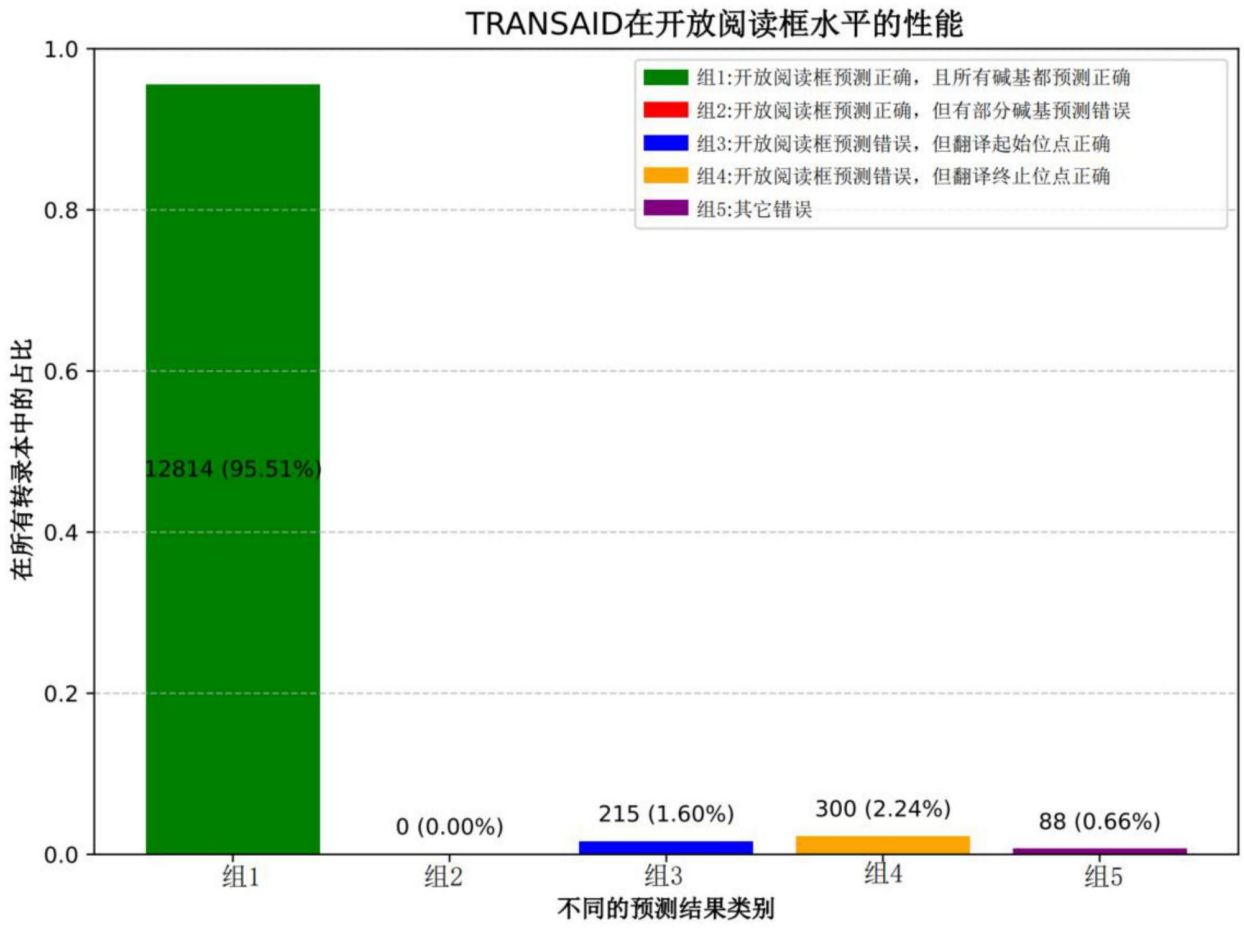
TRANSAID在开放阅读框水平的性能
组1:开放阅读框预测正确，且所有碱基都预测正确
组2:开放阅读框预测正确，但有部分碱基预测错误
组3:开放阅读框预测错误，但翻译起始位点正确
组4:开放阅读框预测错误，但翻译终止位点正确
组5:其它错误
在所有转录本中的占比
1.0
0.8
0.6
0.4
0.2
0.0
12814 (95.51%)
0 (0.00%)
215 (1.60%)
300 (2.24%)
88 (0.66%)
组1
组2
组3
组4
组5
不同的预测结果类别

