[文章编号] 1007－7405(2023)03－0205－09 DOI:10.19715/j.jmuzr.2023.03.02

# 一种基于人工智能的基因组选择方法

顾林林[1,2]，王志勇[1,2]，方 铭[1,2]

（1. 集美大学水产学院，福建 厦门 361021；2. 农业农村部东海海水健康养殖重点实验室，福建 厦门 361021）

[摘要] 开发了一种新的深度学习基因组选择（genomic selection，GS）方法，并命名为DRNGS（deep residual network genomic selection）。新方法的特点有：1）以深度残差网络来预测基因组估计育种值（genomic estimated breeding value，GEBV），可捕获基因型内部的复杂关系，提高预测准确性；2）采用卷积和池化策略来降低高维基因型数据的复杂性，加快计算速度；3）方法中引入批量归一化层，加快了收敛速度。将新方法应用于CIMMYT小麦数据集，实验结果表明：DRNGS的效果比前馈神经网络（feedforward neural network，FNN）提高了101.59%～130.83%；在对大部分性状的表型预测中，DRNGS比GBLUP（genomic best linear unbiased prediction）提高了2.24%～20.19%；在计算耗时方面，DRNGS仅次于GBLUP，比DeepGS快了大约18～22倍，比FNN快了24～26倍。为进一步比较DRNGS和DeepGS，用伊朗面包小麦（*Triticum aestivum*）数据集进行测试，结果表明：DRNGS收敛速度优于DeepGS；在对所有性状的表型预测过程中，DRNGS的计算耗时始终较DeepGS短；而且DRNGS在预测准确性方面优于DeepGS，在8个性状中，DRNGS较DeepGS提高0.12%～1.59%。并将DRNGS开发成R包，可通过https://github.com/GuLinLin-JMU/DRNGS访问。

[关键词] 基因组选择；基因组估计育种值；神经网络；深度学习

[中图分类号] S 965.325；Q 348

# A Genomic Selection Method Based on Artificial Intelligence

GU Linlin[1,2], WANG Zhiyong[1,2], FANG Ming[1,2]

(1. Fisheries College, Jimei University, Xiamen 361021, China;
2. Key Laboratory of Healthy Mariculture for the East China Sea of Ministry of Agriculture, Xiamen 361021, China)

**Abstract:** This paper has improved and developed a new deep learning Genomic selection (GS) method named deep residual network genomic selectio (DRNGS). The features of the new method were: 1) a deep residual network was used to predict genomic estimated breeding value (GEBV), which could capture the complex relationships within genotypes and improve the prediction accuracy; 2) convolution and pooling strategies were used to reduce the complexity of high-dimensional genotype data and speed up the computation; 3) a batch normalization layer was introduced in the method to speed up the convergence speed. The new method applied the CIMMYT wheat dataset, and the experimental results showed that DRNGS outperformed Feedforward Neural Network (FNN) with a relative improvement of 101.59%-130.83%. DRNGS outperformed Genomic Best Linear Unbiased Prediction (GBLUP) by 2.24% to 20.19% for phenotypic prediction of most

[收稿日期] 2022－08－30

[基金项目] 国家重点研究发展计划项目（2018YFD0901201）；国家自然科学基金项目（31672399，31872560）；厦门市科技计划项目（2019SH400133）；福建省自然科学基金项目（2020J01672）；国家海水鱼产业技术体系岗位科学家项目项目（CARS－47－G04）

[作者简介] 顾林林（1994—），男，硕士生，从事统计遗传与基因组方向研究。通信作者：方铭（1979—），男，博士，教授，博导，从事鱼类遗传育种研究。E-mail:fangmign618@126.com

traits. It was the second only to GBLUP in terms of computational time consumption, and was approximately 18-22 times faster than DeepGS and 24-26 times faster than FNN. To further compare DRNGS with DeepGS, we applied the Iranian bread wheat (*Triticum aestivum*) dataset for testing and showed that DRNGS converged faster than DeepGS, consistently took less time to compute than DeepGS in predicting phenotypes for all traits, and that DRNGS outperformed DeepGS in terms of prediction accuracy. For eight traits, DRNGS improved 0.12%-1.59% over DeepGS. DRNGS has been developed as an R package, which can be accessed at https://github.com/GuLinLin-JMU/DRNGS.

**Keywords:** genomic selection; genomic estimation breeding value; neural network; deep learning

## 0 引言

基因组选择作为一种新的育种策略，应用广泛。常规的基因组选择先是通过混合线性统计模型对参考群体的基因型和表型数据建模，然后对只有基因型信息的候选群体预测基因组估计育种值（genomic estimated breeding value，GEBV）。它被应用于早期的育种选择，可以极大缩短育种周期、加快育种进展[1]。目前已经开发了多种基因组选择（genomic selection，GS）方法，有基因组最佳线性无偏预测（genomic best linear unbiased prediction，GBLUP）[2]、BayesA[1]、BayesB[1]、BayesCπ[3] 和 BayesR[4] 等方法。然而不同的统计方法通常需要做出假设并执行线性回归分析，这影响了它们的准确性。GBLUP 通常认为基因组所有的单核苷酸多态性（single nucleotide polymorphism，SNP）标记效应值的方差相等，并利用全基因组 SNP 构建个体间基因组关系矩阵（genomic relationship matrix，G matrix），还用 BLUP 推导所获得的混合模型方程组来估算遗传效应。而 BayesA、BayesB 和 BayesCπ 等贝叶斯方法则需要对不同 SNP 位点效应值的方差做出假设。各种贝叶斯方法之间的主要区别在于：为每个变量都分配超参数，不同的超参数可能会导致不同的结果。基于统计方法的基因组选择，不仅需要克服高维数据集所带来的挑战，即基因型标记 的数量（$p$）远大于种群规模（$n$）[5-8]，而且难以捕获基因型内部之间、基因型与表型之间的非线性关系[5,9]。

