计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(7): 1712−1725, 2026

基于 Evo2 模型微调的主粮作物基因组序列生成模型构建与应用

何曼曼[1] 刘龙正[1] 丁开锋[1] 葛明亮[1] 张雨琪[2] 任澳雨[2] 郭伟龙[2] 王耀君[1]

[1](中国农业大学信息与电气工程学院 北京 100083)

[2](中国农业大学农学院 北京 100193)

(wangyaojun@cau.edu.cn)

Construction and Application of Fine-Tuned Evo2-Based Models for Genomic Sequence Generation Models in Major Staple Crops

He Manman[1], Liu Longzheng[1], Ding Kaifeng[1], Ge Mingliang[1], Zhang Yuqi[2], Ren Aoyu[2], Guo Weilong[2], and Wang Yaojun[1]

[1] (*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083)

[2] (*College of Agronomy and Biotechnology*, *China Agricultural University*, *Beijing* 100193)

Abstract To construct genomic functional-element sequence generation models for major staple crops and evaluate their cross-species generalization ability, this study investigates fine-tuning and performance assessment of the Evo2 genomic language model for conditional sequence generation. Six representative crop genomes, including potato, peanut, indica rice cultivars (MH63 and ZS97), japonica rice, and wheat, are selected. A unified conditional generation and evaluation framework is established for six core genomic functional elements, including coding sequences (CDS), exons, introns, mRNA, 5′ untranslated regions (5′ UTR), and 3′ untranslated regions (3′ UTR). Based on this framework, Evo2 models are fine-tuned on crop-specific genomic data, while the pretrained Evo2 model without fine-tuning is retained as a zero-shot baseline. Experimental results show that fine-tuning consistently improves generation performance across all functional elements and crop species, an average absolute gain of about 0.49 percentage points, corresponding to a relative gain of approximately 0.79%. Cross-species evaluation further demonstrates that the fine-tuned models maintain good generalization ability when applied to unseen crop species, and the cross-species generalization gap remains below 1 percentage point for all six functional elements. Among them, coding-related elements such as CDS and mRNA show the most stable generation performance across species, whereas 5′ UTR is more sensitive to species differences. Additional bioinformatics analyses indicate that the generated sequences exhibit some degree of biological plausibility in ORF continuity, but still show limitations in codon preference, regulatory motifs in UTRs, and splice-site rules in introns. Overall, the results verify the effectiveness and cross-species applicability of Evo2-based fine-tuned genomic language models for functional-element sequence generation in major staple crops, providing a potential technical basis for genome-assisted breeding and crop improvement.

Key words genomic language model; transfer learning; crop genomics; sequence generation; Evo2; major staple crops

收稿日期: 2025-12-31; **修回日期**: 2026-05-20

基金项目: 国家自然科学基金联合基金项目(U25B2019-1); 中央高校建设世界一流大学(学科)和特色发展引导专项资金(2025AC030)

The work was supported by the Joint Fund of the National Natural Science Foundation of China (U25B2019-1), and the Guiding Special Fund for Central Universities to Build World-Class Universities (Disciplines) and Promote Characteristic Development (2025AC030).

通信作者: 王耀君(wangyaojun@cau.edu.cn)

摘　要　为构建适用于主粮作物的基因组功能元件序列生成模型，并评估其跨物种泛化能力，以Evo2基因组语言模型为基础，针对多种主粮作物开展微调与生成性能评测研究。选取马铃薯、花生、籼稻（MH63、ZS97）、粳稻和小麦6个作物品种，围绕编码序列（CDS）、外显子、内含子、mRNA、5′非翻译区（5′ UTR）和3′非翻译区（3′ UTR）六类核心基因组功能元件，构建统一的条件生成与评测体系。实验结果表明，经作物数据微调后，Evo2模型在所有功能元件上的生成性能均较零样本基线获得稳定提升，平均序列相似度绝对提升约0.49个百分点，对应相对提升约0.79%，表明微调能够有效增强模型对作物基因组序列特征的适配能力。进一步的跨物种测试显示，模型在不同作物之间仍保持良好的泛化性能，各类功能元件的跨物种泛化差距均小于1个百分点。其中，编码相关元件（CDS和mRNA）生成最为稳定，而5′ UTR对物种差异相对更为敏感。综上，验证了基于Evo2微调的基因组语言模型在主粮作物多功能元件序列生成任务中的有效性与跨物种适用性。补充生物信息学分析进一步表明，模型在ORF连续性等方面具有一定合理性，但在密码子偏好、UTR调控基序及内含子剪接规则等精细生物学特征上仍存在改进空间。

关键词　基因组语言模型；迁移学习；作物基因组学；序列生成；Evo2；主粮作物

中图法分类号　TP311.13；TP309

DOI: 10.7544/issn1000-1239.202660003　　**CSTR:** 32373.14.issn1000-1239.202660003

随着高通量测序技术的飞速发展，作物基因组数据呈爆炸式增长，为理解生命规律和指导分子育种提供了前所未有的资源[1]。如何利用这些海量数据，通过计算方法生成、分析与设计具有特定功能的基因组序列，已成为计算生物学和智慧农业交叉领域的前沿挑战[2]。传统序列分析方法，如基于比对的同源搜索和基于隐马尔可夫模型的基因预测，严重依赖于手工特征和特定物种的先验知识，泛化能力和适应性有限[3]。

近年来，基于Transformer架构的大规模语言模型在自然语言处理领域取得突破性进展，其强大的序列建模能力被成功引入基因组学，催生了“基因组语言模型”[4-6]。这类模型通过对数十亿至数千亿碱基的基因组数据进行自监督预训练，能够深刻理解DNA序列的语法、语义乃至进化规律[7-8]。其中，Evo2模型作为新一代基因组语言模型的代表，在多项序列理解和生成任务上展现了卓越的性能[9-12]。然而，现有研究多集中于人类或模式生物，关于此类模型在主粮作物（如水稻、小麦、花生）中系统性应用的研究仍属空白。此外，基因组包含功能各异的元件（如编码区、调控区），现有工作往往未加以区分，缺乏对模型在不同类型元件上生成能力的深入分析[13-14]。

针对上述问题，本文提出一种面向主粮作物的基因组语言模型迁移学习与应用框架。本研究以Evo2为基础模型，在花生、珍汕97（ZS97）、明恢63（MH63）、粳稻日本晴、小麦和马铃薯6个作物品种上，围绕CDS、外显子、内含子、mRNA、5′ UTR和3′ UTR六类核心基因组元件开展系统微调与评估。本文的主要贡献如下：

1）构建了覆盖多主粮作物、多功能元件类型的基因组序列生成评估体系，并系统考察了模型规模、训练设置与元件类型对迁移学习效果的影响。

2）建立了作物专属微调与跨物种迁移相结合的实验框架，系统分析了模型在同物种与跨物种场景下的生成一致性与稳定性差异。

3）揭示了基因组语言模型在不同功能元件上的生成规律，发现编码相关元件整体更稳定，而非编码调控元件对物种差异更为敏感，为后续优化提供了依据。

4）在序列相似性评价基础上，进一步引入ORF完整性、密码子分布、UTR调控基序和内含子剪接规则等补充性生物信息学分析，从多维角度评估生成序列的合理性。

本文工作为人工智能大模型在作物基因组研究中的应用提供了新的技术路径，对作物遗传改良和功能基因挖掘具有一定参考价值。

1　相关工作

1.1　基因组语言模型

基因组语言模型借鉴自然语言处理的思想，将DNA碱基（A、T、C、G）或k-mer作为基本token进行建模。早期工作如DNABERT[15]采用BERT架构进行掩码语言建模预训练，用于启动子预测、染色质可及性预测等下游任务。Nucleotide Transformer[16]进一步沿用编码器/掩码语言建模范式，构建了面向基因

组序列表示学习的基础模型；而 Evo[13]等自回归生成模型则被用于基因组序列建模与生成任务。Evo2[14]作为最新进展，通过更大规模的预训练数据和改进的模型架构，显著提升了长序列理解和生成的准确性。这些模型虽展现了强大潜力，但其在作物基因组学，特别是跨作物应用中的效能尚未得到充分探索。

1.2 计算生物学中的迁移学习

迁移学习通过将在大规模源域（如多物种混合基因组）上学到的知识迁移到目标域（如特定作物），以解决目标域数据稀缺问题[17]。在生物序列分析中，迁移学习已被用于蛋白质功能预测和基因表达推断[18]。然而，在基因组序列生成任务中，如何设计有效的迁移策略以适应作物间巨大的遗传差异，仍是一个开放性问题[19-20]。本文探索了全参数微调的迁移路径，旨在实现从通用模型到作物专用生成器的平滑过渡。