图4

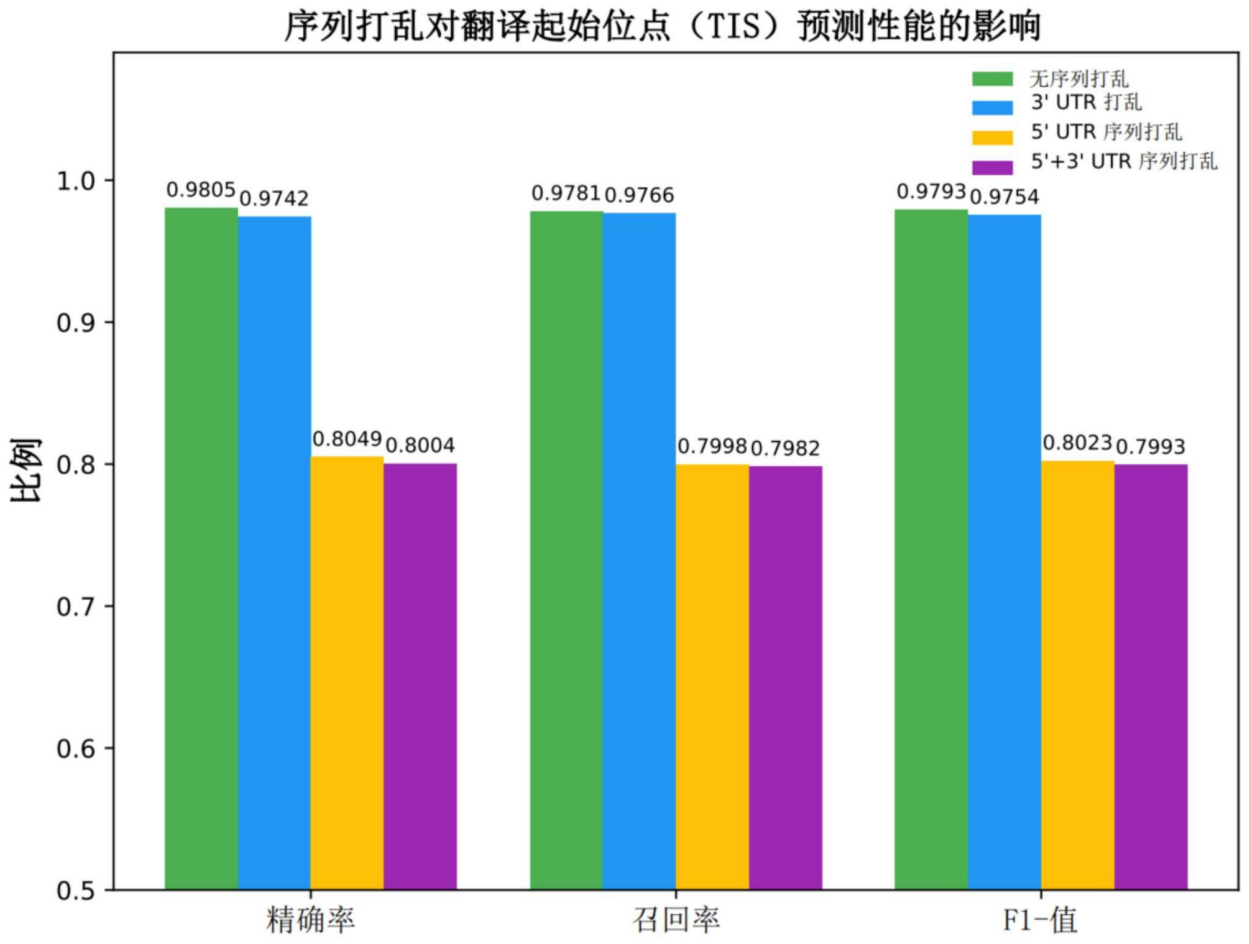
序列打乱对翻译起始位点（TIS）预测性能的影响
无序列打乱
3' UTR 打乱
5' UTR 序列打乱
5'+3' UTR 序列打乱
比例
1.0
0.9
0.8
0.7
0.6
0.5
0.9805
0.9742
0.8049
0.8004
0.9781
0.9766
0.7998
0.7982
0.9793
0.9754
0.8023
0.7993
精确率
召回率
F1-值