多层神经网络架构具有对许多高级特征良好的预测能力。在多层神经网络中，大量神经元被用于捕获大数据中复杂的非线性关系[10]。最近，深度学习也在生物学中开始应用，主要用于基因表达的推断[11-12]、遗传变异的功能注释[13-16]、蛋白质折叠的识别[17-18]和增强子的基因组预测[19]。这些在计算生物学和系统生物学领域的成功应用，证明了深度学习具有从生物学数据中学习复杂关系的强大能力[20-22]。同样深度学习在基因组选择领域的应用，也随之而来。2018 年，Ma 等[23]提出基于深度卷积神经网络的基因组选择方法 DeepGS，并通过实验证明了 DeepGS 在基因组选择中的有效性和稳健性。但是，DeepGS 在实际操作中存在收敛速度慢、耗费时间长等问题，这限制了它在基因组选择中的应用。因此，迫切需要新的应对方法来增强深度学习 GS，以及挖掘其在育种中的潜力。

## 1 深度学习基因组选择方法 DRNGS

本研究提出一种新的深度学习基因组选择方法——DRNGS（deep residual network genomic selection，DRNGS）。DRNGS 方法是基于深度残差网络构建的，包括输入层、卷积层、归一化层、激励层、残差层、池化层、完全连接层和输出层，其具体结构如图 1 所示。

输入层输入标记矩阵。卷积层实现对数据特征的学习，它由 8 个 $1\times18$ 大小的卷积核构成，滑动步长设置为 $1\times1$。卷积是一种特殊的积分变换，可简单理解为某一标记序列经过翻转和平移之后与卷积核序列的重叠部分函数值的乘积对重叠长度的积分。

随后经过归一化层（batch normalization，BN），把卷积层的输出都转换在均值为 0、方差为 1 的状态[24]，在训练过程中输出服从标准的正态分布，模型可以更快收敛。归一化层具体计算公式为：

$$\mu_{\mathrm{BN}}=\frac{1}{m}\sum_{i=1}^{m}x_i,\sigma_{\mathrm{BN}}^2=\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_{\mathrm{BN}})^2,\hat{x}_i=(x_i-\mu_{\mathrm{BN}}/\sqrt{\sigma_{\mathrm{BN}}^2+\varepsilon}),y_i=\gamma_i\hat{x}_i+\beta_i。$$

其中：$m$ 是上一层神经元的数目；$x_i$ 是上一层第 $i$ 个神经元的输出；$\varepsilon$ 为随机误差值；$y_i$ 为归一化层输出；$\gamma_i=\sqrt{\mathrm{var}(x_i)}$；$\beta_i=E(x_i)$。

将归一化层的输出馈送到激励层，激励函数为线性整流函数（rectified linear unit，ReLU），以学习非线性。

接下来是残差层的学习，本研究使用一层恒等残差块。残差块由一个卷积层、一个归一化层和两个激励层构成，具体操作是将上一层学习到的原始特征与经过卷积、非线性转换之后的特征进行融合，这不仅可以增加特征的多样性，提高模型的准确性，同时也可以防止因层数过多而导致过拟合。其中残差层的卷积核大小为 $1\times17$，核数为 8，采用“same”方式填充。残差层的激励函数也为 ReLU 函数，$\mathrm{ReLU}(x)=\begin{cases}x & \text{if } x>0,\\ 0 & \text{else}。\end{cases}$ 其中 $x$ 为激励层的输入。

把残差层学习到的数据特征输入到池化层，最大池化设置为 $1\times4$，步长也为 $1\times4$。最大池化是继卷积层之后进一步实现了数据的降维。最大池化之后与含有 64 个神经元的第一个连接层连接，将学习到的特征整合到一起，再经过归一化层和激励层后输入到第二个连接层，第二个连接层含有 1 个神经元。最后通过回归分析得到预测值。DRNGS 中的参数采用反向传播算法进行优化，学习率设为 0.01，迭代次数设为 500。损失采用平均绝对误差（mean absolute error，MAE）来计算。

