基于Transformer的大豆基因组预测方法

陈艾灵

四川大学

DOI:10.12238/as.v8i3.2800

[摘要] 随着大豆育种的需求日益增加,传统育种方法面临着时间长、成本高以及环境因素影响较大的挑战。因此,基因组预测技术成为了一种提升大豆育种效率的关键工具。研究的目的在于应用Transformer模型,结合大豆SNP数据,构建一个准确、高效的大豆基因组预测模型,探索其在基因型到表型预测中的应用潜力。首先,通过数据预处理和特征选择,将大豆SNP数据转换为适合Transformer模型输入的格式。其次,结合Transformer模型中的自注意力机制,探索如何捕捉SNP间的长距离依赖关系,从而提高性状预测的准确性。为了进一步优化模型性能,采用了贝叶斯优化算法,自动化搜索超参数配置,以提高计算效率和预测精度。实验结果表明,与传统的CNN、LSTM等模型相比,基于Transformer模型的大豆基因组预测框架,在单性状预测及多性状联合预测任务中均表现出较高的精度。此外,贝叶斯优化进一步提升了模型的超参数选择效率,减少了手动调参的工作量。

[关键词] 大豆; 单核苷酸多态性; Transformer; Bayes优化

中图分类号: S565.1 **文献标识码:** A

Soybean Genome Prediction Method Based on Transformer

Ailing Chen

Sichuan University

[Abstract] With the increasing demand of soybean breeding, traditional breeding methods are faced with the challenges of long time, high cost and great influence of environmental factors. Therefore, genome prediction technology has become a key tool to improve the efficiency of soybean breeding. The purpose of this study is to build an accurate and efficient soybean genome prediction model with Transformer model combined with soybean SNP data, and explore its application potential in genotypic to phenotypic prediction. First, the soybean SNP data is converted into a format suitable for Transformer model inputs through data preprocessing and feature selection. Secondly, combined with the self-attention mechanism in Transformer model, explore how to capture the long-distance dependency between SNPs to improve the accuracy of trait prediction. In order to further optimize the performance of the model, a Bayesian optimization algorithm is used to automatically search for hyperparameter configurations to improve the computational efficiency and prediction accuracy. The experimental results show that compared with traditional models such as CNN and LSTM, the soybean genome prediction framework based on Transformer model has higher accuracy in both single trait prediction and multi-trait combined prediction tasks. In addition, Bayesian optimization further improves the efficiency of hyperparameter selection and reduces the workload of manual parameter adjustment.

[Key words] Soybean; Single Nucleotide Polymorphism; Transformer; Bayes Optimization

引言

大豆是全球三大主要农作物之一,其所含的蛋白质比玉米、高粱、燕麦等更易于动物消化^[1]。此外,大豆还富含有益健康的饱和和脂肪酸,特别是亚油酸^[2]。基因组预测(Genomic Prediction, GP)是一种通过统计学方法,分析基因型与表型之

间的关系,从而实现性状预测的生物信息学技术^[3]。而基因组学研究的迅猛发展,带动了基于大豆基因组数据开展性状预测与遗传改良,这不仅为大豆遗传改良及育种实践迈向新高度奠定了理论和技术基础,同时也为加速大豆育种进程提供了崭新的机遇。通过日趋成熟的高通量测序技术,我们更容易获取到基因

数据。对于大豆这种基因高度纯合、遗传背景简单的自交植物,更是锦上添花。

近年来,基于深度学习的模型,特别是Transformer架构,已经取得了突破性进展。该模型的核心创新在于引入了自注意力机制,能够捕捉输入序列中不同位置之间的远距离依赖关系,而无需依赖传统的卷积或递归结构。这种创新使得Transformer模型在处理长序列建模任务时展现出显著的优越性。自Vaswani等人于2017年首次提出Transformer架构以来,该研究工作为自然语言处理领域开启了新的研究范式,其应用范围涵盖了机器翻译、文本生成、情感分析等多个重要任务^[4]。

1 实验材料

植物材料是从公开数据库Soy获得的WM82. A1数据集的大豆SNP数据。它对20087份大豆材料进行基因分型,删除等位基杂合率超过10%的样本,得到19648份大豆种质(包括1168份野生和18480份栽培大豆)。需要对原始数据集上进行统一的质控:首先,使用Plink软件删除等位基因缺失率超过10%的样本;再删除次等位基因频率低于5%的SNP点位;最后对缺失的基因型数据进行推断填充。最终我们保留11779份样本,每份样本含42453个SNP位点,84906个碱基。