图5

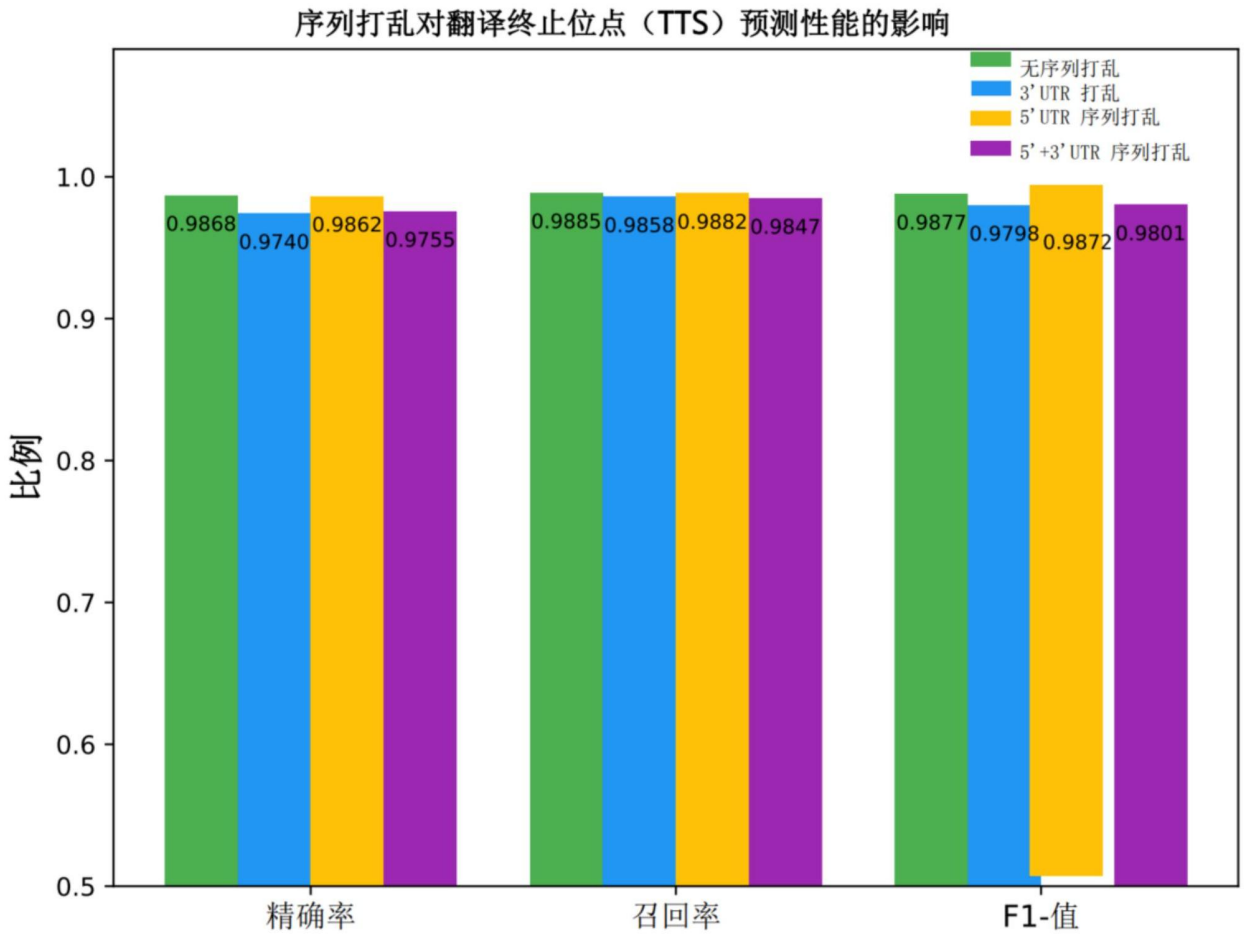
序列打乱对翻译终止位点（TTS）预测性能的影响
无序列打乱
3'UTR 打乱
5'UTR 序列打乱
5'+3'UTR 序列打乱
比例
1.0
0.9
0.8
0.7
0.6
0.5
0.9868
0.9740
0.9862
0.9755
0.9885
0.9858
0.9882
0.9847
0.9877
0.9798
0.9872
0.9801
精确率
召回率
F1-值