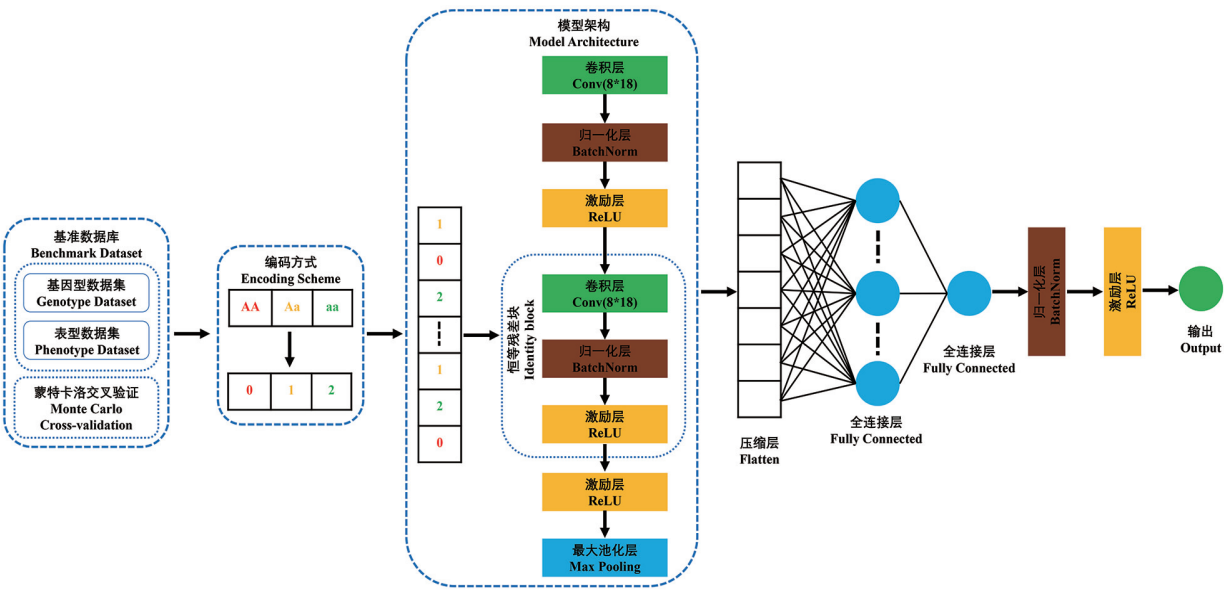



图 1　DRNGS 模型流程图

Fig.1　Flowchart of the DRNGS model

## 2　实验验证

### 2.1　蒙特卡洛交叉验证

使用蒙特卡洛交叉验证来评估新方法的预测准确性。首先将每个数据集的个体随机分成比例为 9∶1的两部分，分别作为参考群体和候选群体。重复该过程 100 次，以便更准确地比较每种方法。对于每个生成的数据集，用候选群体进行预测。在预测过程中其表型值被掩盖。本研究计算了候选群体的预测值和掩盖的表型值之间的皮尔逊相关系数，以评估每种方法的预测准确性。使用差异百分比来说明方法之间的增益程度。

### 2.2 数据集

本研究使用的 GS 数据来自 CIMMYT 的小麦数据集，包含 599 个个体，以及 4 个地点的小麦平均产量。数据集有四个对象：wheat. Y、wheat. A、wheat. X 和 wheat. set[25]。wheat. Y 为小麦两年平均产量；wheat. A 为同一家系的亲缘关系矩阵；wheat. X 为 DArT 标记矩阵。DArT 标记中，1 个等位基因分别被编码为 1 或 0，表示存在（1）或不存在（0）该等位基因。数据来源于 https://www. diversityarrays. com。本研究使用了三种具有代表性的基因组选择方法（GBLUP、FNN、DeepGS）与 DRNGS 进行对比。

为了进一步比较 DeepGS 和 DRNGS 两个基因组选择方法，本研究还使用了伊朗面包小麦数据集加以验证。该数据集由 2000 个个体的伊朗面包小麦组成，每个个体包含 33 709 个 DArT 标记。数据集包含了 8 个性状：谷物长度（grain length，GL）、谷物宽度（grain width，GW）、谷物硬度（grain hardness，GH）、千粒质量（thousand-mernel mass，TKM）、测试质量（test mass，TM）、十二烷基硫酸钠-沉降（sodium dodecyl sulphate sedimentation，SDS）、谷物蛋白质（grain protein，GP）和植物高度（plant height，PHT）。完整的数据可以从 http://genomics. cimmyt. org/mexican_iranian/traverse/iranian/standarizedData_univariate. RData 获得[26]。

## 3 实验结果与分析

### 3.1 DRNGS 的预测效果

本研究首先用 CIMMYT 的小麦数据集来验证 DRNGS 的性能。小麦数据集包含 599 个个体，每个个体都用了 1279 个 DArT 标记，表型记录为不同地点的小麦平均产量（grain yield，GY）。表 1 列出了 100 次交叉验证的平均预测准确性，DRNGS 的预测准确性高于其他三种方法。在第一个地点中，DRNGS 的预测准确性最高（$r$ = 0.5334），FNN 的最低（$r$ = 0.2646），GBLUP 的为 0.4438，DeepGS 的为 0.5126。DRNGS 比其他三种方法提高了 4.06% ~101.59%。在第二个地点中，DRNGS 始终比其他三种方法更准确（$r$ =0.4605），相比于其他三种方法提高了 2.24% ~130.83%。在第三个地点中，预测准确性 DRNGS（$r$ =0.3955）优于 FNN（$r$ =0.1908）和 DeepGS（$r$ =0.3779），但是稍低于 GBLUP（$r$ =0.4085）。在第四个地点中，DRNGS（$r$ =0.4946）的预测准确性最高，比 FNN（$r$ =0.2387）提高了 107.21%，比 GBLUP（$r$ =0.4529）和 DeepGS（$r$ =0.4863）分别提高了 9.21% 和 1.71%。总体而言，DRNGS 的预测效果较其他三种基因组选择方法更优。