将SNP数据进行One-Hot编码,使得每个SNP位点的基因型表示为一个二进制向量,例如在某SNP位点,A为主等位基因,G为次等位基因,编码方式为:AA→0,AG→1,GG→2。性状数据采用含油量、蛋白质含量,经过Z-score标准化处理,标准化公式为:

$$Z = \frac{X - \mu}{\sigma}$$

其中, X 为原始数据, μ 为数据的均值, σ 为数据的标准差。

2 实验方法

Transformer的输入由One-Hot编码后的SNP数据和位置编码共同构成^[5]。位置编码通过正弦和余弦函数生成,公式为:

$$P_i^j = \begin{cases} \sin(\omega_k i), & \text{若 } j \text{ 为偶数} \\ \cos(\omega_k i), & \text{若 } j \text{ 为奇数} \end{cases}$$

其中, $\omega = \frac{1}{10000^{\frac{2k}{d}}}$, i 为位置索引, j 为维度索引, d 为位置编码的维度大小。

Transformer通过自注意力机制捕捉不同SNP位点间的长距离依赖关系,并通过前馈神经网络进一步编码。自注意力机制的核心计算公式为:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

其中, Q, K, V 分别为查询、 d_k 键和值矩阵,为键的维度大小。多头注意力机制通过多个头的并行计算得到不同的上下文信息。

为了使模型更好的训练,对于含油量和蛋白质含量的单性状预测任务采用了均方误差作为损失函数,

$$MSE = \frac{1}{N} \sum_{k=1}^N (y_k - \hat{y}_k)^2$$

其中, y_k 是真实值, $\hat{y}_k$ 是预测值, N 是样本量。通过最小化损失函数,模型能够逐步优化预测结果。

对于多种性状联合预测任务,针对含油量和蛋白质含量进行联合预测。采用损失函数为:

$$Loss = \frac{1}{2\sigma_1^2} MSE_{\text{含油量}} + \frac{1}{2\sigma_2^2} MSE_{\text{蛋白质含量}} + \log(\sigma_1 \sigma_2)$$

其中, σ 是对该任务预测的不确定性的一个度量。

Bayes优化是一种用于高效优化目标函数的方法,通过构建一个代理模型来估计目标函数,并基于该代理模型来选择下一步的超参数配置,从而在有限时间内找到目标函数的较优解。在Transformer模型的训练中,超参数的选择对模型性能具有决定性影响。Bayes优化采用高斯过程作为代理模型,结合期望改进作为采集函数,从而寻找整个模型的超参数。

3 实验结果与分析

所有实验均在RTX3090显卡的Python3.8环境下完成。评价指标为皮尔逊相关系数(PCC)和平均绝对误差(MAE)。MAE越小,表示模型预测误差越低,预测性能越好。PCC值越接近1,说明模型的预测结果与真实值之间的正线性相关性越强。

从Table1和Table2的含油量和蛋白质含量预测结果来看,Transformer模型在单一性状预测中的表现显著优于传统的LSTM、GRU和CNN模型。其中CoTransformer是联合两种性状共同建模。尤其在PCC和MAE这两个评价指标中,Transformer模型较LSTM和GRU具有明显的优势,显示了其强大的全局建模能力。此外,CoTransformer模型通过联合多性状预测策略进一步优化了预测效果,尤其在皮尔逊相关系数上有显著提升。

实验结果表明,基于Transformer模型的大豆基因组预测框架具有较高的预测精度,特别是在处理大规模、高维度的SNP数据时,Transformer模型能够有效捕捉基因型与表型之间的复杂依赖关系,提供了比传统方法更为准确和高效的解决方案。未来,随着基因组测序技术的发展,Transformer模型在大豆育种和遗传改良中的应用潜力将进一步得到释放。

Table1含油量预测结果

	PCC	MAE
LSTM	0.3212	0.4345
GRU	0.3314	0.3712
CNN	0.1455	0.2391
Trasnformer	0.3239	0.2474
CoTrasnformer	0.3882	0.2303

Table 2蛋白质含量预测结果

	PCC	MAE
LSTM	0.1544	0.3794
GRU	0.1655	0.3972
CNN	0.2401	0.2514
Transformer	0.2573	0.3525
CoTransformer	0.2655	0.3321

[参考文献]

[1] Cabanos C, Matsuoka Y, Maruyama N. Soybean proteins/peptides: A review on their importance, biosynthesis, vacuolar sorting, and accumulation in seeds[J]. Peptides, 2021, 143: 170598.

[2] Vargas-Almendra A, Ruiz-Medrano R, Núñez-Muñoz L A, et al. Advances in Soybean Genetic Improvement[J]. Plants, 2024,

13(21):3073.

[3] Du H, Fang C, Li Y, et al. Understandings and future challenges in soybean functional genomics and molecular breeding[J]. Journal of Integrative Plant Biology, 2023, 65(2): 468-495.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[5] Wu C, Zhang Y, Ying Z, et al. A transformer-based genomic prediction method fused with knowledge-guided module[J]. Briefings in Bioinformatics, 2024, 25(1): bbad438.

作者简介:

陈艾灵(1999--), 汉族, 四川人, 四川大学数学学院硕士研究生, 研究方向为数理统计。