1.3 作物基因组序列分析

传统作物基因组分析工具如 BLAST、GMAP 等依赖于序列比对，不适用于从头（de novo）序列的生成。近年来，深度学习和基因组语言模型被广泛用于基因组功能预测、调控信息挖掘和作物基因组辅助育种等任务[21-23]。在生成方面，生成对抗网络（generative adversarial network, GAN）和变分自编码器（variational auto-encoders, VAE）等生成式模型也曾被用于生物序列生成探索，但在长序列连贯性和功能约束保持方面仍存在不足。本文系统地将大规模基因组语言模型应用于多种作物完整基因元件的生成，并进行了多维度评估。

2 预备知识

2.1 Evo2 模型概述

Evo2[14]（Evolutionary Generative Model 2）是由美国 Arc Institute 与英伟达（NVIDIA）合作发布的新一代基因组尺度语言模型，相关模型能力也已通过 NVIDIA BioNeMo 平台开放[24]。该模型基于改进的 StripedHyena 架构，融合注意力机制，支持从单核苷酸到基因组级别的长序列建模与生成。Evo2 在包含细菌、真核生物和古细菌的多物种基因组数据集上进行预训练，数据规模达 9.3 万亿个核苷酸，上下文长度可支持 100 万个碱基，是目前规模最大、覆盖最广的基因组基础模型之一。

在多项生物序列理解与生成基准测试中，Evo2 展现出优异的跨物种泛化能力和长程依赖建模能力，尤其在保守区域识别、功能元件预测及进化模式捕获等方面表现突出。然而，现有研究多聚焦于模式生物或微生物基因组，其在具有复杂农艺性状、驯化历史和大量结构变异的主粮作物基因组中的应用尚未得到系统探索。作物基因组具有更高的重复序列比例、物种特异性调控网络及育种相关功能模块，这对预训练模型的适应性、领域泛化与功能可解释性提出了更高要求。

因此，本文选择 Evo2 作为基础模型，系统研究其在不同主粮作物、不同功能基因组元件上的迁移学习特性与序列生成能力，旨在填补该模型在农业基因组学场景中的研究空白，并为作物智能设计提供可靠的算法基础。

具体而言，Evo2 采用自回归生成范式。给定一个 DNA 序列 $X=(x_1,x_2,\cdots,x_n)$，其中 $x_i\in\{\mathrm{A,T,G,C}\}$，模型通过最大化序列的似然 $P(X)=\prod_{i=1}^{n}P(x_i|x_{<i};\Theta)$ 进行预训练，其中 Θ 为模型参数。该预训练过程使模型获得对 DNA 序列“语言”的通用理解能力，从而可通过下游微调快速适应特定生成任务。本文使用 Evo2 的 1B（十亿）与 7B（七十亿）参数版本，采用主粮作物基因组序列数据集对模型进行微调，增强其在主粮作物基因组学中的应用效果。

2.2 问题定义：基于语言模型的作物基因组序列生成

本文将作物基因组序列生成定义为条件生成任务。给定一个目标作物物种 S、一个基因组元件类型 M（如“CDS”）、一个简短的上下文提示序列 c（如前 200 个碱基）以及目标生成长度 L，目标是生成符合该物种和元件类型生物学特性的后续序列 $\hat{y}$，使得 $\hat{y}$ 与真实参考序列 y 在序列特征和潜在功能上尽可能接近。通过构建有标注的海量作物基因组数据，进一步对模型微调学习条件分布 $P(y|c,S,M;\Theta)$，从而得到一个可以生成目标基因组元件对应序列的模型。

2.3 基因组元件类型

基因组由功能各异的元件组成，其序列特征和进化约束差异显著。表 1 详细列出了本研究所涉及的 6 类核心元件的生物学定义与功能，理解这些差异是训练及分析模型生成行为的基础。

3 基于 Evo2 微调的主粮作物基因组学生成方法

为构建适用于主粮作物场景的基因组功能元件序列生成模型，本文在 Evo2 基因组语言模型的基础

Table 1 Description of Genomic Element Types

表 1 基因组元件类型说明

元件名称	英文简称	整合描述（功能与特征）
编码序列	CDS	编码蛋白质的 DNA 区域，起止于起始与终止密码子。直接决定氨基酸序列，受密码子偏好和强纯化选择约束，序列高度保守。
外显子	Exon	基因中构成成熟 mRNA 的片段，含 CDS 和部分 UTR。序列相对保守，剪切位点附近具有特异信号。
内含子	Intron	基因中不编码蛋白质的间隔序列，转录后被剪接去除。序列长度与碱基组成变异大、进化快，但保留剪切必需保守位点（如 GT-AG 规则）。
信使 RNA	mRNA	成熟转录本，包含 5′ UTR、CDS 和 3′ UTR。作为蛋白质合成模板，由外显子连接形成，承载完整编码与调控信息。
5′ 非翻译区	5′ UTR	mRNA 起始密码子上游非编码区，调控翻译起始、mRNA 稳定性与定位。通常较短，可形成特定二级结构影响翻译效率。
3′ 非翻译区	3′ UTR	mRNA 终止密码子下游非编码区，参与 mRNA 稳定性、翻译效率与亚细胞定位调控。常含 miRNA 结合位点与 poly(A) 加尾信号。

上，采用全参数微调与迁移学习策略，对模型进行作物层面与功能元件层面的适配。本节围绕作物基因组序列生成任务建模、模型微调与迁移方式，以及序列生成与性能评估方法，对所采用的生成方法进行说明，为后续实验设计与结果分析提供统一的方法依据。

3.1 作物基因组序列生成任务建模

基于第 2 节中对 DNA 序列表示与建模方式的定义，本文将主粮作物基因组功能元件序列生成建模为条件自回归生成问题[15]。在训练与推理过程中，模型通过学习由不同作物与功能元件类型划分的序列样本分布，在给定真实序列起始部分作为提示的条件下，以自回归方式逐碱基生成后续 DNA 序列，使生成结果在整体结构和统计特征上尽可能接近真实参考序列[16]。

该建模方式能够统一描述 CDS、外显子（Exon）、内含子（Intron）、mRNA、5′非翻译区（5′ UTR）与 3′非翻译区（3′ UTR）等不同类型功能元件的生成过程，并兼顾不同主粮作物在序列长度分布、碱基组成及功能约束方面的差异。通过在多作物、多元件数据上的训练与评估，模型生成能力可以在不同作物与元件之间进行系统对比分析。

3.2 基于 Evo2 的全参数微调与迁移学习

Evo2 是在大规模多物种基因组序列上预训练的自回归基因组语言模型，能够学习 DNA 序列在不同生物尺度上的统计规律与长程依赖结构。然而，其预训练目标主要侧重于跨物种的通用序列分布建模，尚不能直接适配具体主粮作物及不同基因组功能元件在序列组成、长度分布和进化约束方面的差异。

针对上述问题，本文采用基于全参数微调的迁移学习策略，在保持 Evo2 原有模型结构不变的前提下，利用主粮作物的基因组功能元件数据对模型参数进行进一步优化，使模型能够更准确地刻画目标作物及功能元件的条件生成分布[17-19]。

在微调过程中，本文 Evo2 官方发布的预训练模型作为初始化权重，分别在不同作物数据集上构建物种专属微调模型。每个训练样本由功能元件类型与作物来源隐式约束，并由前缀提示序列（prompt）和目标生成序列（target）组成。模型通过最小化目标序列在给定提示条件下的负对数似然损失，学习作物特异性的碱基组成特征、局部结构模式及功能相关约束。

为系统评估微调策略的有效性，本文在实验设计中同时保留未经任何作物数据微调的 Evo2 预训练模型作为零样本基线模型。通过将微调模型与零样本模型在相同测试集和生成设置下的结果进行对比，可以定量分析微调对生成序列一致性的影响程度，从而验证迁移学习在作物基因组场景中的必要性。

在此基础上，本文进一步构建了跨物种迁移评估框架。具体而言，模型在某一作物数据上完成微调后，被直接应用于其他作物的功能元件序列生成任务，而不再进行额外参数更新。通过比较同物种测试与跨物种测试条件下生成序列与真实参考序列之间的相似性差异，可以评估模型所学习到的序列表示在不同作物间的可迁移性与泛化能力。

此外，考虑到模型容量可能对迁移学习效果产生影响，本文选取 Evo2 的 1B 与 7B 两种参数规模版本，在相同数据划分和训练策略下分别进行微调与评估。通过对不同模型规模在各类功能元件上的生成一致性进行对比分析，可以揭示模型规模变化对作物基因组序列生成精度与稳定性的影响规律。