**表 1 DRNGS 与其他 GS 方法的预测准确性及模型最大差异百分比**

Tab. 1 The predictive accuracies and the maximum difference percentage of other GS methods and DRNGS

| 方法 Method | 预测准确性 Predictive accuracy | | | |
|---|---|---|---|---|
| | 产地 1 Environment 1 | 产地 2 Environment 2 | 产地 3 Environment 3 | 产地 4 Environment 4 |
| DRNGS | 0.5334 ±0.0124 | 0.4605 ±0.0109 | 0.3955 ±0.0127 | 0.4946 ±0.0126 |
| DeepGS | 0.5126 ±0.0135 | 0.4504 ±0.0118 | 0.3779 ±0.0131 | 0.4863 ±0.0118 |
| FNN | 0.2646 ±0.0129 | 0.1995 ±0.0125 | 0.1908 ±0.0127 | 0.2387 ±0.0128 |
| GBLUP | 0.4438 ±0.0105 | 0.4504 ±0.0118 | 0.4085 ±0.0097 | 0.4529 ±0.0110 |
| 最大差异的百分比 Maximum difference percentage | 0.2688 (101.6%) | 0.2610 (130.8%) | 0.2177 (114.1%) | 0.2559 (107.2%) |

表 2 列出了 100 次交叉验证的平均预测准确性，DRNGS 的预测准确性普遍优于 DeepGS。在 PHT 预测中，DRNGS 的预测准确性为 0.3076，优于 DeepGS（$r$ =0.3028），相对提高了 1.59%。在 GH 预

测中，DRNGS 的预测准确性（$r=0.6771$）优于 DeepGS（$r=0.6728$），相对提高了 0.64%。同理，在 GL、SDS、TM、TKM 的预测过程中，DRNGS 相对于 DeepGS 分别提高了 0.54%、0.22%、0.36%、0.12%。但对 GP 和 GW 的预测，DRNGS 略差于 DeepGS。

**表 2　DRNGS 与 DeepGS 的预测准确性及最大差异百分比**

Tab. 2 The predictive accuracies and the maximum difference percentage of DeepGS and DRNGS

| 性状 Trait | 预测准确性 Predictive accuracy | | 最大差异百分比 Maximum difference percentage |
|---|---|---|---|
| | DeepGS | DRNGS | |
| 谷物硬度 GH | 0.6728 ± 0.0043 | 0.6771 ± 0.0040 | 0.0064（0.64%） |
| 谷物长度 GL | 0.7391 ± 0.0037 | 0.7431 ± 0.0039 | 0.004（0.54%） |
| 植物高度 PH | 0.3028 ± 0.0067 | 0.3076 ± 0.0073 | 0.0048（1.59%） |
| 谷物蛋白质 GP | 0.5404 ± 0.0046 | 0.5343 ± 0.0048 | 0.0061（1.14%） |
| 十二烷基硫酸钠 – 沉降 SDS | 0.5032 ± 0.0058 | 0.5043 ± 0.0059 | 0.0011（0.22%） |
| 测试质量 TM | 0.6154 ± 0.0044 | 0.6176 ± 0.0044 | 0.0022（0.36%） |
| 千粒质量 TKM | 0.6484 ± 0.0032 | 0.6492 ± 0.0032 | 0.0008（0.12%） |
| 谷物宽度 GW | 0.7249 ± 0.0033 | 0.7246 ± 0.0032 | 0.0003（0.04%） |

## 3.2　DRNGS 收敛速度优于 DeepGS

在 DeepGS 的基础上，DRNGS 做了进一步的改进，增加了归一化层，加快了方法的收敛速度，节省了训练时间。本研究先用 CIMMYT 的小麦数据集来验证 DRNGS 的收敛性能。图 2 是对 DeepGS 和 DRNGS 收敛性的记录，横坐标为迭代次数，纵坐标为两种方法在每次迭代中所计算的平均绝对误差。