图6

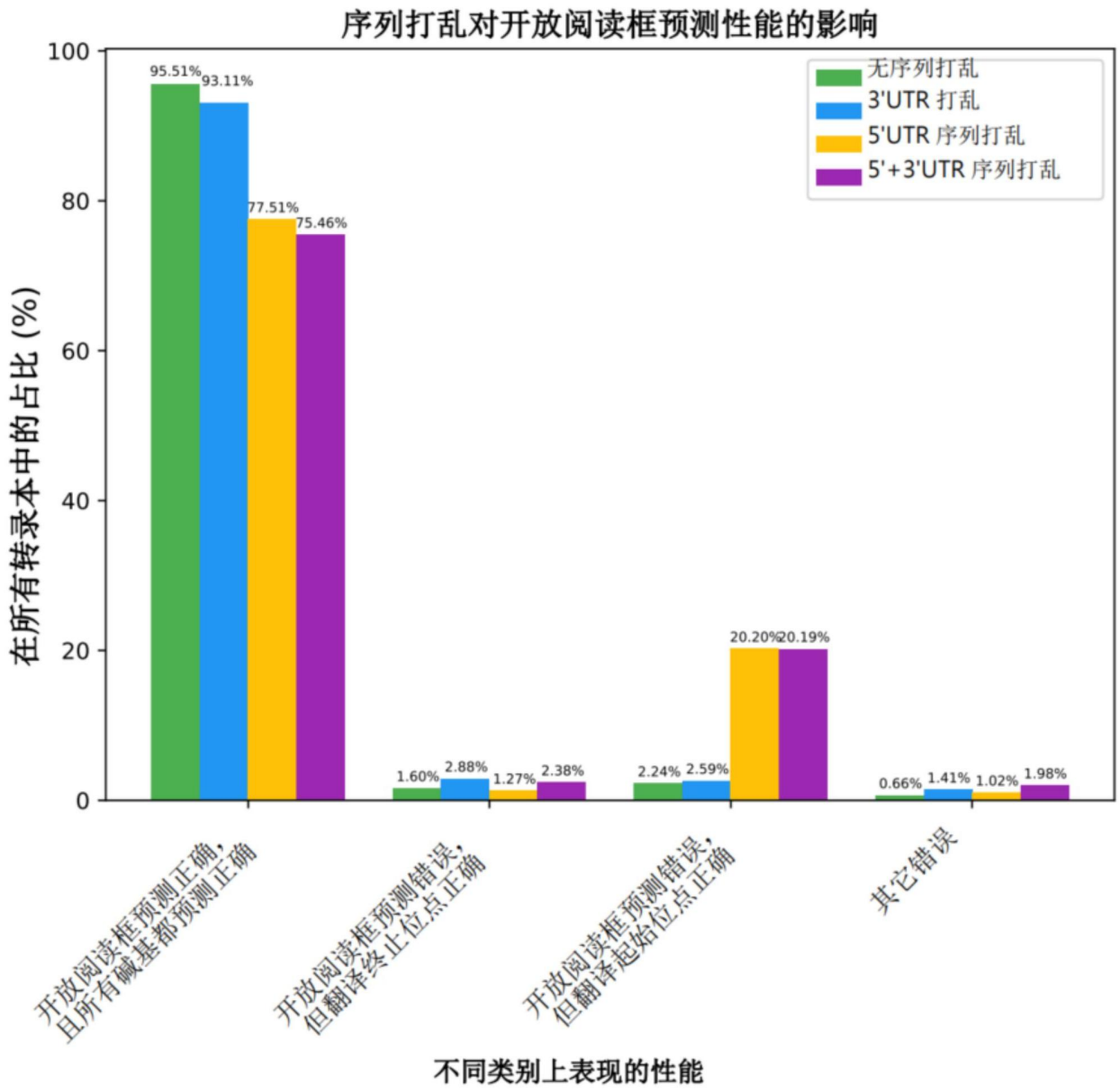
序列打乱对开放阅读框预测性能的影响
无序列打乱
3'UTR 打乱
5'UTR 序列打乱
5'+3'UTR 序列打乱
在所有转录本中的占比 (%)
100
80
60
40
20
0
95.51%
93.11%
77.51%
75.46%
1.60%
2.88%
1.27%
2.38%
2.24%
2.59%
20.20%
20.19%
0.66%
1.41%
1.02%
1.98%
开放阅读框预测正确，且所有碱基都预测正确
开放阅读框预测错误，但翻译终止位点正确
开放阅读框预测错误，但翻译起始位点正确
其它错误
不同类别上表现的性能