需要强调的是，本文在方法层面并不假设更大模型一定带来更优结果，而是通过统一的实验设置，对模型微调效果、跨物种迁移能力以及模型规模影响进行客观量化分析。所有模型均采用一致的序列生成策略与评价指标，其具体定义与统计口径在 4.1 节中给出。

3.3 序列生成与相似性指标评估方法

在序列生成阶段，模型在测试集上采用统一的生成设置进行推理，包括一致的提示序列长度、生成长度及采样策略，以保证不同模型和不同实验结果之间具有可比性。生成序列与对应的真实参考序列一一配对，用于后续性能评估与分析。

在性能评估方面，本文将生成序列与真实参考序列之间的平均全局序列相似度作为主评价指标。具体做法是对生成序列与参考序列进行全局比对，计算匹配碱基数占总长度的比例，并在测试集样本上取平均。该指标能够较直观地反映生成序列在整体结构层面与真实序列的接近程度，因此作为全文统一使用的主要性能指标。

为补充分析模型生成结果的可靠性与生物学合理性，本文还引入 2 类辅助分析。第 1 类为基于对数概率的内部一致性指标，用于比较不同模型规模和迁移策略下生成行为的稳定性。第 2 类为面向不同功能元件的补充性生物信息学一致性分析，包括对 CDS 的开放阅读框完整性、最长 ORF 比例及密码子使用分布进行评估，对 5′ UTR 和 3′ UTR 分别统计 Kozak 核心/完整模式与 Poly(A) 信号命中情况，及内含子序列分析供体端 GT、受体端 AG 及 GT-AG 剪接规则的符合度。需要说明的是，这些指标的适用对象和统计属性与全文统一使用的连续型序列相似度指标并不完全一致，因此不纳入主要性能数值计算，而作为独立的辅助分析结果在实验部分报告。

4 实验设计与结果分析

4.1 实验设置

4.1.1 数据集构建

本文实验基于 6 种主粮作物的参考基因组数据开展，包括马铃薯、花生、籼稻(MH63、ZS97)、粳稻以及小麦。各数据集所采用的基因组版本、数据来源及最终用于实验的有效样本规模统计如表 2 所示。

Table 2 Basic Statistics of Experimental Datasets

表 2 实验数据集基本信息统计

数据集	物种类型	基因组版本（Assembly）	来源	用于实验的有效样本条数
MH63	籼稻	MH63RS3[25]	CNGB/Ensembl	12
ZS97	籼稻	ZS97RS3[25]	CNGB/Ensembl	12
Nipponbare	粳稻	IRGSP-1.0[26]	Ensembl Plants	12
Peanut	花生	Tifrunner[27]	Ensembl Plants	20
Wheat	小麦	IWGSC RefSeq v1.0[27]	Ensembl Plants	21
Potato	马铃薯	SolTub_3.0[27]	Ensembl Plants	12

依据参考基因组 FASTA 文件及对应的 GFF3 注释信息[22-23]，从每个基因中提取 CDS、外显子、内含子、mRNA、5′ UTR 和 3′ UTR 六类基因组功能元件。对于由多个片段构成的 CDS 与 UTR 序列，按照基因组坐标顺序进行拼接，并统一根据链方向进行反向互补校正；内含子序列由相邻外显子之间的区间推断获得。

为适配基因组语言模型的条件生成任务，对完整元件序列进一步划分为前缀提示序列与目标生成序列。不同类型元件采用差异化长度设置，以兼顾生物学合理性与模型训练稳定性。仅保留长度满足实验要求的序列样本，并在此基础上对每个作物数据集随机抽取固定规模样本用于模型训练与评测。不同作物在基因组规模及结构复杂度上的差异反映在各类元件的样本规模与长度分布中(见表 2)，为后续跨物种迁移学习实验提供了具有代表性的对照数据基础。

4.1.2 模型配置与训练策略

本文选用 Evo2 基因组语言模型作为基础模型，分别采用 1B 与 7B 参数规模版本，以分析模型规模对基因组序列生成性能的影响。针对每个作物数据集，分别在对应数据上对 Evo2 模型进行微调，获得物种专属模型；同时保留未经微调的 Evo2 预训练模型作为零样本对照。为尽量降低外部干扰，所有模型均使用统一的数据预处理流程、数据划分规则、损失函数定义与推理设置。

模型训练前，对各作物数据集中的序列样本分别划分为训练集、验证集与测试集。不同模型在训练与推理阶段均采用一致的序列长度设置与生成策略，以保证实验结果之间的可比性。除特别说明外，主要结果在固定数据划分和固定随机种子条件下进行单次微调实验，最佳模型依据验证集负对数似然表现进行选择。为保证对比实验的公平性，CVAE 与 HyenaDNA 基线模型也在相同数据划分、相同输入输出设置和相同评价标准下完成训练与测试。

4.1.3 模型微调方法

本文采用 Evo2 基因组语言模型对不同作物基因组数据集进行微调训练。微调过程中以 Evo2 官方发布的预训练模型作为初始化权重，在保持模型网络结构不变的前提下，对模型参数进行全参数微调，以适配特定作物及功能元件的序列分布特征。

Evo2 模型基于自回归语言建模框架，对基因组序列进行逐碱基建模。在微调阶段，每个训练样本由前缀提示序列 P 与目标生成序列 $A=(t_1,t_2,\cdots,t_L)$ 组成。设模型参数为 θ，则模型在给定前缀条件下对目标序列的条件概率定义为：

$$p(A|P;\theta)=\prod_{i=1}^{L}p(t_i|P,t_{<i};\theta)\ ,$$

其中，$t_{<i}=(t_1,t_2,\cdots,t_{i-1})$。

模型微调的优化目标为最小化训练数据集上的平均负对数似然损失函数：

$$\mathcal{L}(\theta)=-\frac{1}{|D|}\sum_{(P,A)\in D}\log_p(A|P;\theta)\ ,$$

其中 D 表示训练样本集合，$|D|$为训练样本数量。

针对不同模型规模，分别采用分阶段学习率调度策略，并结合梯度裁剪与权重衰减机制，以提升训练稳定性并降低过拟合风险。

2 种参数规模模型均采用全参数微调策略，训练过程中使用 Warmup→Constant→Cosine 学习率调度，并结合权重衰减、梯度裁剪和 BF16 混合精度训练以提升训练稳定性。所有实验均在 H800 GPU 环境下完成，具体超参数设置见表 3。

Table 3　Fine-Tuning Hyperparameters Settings

表 3　微调超参数设置

参数项	Evo2-1B	Evo2-7B
峰值学习率	1.5×10^{-5}	2.5×10^{-5}
最小学习率	1.5×10^{-6}	2.5×10^{-6}
学习率调度	Warmup → Constant → Cosine	Warmup → Constant → Cosine
Warmup 步数	200	200
Constant 阶段占比	约 70%	约 70%
最大训练步数	10 000	28 000
验证间隔	1 000 steps	2 000 steps
微批大小	4	1
输入序列长度	8 192 tokens	8 192 tokens
权重衰减系数	0.001	0.001
梯度裁剪阈值	0.5	0.5
Dropout（隐藏层/注意力）	0.02/0.02	0.02/0.02
张量并行数	1	2
激活检查点层数	1	16
混合精度训练	BF16	BF16
随机种子	12 342	12 342
GPU 配置	1 × H800	4 × H800

4.1.4　评价指标

模型生成性能从序列相似度与概率一致性 2 个层面进行评估。

1）序列相似度（*Similarity*）

设对生成序列与对应参考序列进行全局序列比对，在非缺失位置统计匹配碱基数 M，设目标序列长度为 L，则相似度定义为：

$$Similarity=\left(\frac{M}{L}\right)\times100\%\ 。$$

该指标反映模型在给定前缀提示序列条件下，对真实基因组序列局部结构的复现能力。

2）概率差分指标Δ

在给定前缀提示序列 P 的条件下，模型对参考序列与生成序列的平均对数概率分别记为$\log_p(A|P;\theta)$ 与$\log_p(G|P;\theta)$，概率差分定义为：

$$\Delta=\log_p(G|P;\theta)-\log_p(A|P;\theta)\ 。$$

该指标用于衡量模型内部概率空间中，生成序列相对于参考序列的偏离程度。当Δ接近 0 时，表明模型对生成序列与真实序列的概率评估较为一致；当Δ为较大负值时，说明生成序列在模型概率分布中偏离真实基因组分布较多。需要说明的是，Δ受模型校准状态、序列长度及解码策略等因素影响，因此本文将其作为辅助指标，用于从模型内部置信度角度补充分析生成结果的稳定性与一致性，而不作为全文统一主评价指标。

统计口径说明：文中涉及的“平均相似度”均采用两级算术平均方式计算：首先在每一类功能元件内部，对测试样本的相似度取算术平均；随后对 6 类功能元件的平均值再次取算术平均，得到整体平均相似度。