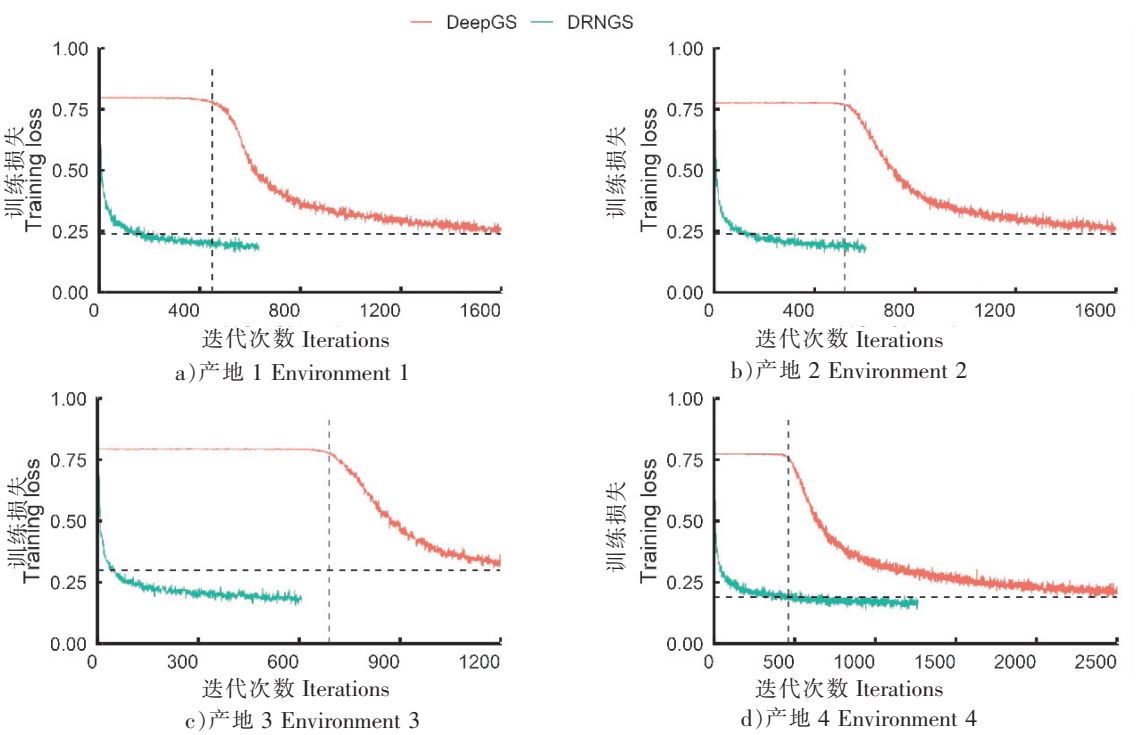



**图 2　DeepGS 和 DRNGS 在 CIMMYT 小麦数据集中收敛性能的比较**

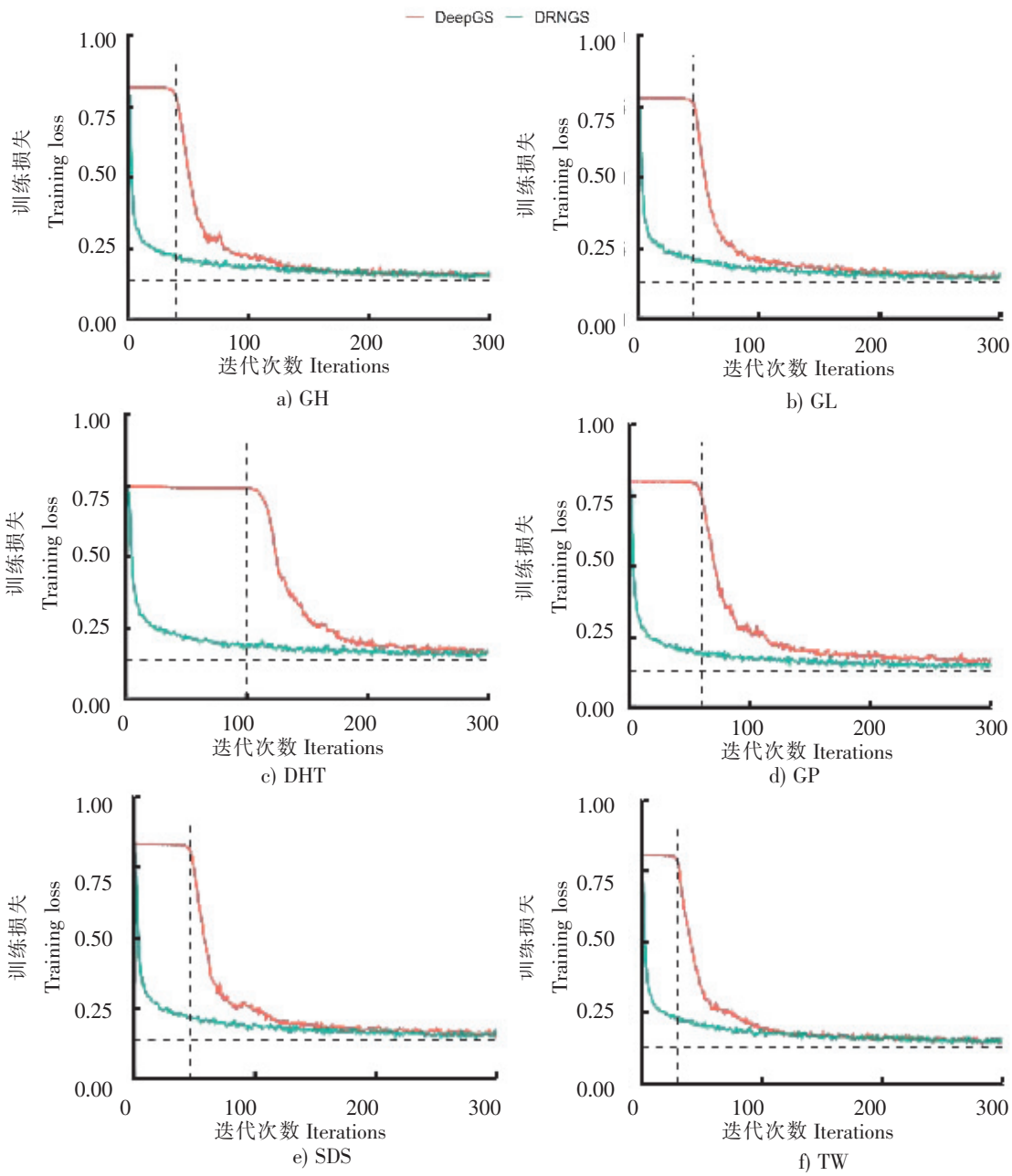
Fig.2　Comparison of convergence performances of DeepGS and DRNGS in CIMMYT wheat datasets