图7

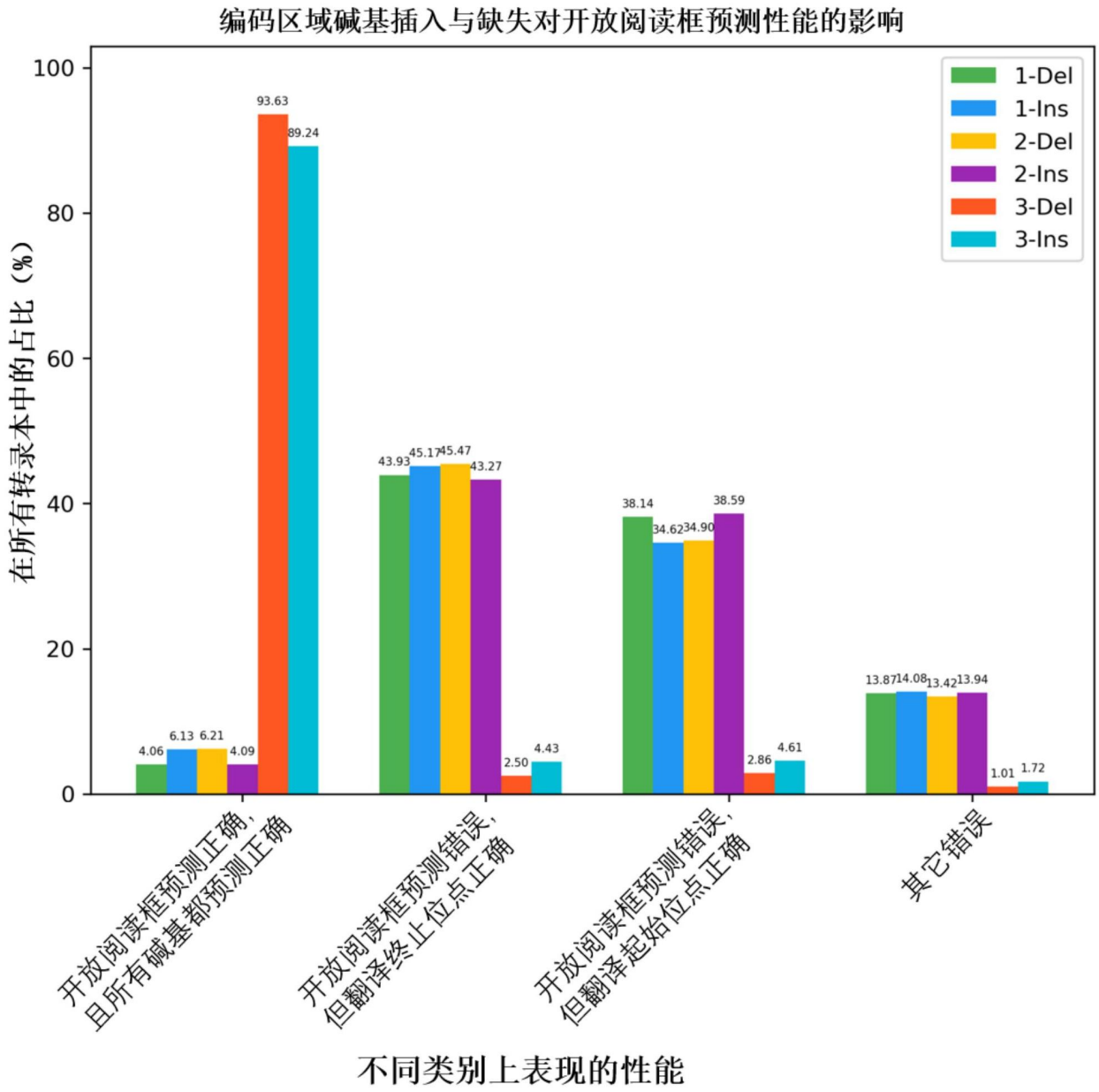
编码区域碱基插入与缺失对开放阅读框预测性能的影响
在所有转录本中的占比（%）
100
80
60
40
20
0
1-Del
1-Ins
2-Del
2-Ins
3-Del
3-Ins
4.06
6.13
6.21
4.09
93.63
89.24
43.93
45.17
45.47
43.27
2.50
4.43
38.14
34.62
34.90
38.59
2.86
4.61
13.87
14.08
13.42
13.94
1.01
1.72
开放阅读框预测正确，且所有碱基都预测正确
开放阅读框预测错误，但翻译终止位点正确
开放阅读框预测错误，但翻译起始位点正确
其它错误
不同类别上表现的性能