4.1.5　补充生物信息学一致性分析设置

本文进一步针对生成序列开展补充性生物信息学一致性分析，以评估模型在序列相似性之外对基本生物学规则的保持情况。分析对象包括 CDS、5′ UTR、3′ UTR 和内含子这 4 类功能元件。对于 CDS，重点考察开放阅读框相关指标，包括固定阅读框内无终止密码子的比例、严格完整 ORF 的比例、最长 ORF 长度占序列长度的平均比例，以及生成序列与真实序列在密码子使用分布上的差异。对于 5′ UTR 和 3′ UTR，分别统计 Kozak 核心/完整模式和 Poly(A) 信号的命中情况，以评估模型对典型调控基序的保留能力。对于内含子，则统计供体端 GT、受体端 AG 及完整 GT-AG 剪接规则的符合度。上述分析主要用于从生物学规则层面对生成结果进行补充说明，不参与全文统一主评价指标的统计。

4.1.6　实验设备与运行环境

本文所有实验均在统一的高性能计算服务器上完成。服务器配备 NVIDIA H800 GPU，不同模型规模采用不同 GPU 配置，其中 Evo2-1B 使用 1×H800，Evo2-7B 使用 4×H800，具体设置见表 3。同时配置多核 CPU 与大容量内存，用于支持大规模基因组序列的数据预处理、模型训练与推理任务。所有实验均

运行于 Linux 操作系统环境下, 模型训练与推理基于 Evo2 官方源代码框架完成。

4.1.7 实验重复性与确定性设置

为保证实验结果的可重复性, 本文在训练阶段采用固定数据划分和固定随机种子设置, 除特别说明外, 主要结果均来自单次微调实验中验证集负对数似然表现最优的检查点。

在实验开发过程中, 我们对 Evo2-1B 和 Evo2-7B 的主要配置进行了多次重复训练, 结果差异保持在可接受范围内, 表明方法具有较好的稳定性。

在推理阶段, 本文统一采用贪心解码, 不使用温度采样或束搜索, 因此在相同模型检查点和相同输入条件下, 评测结果可稳定复现。

4.2 实验 1: 与其他基线方法的对比分析

4.2.1 基线方法与实验设置

在主粮作物基因组功能元件序列生成任务中, 本文在与 Evo2 相同的数据集、数据划分、prompt/target 设置和评价指标下, 引入 2 类代表性基线方法进行对比实验, 共覆盖 6 个作物基因组数据集与 6 类基因组功能元件, 总计 36 个评测组合。

第 1 类基线为条件变分自编码器(CVAE)。该模型采用双层 GRU 编码器—解码器结构, 嵌入维度为 128, 隐藏层维度为 256, 潜变量维度为 128, 参数量约为 0.8×10^{6}。训练过程中采用 AdamW 优化器, 初始学习率为 3×10^{-4}, batch size 为 64, 训练 50 个 epoch, 并结合余弦退火学习率调度与早停策略。

第 2 类基线为预训练基因组语言模型 HyenaDNA。本文选用 HyenaDNA-small-32k-seqlen 作为基座模型, 并采用因果语言建模续写方式进行微调。为与本文方法保持一致, 训练时 prompt 部分不参与损失计算, 仅在 target 区域计算损失; 训练过程中采用 AdamW 优化器, 初始学习率为 1×10^{-4}, weight decay 为 0.01, batch size 为 16, 训练 10 个 epoch, 并结合 warmup 与余弦退火调度策略。

需要说明的是, 以上基线模型均在与 Evo2 完全一致的主粮作物基因组数据上训练, 并在统一测试集上按照相同评价指标进行评测。

4.2.2 对比结果与分析

为更全面评估 Evo2 在主粮作物基因组功能元件序列生成任务中的相对优势, 本文在相同数据集、相同输入输出设置和统一评价指标下, 引入 CVAE 和 HyenaDNA 两类基线方法进行对比。表 4 给出了 36 个评测组合上的平均结果。

从整体结果看, Evo2 在平均序列相似度指标上明显优于 2 类基线模型。其平均序列相似度达到 62.60%, 明显高于 HyenaDNA 的 39.83% 和 CVAE 的 38.46%, 说明 Evo2 在主粮作物基因组功能元件序列生成任务中具有更强的整体建模能力。

Table 4 Comparison of Average Similarity on Different Models

表 4 不同模型平均序列相似度对比 %

功能元件	Evo2	HyenaDNA	CVAE
CDS	62.87	39.01	37.85
外显子	61.43	43.31	41.17
内含子	63.94	38.18	37.50
mRNA	63.66	37.81	35.24
3′ UTR	62.88	39.99	39.86
5′ UTR	60.82	40.71	39.16
平均	62.60	39.83	38.46

从不同功能元件来看, Evo2 在 6 类元件上均保持领先, 没有出现某一类元件被基线模型反超的情况。其中, mRNA、内含子和 CDS 上的优势最为明显。以序列相似度为例, Evo2 在 mRNA 上分别比 HyenaDNA 和 CVAE 高 25.85 个百分点和 28.42 个百分点, 在内含子上分别高 25.77 个百分点和 26.44 个百分点, 在 CDS 上分别高 23.86 个百分点和 25.02 个百分点。这说明在较长序列或结构更复杂的功能元件上, Evo2 对长程依赖和条件生成分布的刻画更充分。

相比之下, 外显子和 5′ UTR 虽然同样由 Evo2 取得最优结果, 但相对优势略小。尤其是外显子上, Evo2 相比 HyenaDNA 提升为 18.13 个百分点, 相比 CVAE 提升 20.26 个百分点, 低于 mRNA 和内含子上的差距。这说明对于局部结构相对清晰、长度较短或模式较集中的元件, 基线模型也能够学习到部分统计规律, 但整体上仍明显弱于 Evo2。

综合来看, CVAE 受限于模型容量和长程依赖建模能力, 在较长功能元件序列生成任务上表现最弱; HyenaDNA 虽属于预训练基因组语言模型, 但在当前主粮作物功能元件生成任务中的整体表现仍明显低于 Evo2。上述结果说明, Evo2 在大规模基因组预训练、长序列建模和跨元件模式学习方面具有更强优势, 因此在主粮作物基因组功能元件序列生成任务中表现出更优的整体性能。

4.3 实验 2: 微调模型与零样本基线的对比分析

为定量评估针对目标作物进行参数微调对模型生成结果一致性的影响, 本实验基于 7B 参数规模的 Evo2 基因组语言模型, 对比未经微调的预训练模型

（零样本基线）与在目标作物数据上完成微调的物种专属模型，在 6 类基因组功能元件上的平均序列相似度。本节所报告的平均序列相似度均按照 4.1.4 节中定义的统计口径计算。

为便于量化微调前后的变化幅度，本文同时采用绝对提升与相对提升作为辅助描述指标。

4.3.1 微调对不同功能元件相似性的整体影响

表 5 给出了零样本模型与微调模型在 6 类功能元件上的平均序列相似度对比结果。可以观察到，微调模型在所有功能元件上的平均序列相似度均高于零样本基线，且绝对提升均为正值，表明在目标作物数据上进行参数适配能够在不同类型功能元件上稳定提升生成序列与参考序列的一致程度。

Table 5 Similarity Before and After Fine-Tuning (7B)

表 5 微调前后序列相似度（7B）

元件类型	零样本模型/%	微调模型/%	绝对提升（百分点）	相对提升/%
内含子	61.80	62.63	+0.83	+1.34
mRNA	62.95	63.56	+0.61	+0.96
3′ UTR	61.47	62.04	+0.57	+0.93
5′ UTR	61.14	61.59	+0.46	+0.74
外显子	62.05	62.37	+0.32	+0.51
CDS	63.58	63.75	+0.17	+0.27

从提升幅度看，各功能元件的相似度绝对提升介于 0.17~0.83 个百分点，对应的相对提升范围为 0.27%~1.34%。整体提升幅度相对温和，说明大规模预训练模型已具备较强的通用序列建模能力，微调主要对生成分布进行针对特定物种的精细调整，而非对模型能力的根本性重构。

4.3.2 不同功能元件对微调的响应差异分析

从功能元件角度看，微调带来的收益呈现出明显的元件差异。内含子的提升最为明显，平均相似度由 61.80% 提升至 62.63%，绝对提升为 0.83 个百分点，相对提升为 1.34%。这说明对于变异度较高、物种特征更明显的元件，微调能够更有效地引入目标作物的特定序列模式。相比之下，CDS 的提升最小，仅由 63.58% 提升至 63.75%，绝对提升为 0.17 个百分点，相对提升为 0.27%。由于 CDS 直接参与蛋白编码，进化约束较强，其主要序列模式在预训练阶段已被较充分学习，因此微调带来的额外增益相对有限。外显子、mRNA 以及 5′ UTR 和 3′ UTR 的提升幅度介于内含子和 CDS 二者之间，反映出这些元件在功能约束与序列变异性之间具有过渡特征。