在第一个地点的产量中，DRNGS 在迭代训练初始就快速收敛直至平衡；而 DeepGS 在迭代训练初期 MAE 维持在 0.8 左右，大约到 500 次才开始下降，需要迭代 1500 多次才能达到收敛标准，而且所拟合方法的 MAE 高于 DRNGS（见图 2a）。在其他三个地点的产量中，DRNGS 同样在迭代训练初始就

快速收敛直至平衡，而 DeepGS 在迭代初期 MAE 会维持在一个较高的水平。相同的条件下，DRNGS 比 DeepGS 更快达到收敛标准，而且所拟合的 DRNGS 方法优于 DeepGS，不仅节省了模型的训练时间，而且提高了育种选择的效率（见图 2b ~ 图 2d）。

为了进一步比较 DRNGS 和 DeepGS 之间收敛性能的差别，本研究使用了伊朗面包小麦数据集来进行验证。同样地，针对 DeepGS 和 DRNGS 的收敛性做了记录，如图 3 所示，横坐标为迭代次数，纵坐标为训练迭代的平均绝对误差。

在所有性状中，DRNGS 在迭代训练初始就快速收敛直至平衡，而 DeepGS 在迭代训练初期 MAE 会保持不变。图 3 中显而易见，预测 PHT 性状时，DeepGS 迭代 100 次才开始缓慢下降，需要更多次迭代才能达到收敛标准，这就导致训练时间加长，而育种讲究的是时效性。此外，在对其他 7 个性状的表型预测过程中，DeepGS 在迭代初期 MAE 会维持在一个较高的水平，并不能像 DRNGS 那样在迭代训练初始就快速收敛直至平衡。同样的条件下，DRNGS 比 DeepGS 更快达到收敛标准。



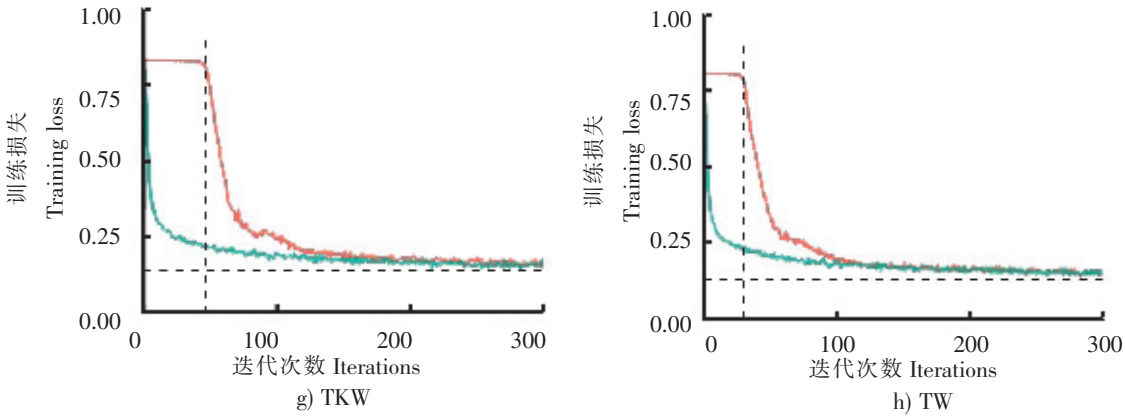



图 3　DeepGS 和 DRNGS 在伊朗小麦数据集中收敛性能的比较

Fig.3　Comparison of convergence performances of DeepGS and DRNGS in Iranian wheat datasets

### 3.3　模型计算耗时的比较

为了解 DRNGS 在计算速度上的优势，本研究对 CIMMYT 的小麦数据集计算耗时做了记录。所有的程序均运行于配置为 1.95 GHz、Hygon C86 7151 16-core Processor CPU 和 500GB 内存的 CentOS Linux 服务器上。4 种方法的 100 次交叉验证总的耗费时间（$t$）如图 4 所示。可以很直观地看到在所有性状中，前馈神经网络 FNN 的耗时最长，然后依次是 DeepGS、DRNGS、GBLUP。其中 DeepGS 和 DRNGS 虽同为深度学习技术，但 DRNGS 比 DeepGS 快了约 18 ~ 22 倍。DRNGS 相对于 FNN 快 24 ~ 26 倍。总体来说，除了 GBLUP 外，在计算速度上 DRNGS 占据了明显的优势，而且 DeepGS 和 DRNGS 在计算时间上相差了至少两个以上的数量级。