图8

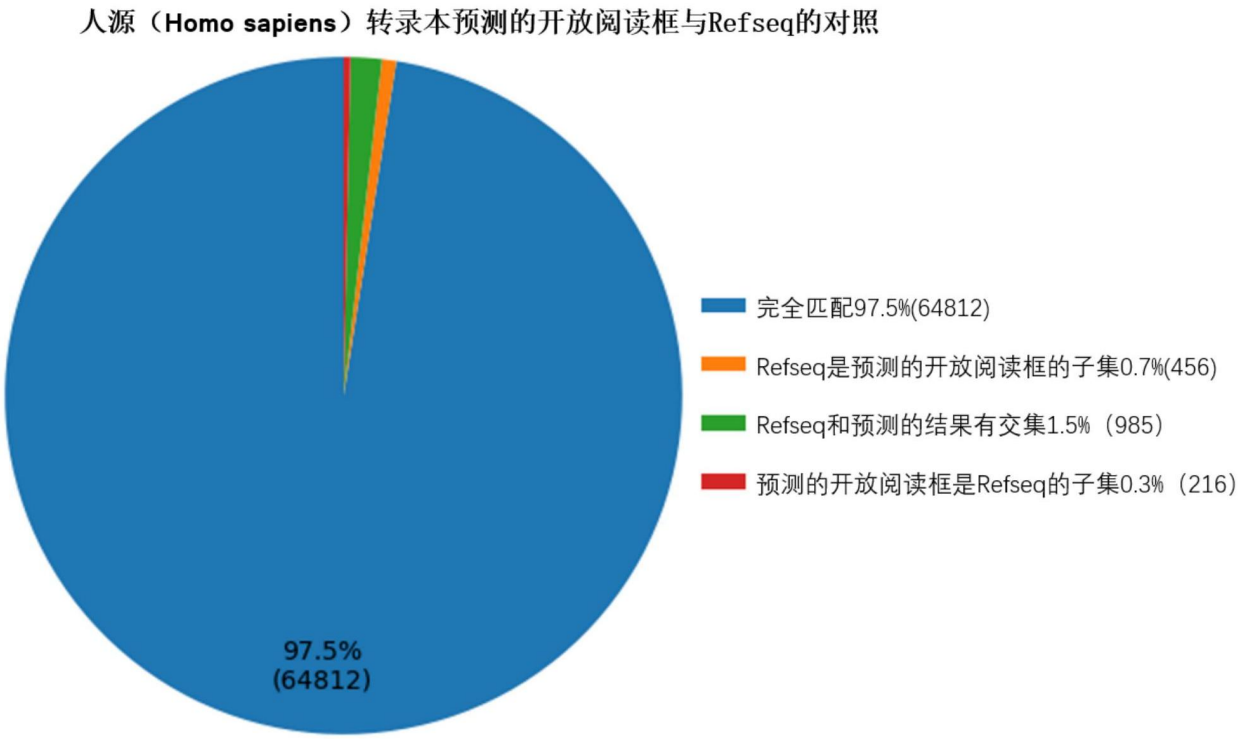
人源（Homo sapiens）转录本预测的开放阅读框与Refseq的对照
97.5%
(64812)
完全匹配97.5%(64812)
Refseq是预测的开放阅读框的子集0.7%(456)
Refseq和预测的结果有交集1.5%（985）
预测的开放阅读框是Refseq的子集0.3%（216）