4.3.3 微调对整体平均相似度与结果稳定性的影响

从整体统计结果看（见表 6），微调使 6 类功能元件的整体平均相似度由 62.17% 提升至 62.66%，绝对提升为 0.49 个百分点，对应相对提升为 0.79%。尽管整体提升幅度有限，但该提升在所有功能元件上均保持一致方向，表明微调带来的增益具有稳定性，而非由少数元件或个别样本驱动。

Table 6 Overall Performance Improvement After Fine-Tuning (7B)

表 6 微调后整体性能提升（7B）

指标	零样本模型	微调模型	绝对提升（百分点）	相对提升/%
平均性能/%	62.17	62.66	+0.49	+0.79
标准差	1.01	0.93		
最小值/%	59.89	61.04		
最大值/%	63.98	64.44		

注：平均相似度为 6 类功能元件相似性的两级算术平均，统计口径同 4.1.4 节。

此外，微调后相似度的标准差由 1.01 降至 0.93，说明在提升平均相似度的同时，不同功能元件之间的结果有所降低，模型生成结果的一致性得到一定改善。

4.3.4 图示结果的直观验证

图 1 进一步给出了各功能元件在微调后的相对提升幅度分布情况，其中内含子的提升最为明显，而 CDS 的提升幅度最小，与表 5 和表 6 中的绝对提升与相对提升结果一致。

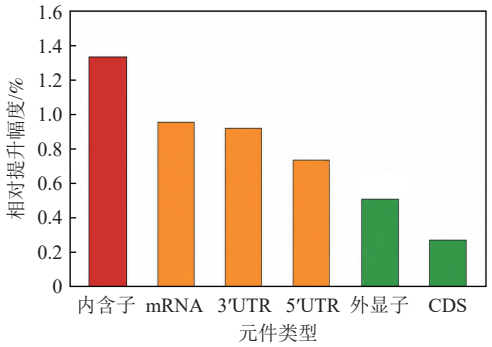


Fig. 1 Similarity improvement after fine-tuning

图 1 微调后的相对提升

4.4 实验 3：跨物种迁移学习性能分析

为评估微调后模型在不同作物间的泛化能力，本实验系统比较了模型在同物种场景（训练与测试数据来自同一作物）与跨物种场景（训练作物与测试作物不同）下的生成结果一致性。评价指标采用 4.1.4 节定义的平均序列相似度，统计口径保持一致。

表 7 汇总了 6 类基因组功能元件在同物种与跨物种条件下的平均相似度及其差异情况。

Table 7 Cross-species Similarity of Different Functional Elements

表 7 不同功能元件跨物种相似度

功能元件类型	同物种平均/%	跨物种平均/%	泛化差距（百分点）	相似度最高/%	相似度最低/%	相似度最高最低差值（百分点）
CDS	63.75	63.33	0.42	64.02	61.89	2.13
外显子	62.37	61.8	0.57	62.85	60.37	2.48
内含子	62.63	62.42	0.21	64.4	60.31	4.09
5′ UTR	61.59	60.94	0.65	62.73	59.85	2.88
3′ UTR	62.04	61.84	0.2	63.24	60.49	2.75
mRNA	63.56	63.19	0.37	64.44	62.48	1.95

从整体结果看，所有功能元件在跨物种测试下的平均相似度均略低于同物种场景，但相似度差距均小于 1 个百分点，表明模型在不同作物间具备较强的迁移能力。其中，3′ UTR 与内含子的跨物种相似度差距最小（均不超过 0.33 个百分点），而 5′ UTR 的差距相对较大（0.65 个百分点），显示其对物种差异更为敏感。

从相似度分布范围看，mRNA 的相似度波动区间最小（最高值与最低值差为 1.95 个百分点），生成结果最为稳定；内含子的波动范围最大（4.09 个百分点），反映其序列变异度较高、结构约束较弱的特性。

综合表 7 及图 2、图 3 可以看出，模型在同物种与跨物种场景下的整体表现较为接近，说明微调后模型具有较好的跨作物迁移能力。即使是在物种特异性更强的 5′ UTR 上，同物种测试的优势也主要表现为轻微提升，而非显著下降。换言之，模型在学习目标作物特征的同时，并未明显损害其对其他作物序列模式的泛化能力。这一结果表明，基于 Evo2 的微调模型不仅适用于作物专属生成，也具备一定的跨物种应用潜力。

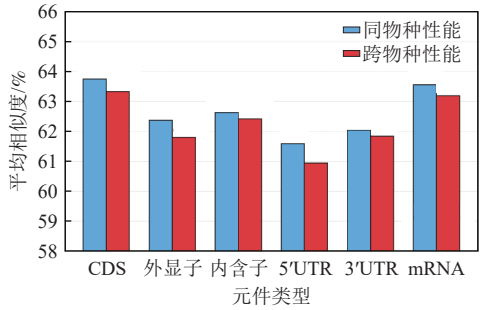


Fig. 2 Similarity in within- and cross-species scenarios

图 2 同物种与跨物种场景下的相似度

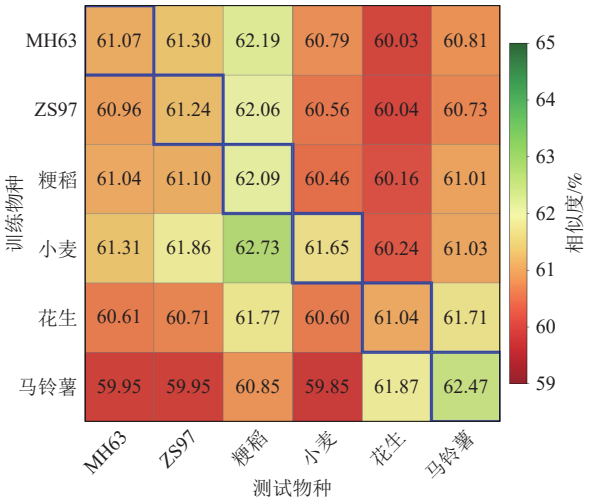


Fig. 3 Heatmap of cross-species similarity for 5′ UTR

图 3 5′ UTR 跨物种相似度热图

4.5 实验 4：模型规模对序列生成一致性的影响分析

为分析模型参数规模对作物基因组序列生成结果的影响，本实验对比了 Evo2 基因组语言模型在 1B 与 7B 参数规模下的微调模型表现。评价指标采用 4.1.4 节中定义的平均序列相似度。

在此基础上，为定量刻画模型规模变化带来的影响，本文进一步引入绝对变化与相对变化 2 个差值指标，用于衡量 7B 模型相对于 1B 模型在不同功能元件上的相似度变化幅度。其定义为：

绝对变化（absolute change）：

$$\Delta_{\text{abs}} = Similarity_{7B} - Similarity_{1B}。$$

该指标反映模型参数规模由 1B 增加至 7B 后，平均序列相似度的绝对差异。

相对变化（relative change）：

$$\Delta_{\text{rel}} = \frac{Similarity_{7B} - Similarity_{1B}}{Similarity_{1B}} \times 100\%。$$

该指标用于描述相似度变化相对于 1B 模型基线的比例大小。

表 8 给出了 6 类基因组功能元件在 1B 与 7B 微调模型下的平均序列相似度、标准差及对应的绝对

Table 8 Similarity of 1B and 7B Models

表 8 1B 与 7B 模型相似度

元件类型	1B 微调模型/%	1B 微调标准差	7B 微调模型/%	7B 微调标准差	绝对变化（百分点）	相对变化/%
外显子	61.6	0.84	62.37	0.41	+0.77	+1.24
CDS	63.21	0.73	63.75	0.29	+0.54	+0.86
5′ UTR	61.08	1.13	61.59	0.59	+0.51	+0.83
mRNA	63.85	0.62	63.56	0.5	−0.28	−0.45
3′ UTR	62.81	0.59	62.04	0.43	−0.77	−1.23
内含子	63.98	0.65	62.63	0.91	−1.35	−2.11

变化与相对变化结果。从整体平均结果看(见表 9)，1B 与 7B 模型的整体平均相似度分别为 62.76% 与 62.66%，绝对变化仅为−0.10%，表明在当前数据规模与任务设置下，单纯增加模型参数并未带来显著的整体相似度提升。