此外，本研究还对伊朗面包小麦数据集的 8 个性状分别记录了 DeepGS 和 DRNGS 的计算耗时。所有程序均用 R 语言代码编写，并运行于 Linux 服务器上。DeepGS 和 DRNGS 的计算时间如图 5 所示，$x$ 轴为不同的性状，$y$ 轴表示计算耗费的时间。在所有的性状预测过程中，DRNGS 的计算耗时始终较 DeepGS 短，大约是 DeepGS 的 1/3。由图 5 易见，在对不同性状的表型预测过程中，DeepGS 的耗时并不十分稳定（59.47 ~ 77.7 min），反之，DRNGS 对所有性状表型预测的耗时都相对稳定（20.80 ~ 20.84 min）。总体来说，同属于深度学习方法的 DeepGS 和 DRNGS，DeepGS 需要消耗的计算时间是 DRNGS 计算时间的 3 ~ 4 倍。

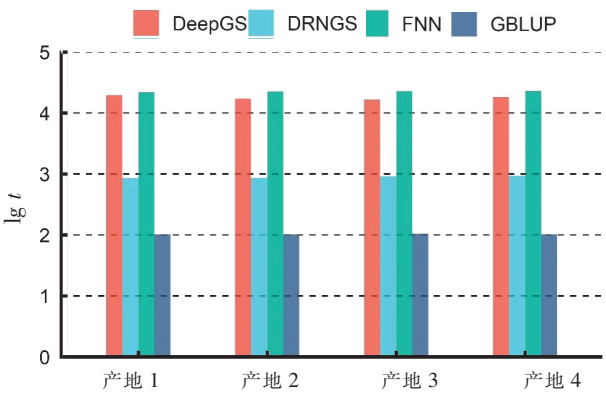



图 4　GBLUP、FNN、DeepGS 和 DRNGS 对 CIMMYT 小麦数据集的计算耗时

Fig.4　The computing times (in seconds) of GBLUP, FNN, DeepGS and DRNGS for CIMMYT wheat datasets

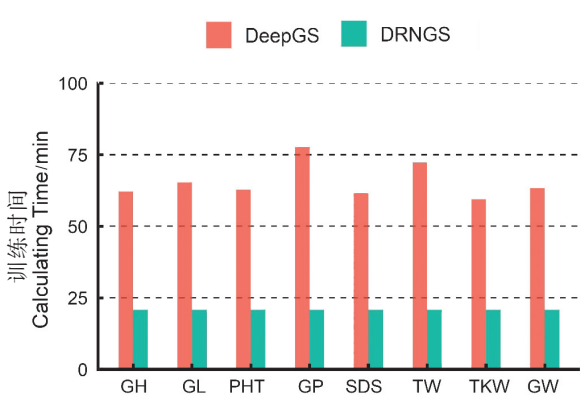



图 5　DeepGS 和 DRNGS 对伊朗小麦数据集的计算耗时

Fig.5　The computing times (in minutes) of DeepGS and DRNGS for Iranian wheat datasets

## 4 讨论

相比于传统的育种来说，基因组选择优势显著，而新的育种方法对于准确预测表型尤为重要。深度学习是近期开发的机器学习技术，它能够捕获变量之间的非线性复杂关系。深度学习技术已经在系统生物学和结构生物学方面应用广泛，目前针对基因组选择领域还处于起步阶段。本研究将深度学习技术应用于基因组选择，并提出了基于深度残差网络的基因组选择方法 DRNGS。本研究已使用多份小麦的经济性状数据集来验证 DRNGS 的性能，所有结果都显示，相对于 GBLUP、FNN、DeepGS 三种基因组选择方法而言，DRNGS 的效果更优。在预测 CIMMYT 的小麦数据集时，DRNGS 相对于另三种方法显示出更大的优势，对于大部分性状的表型预测，DRNGS 都比其他方法更准确。

DRNGS 比 FNN 增益明显，这表明 FNN 不太适合做基因组选择，其中原因可能是：1）FNN 为半监督型的机器学习方法，内部所涉及的超参数对模型的影响结果比较大；2）FNN 各层之间无反馈信息，无法充分学习各个变量之间的复杂关系。

DRNGS 相对于 Ma 等[23]开发的基因组选择方法 DeepGS 而言，有以下优势：1）加入了归一化层，把每层的数据都转换在均值为 0、方差为 1 的状态，这样每层数据的分布都是一样的，训练更容易收敛；2）在常规的深度神经网络中，如果网络的激活输出很大，其对应的梯度就会很小，导致网络的学习速率就会很慢，而 DRNGS 的输出不会很大，梯度也就不会很小，成功地解决了梯度消失的问题；3）在常规的深度神经网络中，第一层偏移量的梯度等于所有激活层斜率与权值的乘积，假如激活层斜率均为最大值 0.25，所有层的权值为 100，这样梯度就会呈指数增加，即梯度爆炸，而 DRNGS 权值的更新不会很大，也就不存在梯度爆炸的问题；4）在网络的训练中，归一化层的使用使得一个批次中所有样本都被关联在一起，因此网络不会从某一个训练样本中生成确定的结果，即同样一个样本的输出不再仅仅取决于样本的本身，还取决于跟这个样本同属一个批次的其他样本，而每次网络都是随机抽取批次，这样就会使得整个网络不会朝一个方向使劲学习。DRNGS 在一定程度上避免了过拟合问题。