图9

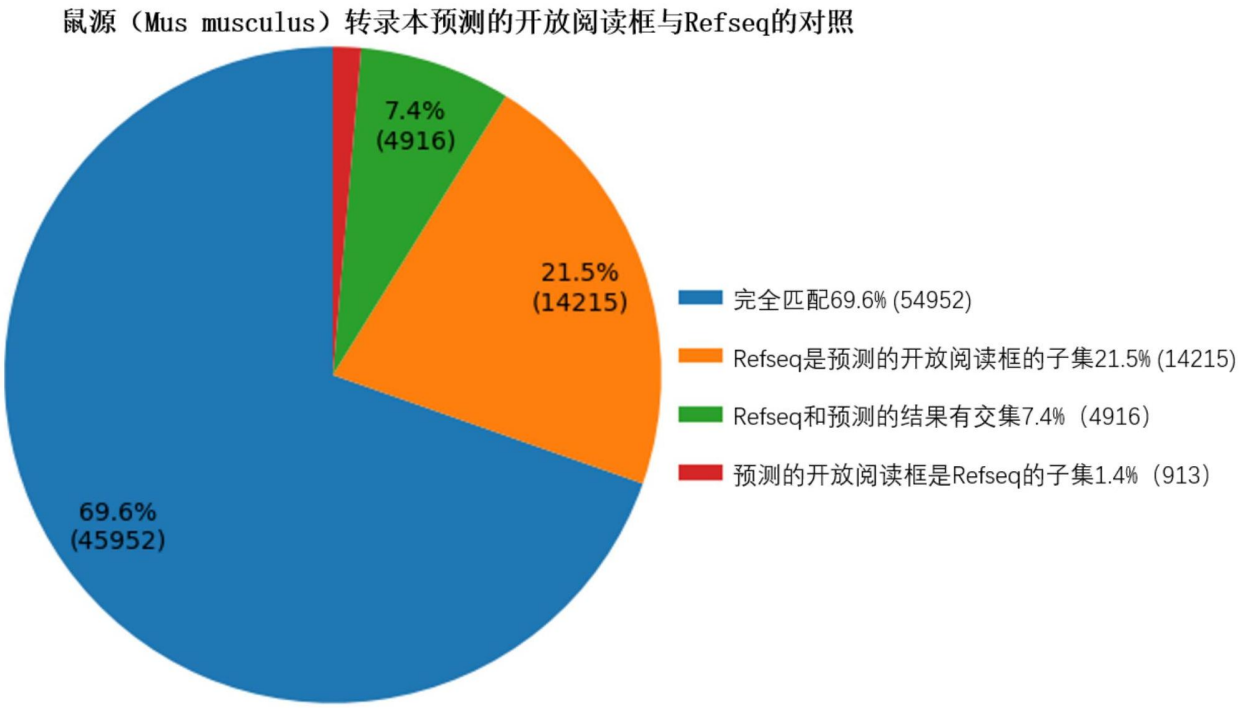
鼠源（Mus musculus）转录本预测的开放阅读框与Refseq的对照
7.4%
(4916)
21.5%
(14215)
69.6%
(45952)
完全匹配69.6% (54952)
Refseq是预测的开放阅读框的子集21.5% (14215)
Refseq和预测的结果有交集7.4%（4916）
预测的开放阅读框是Refseq的子集1.4%（913）

图10