Table 9　Effect of Model Scale on Similarity

表 9　模型规模对相似度的影响

指标	1B 微调模型	7B 微调模型	绝对变化（百分点）	相对变化/%
平均性能/%	62.76	62.66	−0.1	−0.16
标准差	1.3	0.93		
最小值/%	59.57	61.04		
最大值/%	64.75	64.44		

分功能元件结果表明，模型规模的影响并不一致。对于外显子、CDS 和 5′ UTR，7B 模型的平均相似度高于 1B 模型，说明在结构约束较强或局部模式较明确的元件上，更大规模模型在细粒度序列建模方面具有一定优势。相反，在 mRNA、3′ UTR 和内含子上，1B 模型反而略优，其中内含子的差异最为明显。这表明当元件本身变异度较高、训练数据规模又相对有限时，更大规模模型未必能够带来更高的平均相似度，甚至可能因对局部模式拟合过强而影响泛化表现。

尽管 1B 与 7B 模型在整体平均相似度上差异不大，但二者在结果稳定性上存在明显区别。表 9 显示，7B 模型的相似度标准差低于 1B 模型，说明其在不同功能元件上的生成结果更均衡、波动更小。换言之，在当前任务设置下，更大规模模型的优势更多体现在稳定性提升，而不是整体平均性能的显著增长。图 4 与图 5 从分元件和整体 2 个角度直观展示了模型规模变化对相似度的影响。

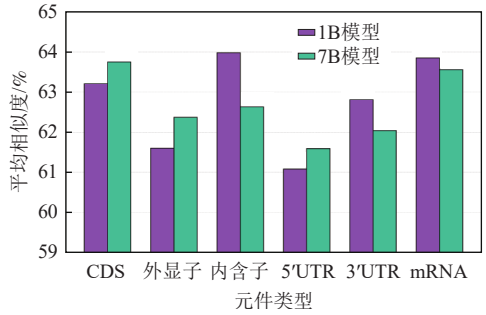


Fig. 4　Similarity comparison between 1B and 7B models

图 4　1B 与 7B 模型相似度对比

4.6　实验 5: 基因组功能元件类型差异化分析

为分析不同功能元件的生物学属性对模型生成

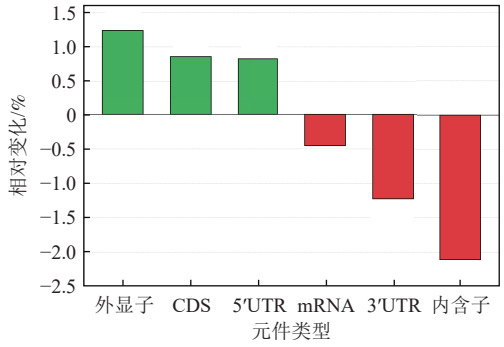


Fig. 5　Effect of model scale on average similarity

图 5　模型规模对平均相似度的影响

行为的影响，本文进一步比较了 6 类功能元件在 6 种作物上的平均序列相似度。表 10 显示，不同功能元件之间存在较稳定的性能分层。总体上，CDS 与 mRNA 在多数作物中均表现出较高的平均相似度，说明编码相关元件由于进化约束更强、序列模式更稳定，更容易被模型一致地学习与生成；相比之下，5′ UTR 的平均相似度整体最低，且不同作物之间波动更明显。

Table 10　Similarity Across Species and Elements

表 10　不同物种和元件的相似度　%

物种	CDS	外显子	内含子	5′ UTR	3′ UTR	mRNA
MH63	63.83	62.2	61.93	61.07	61.68	63.48
ZS97	63.88	62.44	62.05	61.24	62.08	62.97
粳稻	63.52	61.71	62.57	62.09	61.76	63.51
小麦	63.95	62.72	62.66	61.65	61.94	63.74
花生	63.28	62.29	62.16	61.04	61.91	63.22
马铃薯	64.02	62.85	64.4	62.47	62.87	64.44

表 11 进一步从同物种均值、跨物种均值和衰减率等角度对这一现象进行了量化。结果表明，CDS 与 mRNA 的同物种平均相似度较高且标准差较小，说明模型在编码相关元件上具有更好的稳定性；5′ UTR 的同物种均值最低而跨物种衰减率最高，反映

Table 11　Statistics and Attenuation of Element Similarity

表 11　各元件相似度统计与衰减

元件类型	同物种均值/%	同物种标准差	跨物种均值/%	跨物种标准差	跨物种衰减率/%	测试样本数
5′ UTR	61.59	0.59	60.94	0.74	1.06	36
外显子	62.37	0.41	61.80	0.64	0.91	36
CDS	63.75	0.29	63.33	0.57	0.66	36
mRNA	63.56	0.50	63.19	0.39	0.59	36
内含子	62.63	0.91	62.42	0.93	0.33	36
3′ UTR	62.04	0.43	61.84	0.71	0.33	36

出其对物种特异调控机制更为敏感。相比之下,内含子和 3′ UTR 的跨物种衰减率较低,说明虽然这些元件本身变异性较强,但模型在迁移场景下仍能保持相对稳定的整体一致性。

图 6 通过箱线图展示了不同功能元件相似度的分布情况,CDS 与 mRNA 的分布集中,而 5′ UTR 与内含子的分布更为分散。图 7 与图 8 分别从物种与元件交互视角和衰减率角度,进一步验证了不同功能元件在生成难度与跨物种稳定性上的系统性差异。

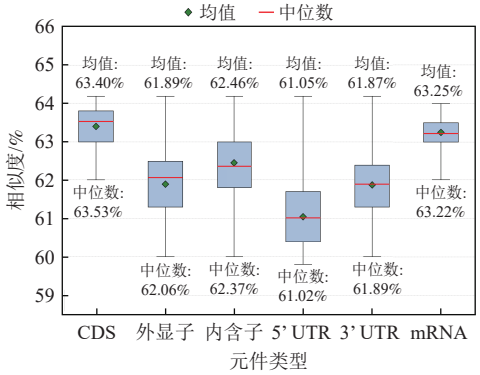


Fig. 6 Boxplot of similarity distribution for element type

图 6 不同元件类型相似度分布箱线图

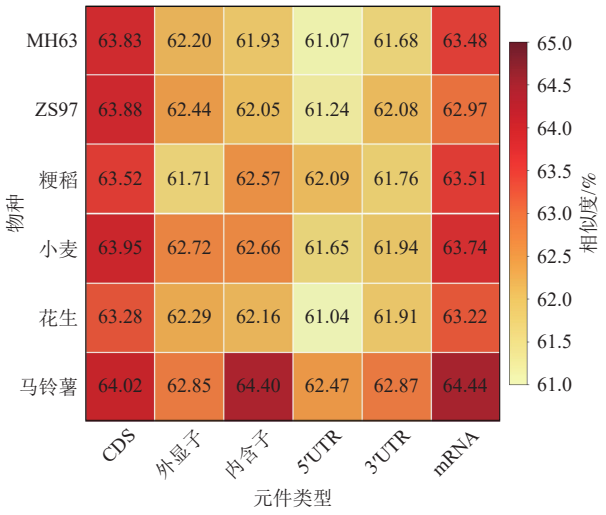


Fig. 7 Heatmap of similarity across species and elements

图 7 不同物种与元件的相似度热图

4.7 生成序列的生物学合理性补充分析

为进一步评估模型生成序列在生物学规则层面的合理性,本文对 CDS、5′ UTR、3′ UTR 和内含子开展了补充性一致性分析。

1)CDS: ORF 完整性与密码子分布

在 CDS 上,生成序列在长 ORF 连续性方面表现出一定合理性。与真实序列相比,生成序列的无终止密码子比例由 0.553 3 提升至 0.630 0,最长 ORF 长度占序列长度的平均比例由 0.763 0 提升至 0.986 8,说明模型在维持阅读框连续性方面具备一定能力。但严格完整 ORF 的比例仍为 0,表明模型尚未稳定恢复完整编码结构。进一步比较密码子使用分布可以发现,生成序列与真实序列之间仍存在明显差异,JSD 为 0.834 1、Pearson 相关系数仅为 0.091 0、L1 距离为 1.873 1,说明模型对编码区局部统计模式有一定学习能力,但对更精细的翻译相关偏好尚未充分捕获。

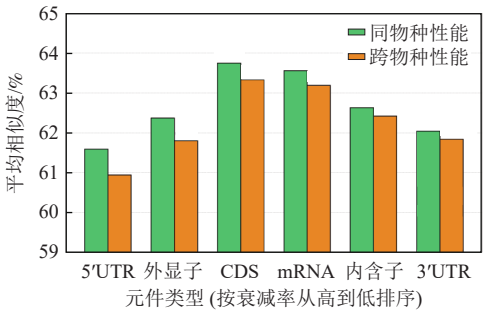


Fig. 8 Cross-species similarity attenuation rate of element type

图 8 不同元件类型的跨物种相似度衰减率

2)5′ UTR: Kozak 基序

在 5′ UTR 上,真实序列的 Kozak 核心模式命中率为 0.356 7,而生成序列为 0; Kozak 完整模式在真实序列中的命中率为 0.003 3,而生成序列同样为 0。该结果表明,模型尚未有效恢复与翻译起始相关的典型调控基序,说明其对 5′ UTR 调控信号的学习仍然较弱。