然而，DRNGS 还存在一些限制。比如更加合适的网络结构对于预测性能尤为重要，如网络中的超参数，网络的层数，卷积层、激励层、归一化层、残差层之间的连接方式，它们都会影响模型的预测性能，这也是深度学习在计算生物学和生物信息学应用中的限制。近期在计算机视觉领域所开发的网络可视化系统有望解决这个难题。

总而言之，相对于 DeepGS 而言，DRNGS 在计算速度和预测准确性上有一个很大的改变。DRNGS 不仅预测准确性高，而且计算速度快。日后会继续改进 DRNGS 存在的问题，最终实现它在 GS 领域的广泛应用。

## [ 参 考 文 献 ]

[1]MEUWISSEN T H E, HAYES B J, GODDARD M E. Prediction of total genetic value using genome-wide dense marker maps [J]. Genetics, 2001, 157(4): 1819-1829.

[2]Van RADEN P M. Efficient methods to compute genomic predictions[J]. Journal of Dairy Science, 2008, 91(11): 4414-4423.

[3]HABIER D, FERNANDO R L, KIZILKAYA K, et al. Extension of the bayesian alphabet for genomic selection[J]. BMC Bioinformatics, 2011, 12(1): 186.

[4]ERBE M, HAYES B J, MATUKUMALLI L K, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels[J]. Journal of Dairy Science, 2008, 95(7): 4114-4129.

[5]CROSSA J, PÉREZ-RODRÍGUEZ P, CUEVAS J, et al. Genomic selection in plant breeding: methods, models, and perspectives[J]. Trends in Plant Science, 2007, 22(11): 961-975.

[6]DESTA Z A, ORTIZ R. Genomic selection: genome-wide prediction in plant improvement[J]. Trends in Plant Science, 2014,19(9):592-601.
[7]JANNINK J L,LORENZ A J,IWATA H. Genomic selection in plant breeding: from theory to practice[J]. Briefings in Functional Genomics, 2010,9(2):166-177.
[8]SCHMIDT M,KOLLERS S,MAASBERG-PRELLE A,et al. Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection[J]. Theoretical and Applied Genetics, 2016,129(2):203-213.
[9]Van EEUWIJK F A,BINK M C,CHENU K,et al. Detection and use of QTL for complex traits in multiple environments[J]. Current Opinion in Plant Biology, 2010,13(2): 193-205.
[10]LE CUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature, 2015,521(7553):436-444.
[11]CHEN Y,LI Y,NARAYAN R,et al. Gene expression inference with deep learning[J]. Bioinformatics, 2016,32(12):1832-1839.
[12]SINGH R,LANCHANTIN J,ROBINS G,et al. DeepChrome: deep-learning for predicting gene expression from histone modifications[J]. Bioinformatics, 2016,32(17):i639-i648.
[13]QUANG D,CHEN Y,XIE X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants[J]. Bioinformatics,2015,31(5):761-763.
[14]QUANG D,XIE X. Dan Q: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences[J]. Nucleic Acids Research, 2016,44(11):e107-e107.
[15]XIONG H Y,ALIPANAHI B,LEE L J,et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science,2015,347(6218):1254806-1254806.
[16]ZHOU J,TROYANSKAYA O G. Predicting effects of noncoding variants with deep learning-based sequence model[J]. Nature Methods, 2015,12(10):931-934.
[17]JO T,HOU J,EICKHOLT J,et al. Improving protein fold recognition by deep learning networks[J]. Scientific Reports, 2015,5(1):17573.
[18]WANG S,PENG J,MA J,et al. Protein secondary structure prediction using deep convolutional neural fields[J]. Scientific Reports,2016,6(1):18962.
[19]KELLEY D R,SNOEK J,RINN J L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks[J]. Genome Research,2016,26(7):990-999.
[20]ANGERMUELLER C,PÄRNAMAA T,PARTS L,et al. Deep learning for computational biology[J]. Molecular Systems Biology,2016,12(7):878.
[21]CAO Y,GEDDES T A,YANG J Y H,et al. Ensemble deep learning in bioinformatics[J]. Nature Machine Intelligence, 2020,2(9): 500-508.
[22]MIN S,LEE B,YOON S. Deep learning in bioinformatics[J]. Brief Bioinform,2017,18(5):851-869.
[23]MA W,QIU Z,SONG J,et al. A deep convolutional neural network approach for predicting phenotypes from genotypes[J]. Planta, 2018,248(5):1307-1318.
[24]IOFFE S,SZEGEDY C. Batchnormalization: accelerating deep network training by reducing internal covariate shift[J]. Journal of Machine Learning Research,2015:448-456.
[25]MCLAREN C G,BRUSKIEWICH R M,PORTUGAL A M,et al. The international rice information system:a platform for meta-analysis of rice crop data[J]. Plant Physiology,2005,139(2):637-642.
[26]CROSSA J,JARQUÍN D,FRANCO J,et al. Genomic prediction of gene bank wheat landraces[J]. G3 Genes|Genomes|Genetics, 2016,6(7):1819-1834.
[27]SRIVASTAVA N,HINTON G,KRIZHEVSKY A,et al. Dropout: a simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014,15(1):1929-1958.

(责任编辑　朱雪莲　英文审校　黄力行)