3)3′ UTR: Poly(A) 信号

在 3′ UTR 上,生成序列能够恢复部分 Poly(A) 信号,命中率为 0.176 7,而真实序列对应值为 0.356 7。虽然模型对转录后加工相关调控模式已有一定捕获能力,但整体命中率仍明显低于真实序列,说明其对 3′ UTR 关键调控特征的恢复仍然有限。

4)内含子: GT-AG 剪接规则

对内含子剪接规则的分析表明,生成序列在供体端 GT 特征上有一定保留,命中率由真实序列的 0.040 0 上升至 0.056 7; 但在受体端 AG 上,生成序列命中率为 0,低于真实序列的 0.056 7,完整 GT-AG 规则命中率同样为 0。总体来看,模型对剪接相关约束的学习仍不充分,尚未稳定形成完整的典型剪接模式。

综合来看,Evo2 微调模型在序列层面的生成能力较好,并在部分 ORF 连续性指标上表现出一定合理性,但在密码子偏好、调控基序和剪接规则等更细粒度的生物学机制层面仍存在明显不足。这说明当

前模型更接近于对序列统计分布的拟合，而距离具有更强机制约束的生物学合理生成仍有进一步提升空间。

4.8 小 结

4.1~4.7 节围绕基线方法对比、微调效果、跨物种迁移、模型规模差异、功能元件类型差异以及补充生物信息学一致性分析 6 个方面，对 Evo2 微调模型在主粮作物基因组序列生成任务中的表现进行了系统评估。结果表明，微调能够在不同功能元件上稳定提升生成序列与参考序列之间的一致性，且模型在跨物种场景下仍保持较好的泛化能力。不同功能元件的生成难度存在明显差异，编码相关元件整体更稳定，5′ UTR 等调控元件则更敏感。模型规模扩大并未带来整体平均性能的显著提升，但有助于提高结果稳定性。补充生物信息学分析进一步说明，当前模型在部分结构连续性指标上具有一定合理性，但在密码子偏好、调控基序和剪接规则等细粒度机制层面仍存在提升空间。

5 讨 论

5.1 框架创新性与应用价值

本研究的核心创新在于将前沿的基因组语言模型与农业基因组学的实际需求相结合，建立了一个标准化、可扩展的作物序列生成与评估流程。本文构建了基于作物专属微调与跨物种迁移评估的实验框架，可从同物种适配、跨物种泛化、模型规模差异和功能元件差异等多个维度系统评估基因组语言模型在主粮作物场景中的应用潜力。该框架可直接应用于等位基因序列设计、合成生物学元件优化及数据增强等场景[28-29]。

5.2 局限性分析

本文仍存在 4 个方面的局限。1)本文的主要评价仍以序列相似度和补充性规则分析为主，尚未通过湿实验或下游功能实验对生成序列的真实生物学活性进行验证，因此当前结论更多反映的是序列层面和规则层面的合理性，而非功能层面的最终有效性。2)研究对象虽然涵盖 6 个作物品种和 6 类功能元件，但总体作物范围仍然有限，尚不足以完全代表更广泛农业基因组场景下的复杂情况。3)本文采用全参数微调策略，能够较充分地适配目标作物分布，但在更大规模模型或更复杂多任务设置下，其训练成本与部署成本仍然较高。4)本文虽然比较了同物种与跨物种场景下的表现差异，但尚未进一步结合系统发育距离、基因组复杂度和调控特征差异，对模型泛化边界的来源进行更细粒度解释。

5.3 未来工作展望

基于本文构建的评估体系与迁移学习框架，后续工作可从 4 个方向进一步展开。

1)扩展作物与数据类型覆盖范围。在现有主粮作物基础上，进一步引入玉米、大豆等重要作物，并结合更多来源的基因组与转录组数据，验证方法在更广泛农业场景中的适用性[30-32]。

2)构建更贴近农业基因组学需求的领域模型。可在大规模多作物基因组数据基础上，进一步融合作物特有的遗传结构、驯化背景与育种知识，推动模型从通用序列生成向农业场景下的专用智能设计发展[28,30]。

3)加强生物学约束驱动的生成优化。后续可将 ORF 完整性、剪接位点规则、保守调控基序等显式纳入微调或生成过程，并结合密码子偏好、表达水平预测、RNA 二级结构稳定性等下游目标，构建“生成—评价—优化”的闭环框架，提升候选序列在结构和功能层面的合理性[33]。

4)探索轻量化与可部署方案。针对农业研究机构和实际应用中的算力限制，未来可进一步研究参数高效微调、模型压缩、知识蒸馏和量化等方法在基因组语言模型中的适用性，为模型在真实农业科研环境中的推广提供支持。

6 结 论

本文系统研究了基于 Evo2 基因组语言模型的迁移学习框架在主粮作物基因组序列生成任务中的应用效果。围绕多作物、多功能元件场景的实验结果表明：针对目标作物进行微调能够稳定提升模型生成序列与真实参考序列之间的一致性；微调后的模型在跨物种测试中仍保持较好的泛化能力，说明其具备一定的迁移应用潜力；不同功能元件的生成表现存在明显差异，其中编码相关元件整体更稳定，5′ UTR 等调控相关元件对物种差异更为敏感；模型规模扩大并未显著提高整体平均相似度，但有助于降低不同元件之间的结果波动。补充生物信息学分析进一步表明，当前模型在 ORF 连续性等方面已表现出一定合理性，但在密码子偏好、调控基序和剪接规则等细粒度生物学机制层面仍有较大改进空间。总体来看，本文为基因组语言模型在作物基因组生成

任务中的应用提供了较为系统的实验依据,也为后续面向功能约束和农业育种需求的序列设计研究奠定了基础。

作者贡献声明:何曼曼负责研究方案设计、实验组织、结果分析和论文撰写;刘龙正负责算法框架设计、模型训练、实验实现和论文撰写;丁开锋参与算法实现、实验方法设计和结果整理;葛明亮参与数据处理、实验辅助和结果可视化;张雨琪负责作物基因组数据整理、注释信息核对和生物学分析支持;任澳雨参与小麦相关数据整理、基因组背景分析和实验结果核对;郭伟龙提出作物基因组学研究思路,指导数据来源、生物学解释和论文修改;王耀君提出总体研究思路,负责技术路线指导、论文审阅和修改完善。何曼曼和刘龙正为共同第一作者。

参 考 文 献

[1] Hu Yunshu, Zhou Jun, Cao Zhenfu, et al. A lightweight multi-user verifiable privacy-preserving genomic sequence analysis scheme[J]. Journal of Computer Research and Development, 2024, 61(10): 2448−2466 (in Chinese)
(胡云舒, 周俊, 曹珍富, 等. 轻量级多用户可验证隐私保护基因序列分析方案[J]. 计算机研究与发展, 2024, 61(10): 2448−2466)

[2] Wu Jiansheng, Feng Qiaoyu, Yuan Jingzhou, et al. Prediction of biological functions of G protein-coupled receptors based on fast multi-instance multi-label learning[J]. Journal of Computer Research and Development, 2018, 55(8): 1674−1682 (in Chinese)
(吴建盛, 冯巧遇, 袁京洲, 等. 基于快速多示例多标记学习的 G 蛋白偶联受体生物学功能预测[J]. 计算机研究与发展, 2018, 55(8): 1674−1682)

[3] Ge Hongwei, Liang Yanchun. A multiple sequence alignment algorithm based on hidden Markov model and immune particle swarm optimization[J]. Journal of Computer Research and Development, 2006, 43(8): 1330−1336 (in Chinese)
(葛宏伟, 梁艳春. 基于隐马尔可夫模型和免疫粒子群优化的多序列比对算法[J]. 计算机研究与发展, 2006, 43(8): 1330−1336)

[4] Wei Xuechao, Zhou Zhe, Xu Yinghui, et al. PetS: Scalable inference service system for parameter-efficient Transformer models[J]. Journal of Computer Research and Development, 2025, 62(7): 1816−1832 (in Chinese)
(魏学超, 周哲, 徐盈辉, 等. PetS: 针对参数高效 Transformer 模型的可扩展推理服务系统[J]. 计算机研究与发展, 2025, 62(7): 1816−1832)

[5] Liu Hai, Peng Changgen, Wu Zhenqiang, et al. A survey of theories and methods for genomic data privacy protection[J]. Chinese Journal of Computers, 2021, 44(7): 1430−1480 (in Chinese)
(刘海, 彭长根, 吴振强, 等. 基因组数据隐私保护理论与方法综述[J]. 计算机学报, 2021, 44(7): 1430−1480)

[6] Wang Yuanrong, Zeng Ping, Zang Dawei, et al. Parallel design and optimization of the Gene Panel workflow[J]. Chinese Journal of Computers, 2019, 42(11): 2429−2446 (in Chinese)
(王元戎, 曾平, 臧大伟, 等. Gene Panel 流程的并行设计与优化研究[J]. 计算机学报, 2019, 42(11): 2429−2446)

[7] Wang Kaili, Zhang Li, Liu Xuejun. Differential expression analysis of transcriptomes by integrating multi-platform expression data[J]. Chinese Journal of Computers, 2018, 41(6): 1415−1430 (in Chinese)
(王凯莉, 张礼, 刘学军. 融合多平台表达数据的转录组差异表达分析[J]. 计算机学报, 2018, 41(6): 1415−1430)

[8] Wu Hao. An algorithm for detecting cancer driver pathways based on mutated gene networks[J]. Chinese Journal of Computers, 2018, 41(6): 1400−1414 (in Chinese)
(吴昊. 基于突变基因网络的致癌驱动通路检测算法[J]. 计算机学报, 2018, 41(6): 1400−1414)

[9] Zhang Yipu, Huo Hongwei, Yu Qiang, et al. A localization projection refinement algorithm for transcription factor binding site identification[J]. Chinese Journal of Computers, 2013, 36(12): 2545−2559 (in Chinese)
(张懿璞, 霍红卫, 于强, 等. 用于转录因子结合位点识别的定位投影求精算法[J]. 计算机学报, 2013, 36(12): 2545−2559)

[10] Dong Zhengyu, Zhang Weifeng, Xu Lei, et al. An automated merge conflict resolution method based on code context and large language models[J]. Chinese Journal of Computers, 2025, 48(10): 2468−2486 (in Chinese)
(董政宇, 张卫丰, 许蕾, 等. 基于代码上下文与大语言模型的自动化合并冲突解决方法[J]. 计算机学报, 2025, 48(10): 2468−2486)

[11] Zhou Hao, Dai Hua, Yang Geng, et al. Privacy-preserving verifiable federated learning based on biometric recognition[J]. Chinese Journal of Computers, 2025, 48(8): 1848−1869 (in Chinese)
(周浩, 戴华, 杨庚, 等. 基于生物特征识别的隐私保护可验证联邦学习[J]. 计算机学报, 2025, 48(8): 1848−1869)

[12] Consens M E, Dufault C, Wainberg M, et al. Transformers and genome language models[J]. Nature Machine Intelligence, 2025, 7: 346−362

[13] Nguyen E, Poli M, Durrant M G, et al. Sequence modeling and design from molecular to genome scale with Evo[J]. Science, 2024, 386(6723): eado9336

[14] Brixi G, Durrant M G, Ku J, et al. Genome modeling and design across all domains of life with Evo 2[J/OL]. bioRxiv, 2025. DOI: 10.1101/2025.02.18.638918

[15] Ji Yanrong, Zhou Zhihan, Liu Han, et al. DNABERT: Pre-trained bidirectional encoder representations from Transformers model for DNA-language in genome[J]. Bioinformatics, 2021, 37(15): 2112−2120

[16] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, et al. Nucleotide Transformer: Building and evaluating robust foundation models for human genomics[J]. Nature Methods, 2025, 22(2): 287−297

[17] Mendoza-Revilla J, Castillo J M, Vanderschuren H, et al. A foundational large language model for edible plant genomes[J]. Communications Biology, 2024, 7: 835

[18] MacNish T R, Danilevicz M F, Bayer P E, et al. Application of machine learning and genomics for orphan crop improvement[J]. Nature Communications, 2025, 16: 982

[19] Pérez-de-Castro A M, Vilanova S, Cañizares J, et al. Application of genomic tools in plant breeding[J]. Current Genomics, 2012, 13(3):

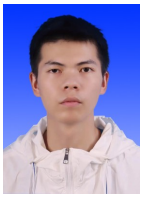
179−195
[20] Kim K D, Kang Y, Kim C. Application of genomic big data in plant breeding: Past, present, and future[J]. Plants, 2020, 9(11): 1454
[21] Balakrishnan P, Anny Leema A, Dhivya Shree V, et al. Gene-LLMs: A comprehensive survey of Transformer-based genomic language models for regulatory and clinical genomics[J]. Frontiers in Genetics, 2025, 16: 1634882
[22] Varshney R K, Bohra A, Yu J, et al. Designing future crops: Genomics-assisted breeding comes of age[J]. Trends in Plant Science, 2021, 26(6): 631−649
[23] Wójcik-Gront E, Zieniuk B, Pawełkowicz M. Harnessing AI-Powered genomic research for sustainable crop improvement[J]. Agriculture, 2024, 14(12): 2299
[24] NVIDIA. Massive Foundation Model for Biomolecular Sciences Now Available via NVIDIA BioNeMo[EB/OL]. NVIDIA Blog, 2025 [2025-12-21]. https://blogs.nvidia.com/blog/evo-2-biomolecular-ai/
[25] Wang S, Gao S, Nie J, et al. Improved 93-11 genome and time-course transcriptome expand resources for rice genomics[J]. Frontiers in Plant Science, 2022, 12: 769700
[26] International Rice Genome Sequencing Project. The map-based sequence of the rice genome[J]. Nature, 2005, 436(7052): 793−800
[27] Yates A D, Allen J, Amode R M, et al. Ensembl Genomes 2022: An expanding genome resource for non-vertebrates[J]. Nucleic Acids Research, 2022, 50(D1): D996−D1003
[28] Liu Guanqing, Chen Long, Wu Yuechao, et al. PDLLMs: A group of tailored DNA large language models for analyzing plant genomes[J]. Molecular Plant, 2025, 18(2): 175−178
[29] Krishna T P A, Veeramuthu D, Maharajan T, et al. The era of plant breeding: Conventional breeding to genomics-assisted breeding for crop improvement[J]. Current Genomics, 2023, 24(1): 24−35
[30] Zhu W, Li W, Zhang H, et al. Big data and artificial intelligence-aided crop breeding: Progress and prospects[J]. Journal of Integrative Plant Biology, 2025, 67(3): 722−739
[31] Benegas G, Batra S S, Song Y S. DNA language models are powerful predictors of genome-wide variant effects[J]. Proceedings of the National Academy of Sciences, 2023, 120(44): e2311219120
[32] Sanabria M, Hirsch J, Joubert P M, et al. DNA language model GROVER learns sequence context in the human genome[J]. Nature Machine Intelligence, 2024, 6: 911–923
[33] Sinha D, Maurya A K, Abdi G, et al. Integrated genomic selection for accelerating breeding programs of climate-smart cereals[J]. Genes, 2023, 14(7): 1484



He Manman, born in 1997. PhD candidate. Her main research interest includes bioinformatics.
何曼曼, 1997 年生。博士研究生。主要研究方向为生物信息学。



Liu Longzheng, born in 2001. Master candidate. His main research interest includes knowledge graphs.
刘龙正, 2001 年生。硕士研究生。主要研究方向为知识图谱。



Ding Kaifeng, born in 2005. Undergraduate. His main research interest includes knowledge graphs.
丁开锋, 2005 年生。本科生。主要研究方向为知识图谱。



Ge Mingliang, born in 1999. Master candidate. His main research interest includes target detection.
葛明亮, 1999 年生。硕士研究生。主要研究方向为目标检测。



Zhang Yuqi, born in 2000. PhD candidate. Her main research interest includes wheat pan-genome analysis.
张雨琪, 2000 年生。博士研究生。主要研究方向为小麦泛基因解析。



Ren Aoyu, born in 2000. PhD candidate. His main research interest includes evolutionary analysis of wheat germplasm resources.
任澳雨, 2000 年生。博士研究生。主要研究方向为小麦种质资源的演化分析。



Guo Weilong, born in 1988. PhD, professor. His main research interest includes analysis of the wheat genomics and intelligent breeding.
郭伟龙, 1988 年生。博士，教授。主要研究方向为小麦基因组解析与智慧育种。



Wang Yaojun, born in 1985. PhD, associate professor. His main research interests include knowledge graphs, artificial intelligence, and large models.
王耀君, 1985 年生。博士，副教授。主要研究方向为知识图谱、人工智能、大模型。