

DOI: 10.3724/SP.J.1006.2011.00000

人工神经网络在作物基因组选择中的应用

束永俊 吴磊 王丹 郭长虹*

哈尔滨师范大学生命科学与技术学院 / 黑龙江省分子细胞遗传与遗传育种重点实验室, 黑龙江哈尔滨 150025

摘要: 目前, 基因组选择育种主要采用线性模型估计遗传育种值指导作物遗传育种的筛选过程, 但是生物体内的基因以及遗传位点的关系主要是复杂的非线性调控。本研究将人工神经网络技术应用到作物基因组选择育种中, 对现有的作物基因组选择育种模型进行优化, 建立了高效的作物基因组选择预测系统, 并与其他线性回归预测模型进行比较。通过分析小麦的育种数据发现, 基于人工神经网络的遗传育种估计效果优于其他线性回归预测模型, 预测育种值与实际育种值间的相关系数平均值达到 0.6636, 相应的岭回归 BLUP、贝叶斯线性回归模型和基于系谱信息的贝叶斯回归模型的预测能力分别为 0.6422、0.6294 和 0.6573; 最优的预测效果达到 0.8379, 远高于其他 2 种模型的最优结果。同时, 基于人工神经网络的基因组选择模型的预测效果稳定, 与传统的统计模型相近, 因此, 利用人工神经网络技术建立基因组选择是可行的。

关键词: 基因组选择; 小麦; 人工神经网络; 岭回归 BLUP; 贝叶斯线性回归

Application of Artificial Neural Network in Genomic Selection for Crop Improvement

SHU Yong-Jun, WU Lei, WANG Dan, and GUO Chang-Hong*

Key Laboratory of Molecular Cytogenetics and Genetic Breeding of Heilongjiang Province / College of Life Science and Technology, Harbin Normal University, Harbin 150025, China

Abstract: With important progress in marker technologies, marker-assisted selection (MAS) has been used broadly for the crop improvement. Biparental populations are designed for the detection of quantitative trait loci (QTLs), but their application is retarded. The association mapping (AM) is applied directly to natural populations, which has been proposed to mitigate the lack of relevance of biparental populations in QTL identification. Many QTLs are identified by the two methods, which have encouraged genetic improvement of crop. However, they are using significant thresholds to identify QTL from estimated means that estimated effects are biased. Therefore, small-effect QTLs can't be identified and missed entirely, while lots of traits of crop are controlled by those small-effect QTLs. Genomic selection (GS) has been proposed to make good for these deficiencies. Genomic selection predicts the breeding values of lines in a population by analyzing their phenotypes and high-density marker scores, and by including all markers in the model, and benefits from unbiased estimation of all chromosome segment effects, even when they are small. The GS incorporates all marker information in the prediction model, which avoids biased marker effect estimates and captures more of the variation from small-effect QTLs. Furthermore, markers carry information on the relatedness among lines, which contributes to prediction accuracy. Such accuracies are sufficient to select parents strictly on the basis of marker scores even for traits such as yield, tolerance to abiotic stress. From the view of perspective of the products from plant breeding, the genomic selection would greatly accelerate the breeding cycle, and enhance annual gains. GS would develop a prediction model from training population, genotyped and phenotyped, by estimated the markers effects. Then GS model would take genotypic data from candidate population to predict genomic estimated breeding values (GEBV), and there are some methods used for GS model, such as best linear unbiased prediction (BLUP), ridge regression BLUP (RR-BLUP), and Bayesian linear regression (BLR). These models are well used for crop genomic selection breeding. However, all the models are developed based on line or regression, while the relationships of genetic sites in life are not non-line or regression. The neural network was introduced to genomic selec-

本研究由哈尔滨师范大学青年骨干教师基金(KGB201010), 哈尔滨师范大学科技创新团队(KJTD201102), 国际科技合作项目(2009DFA32470), 国家重点基础研究计划(973 计划)前期项目(2011CB111505)和黑龙江省科技攻关项目(GA08B104)资助。

* 通讯作者(Corresponding author): 郭长虹, Email: kaku3008@yahoo.com.cn, Tel: 0451-88060576

第一作者联系方式: E-mail: syjun2003@126.com

Received(收稿日期): 2011-05-30; Accepted(接受日期): 2011-09-09; Published online(网络出版日期): 2011-10-09.

URL: <http://www.cnki.net/kcms/detail/11.1809.S.20110929.1552.015.html>

tion in crop improvement in this study. The crop genomic selection model was optimized by non-linear model system. Therefore, the high efficient genomic selection system was established, and the prediction results were compared with these of other linear models, such as RR-BLUP, BLR. In wheat genetic data simulation, the correlation coefficient between the true breeding value of unphenotyped experimental lines and that predicted by genomic selection based on the neural network reached 0.6636, while that of RR-BLUP, BLR and BLR with pedigree information was 0.6422, 0.6294, and 0.6573, respectively. Meanwhile, the best prediction was 0.8379, which indicated the genomic selection based on the neural network is superior to these of other linear regression models. This level of accuracy was sufficient for selecting for agronomic performance using marker information alone. Such selection would substantially accelerate the breeding cycle, and enhance gains per unit time. Therefore, this research showed that GS has more potential for incorporating it into breeding schemes.

Keywords: Genomic selection; Wheat; Artificial neural network; Ridge regression BLUP; Bayesian linear regression

在育种中,一般根据个体的表型性状和个人经验,选育出具有优良性状的后代,在传统的选育过程中,主要依靠表型性状选择。但是,由于可供筛选的性状数据有限,往往需要淘汰大量个体,或者从大量的候选样本中选择具有较强优势性状的个体。整个选育过程可参考的因素有限,对育种实践经验有很大信赖。由于动物具有生长周期长,个体培养成本高等特点,动物育种学家开展经验育种压力较大,因此开发了基于动物表型数据的育种模型,即最适线性无偏估计模型(best linear unbiased prediction, BLUP)^[1]。该模型在动物,特别是一些家畜育种中取得了巨大的进展^[2-3]。近年来,一些植物育种学家对该模型进行了改进,将其应用到作物^[4-8]和林木^[9]育种上,选择效果明显^[10]。然而,生物个体的表型是基因型和环境共同作用下形成的,在选育过程中极易受到环境因素的影响。随着分子生物学技术的发展,分子标记技术逐渐应用到动植物遗传育种领域,分子标记能直接检测个体的遗传物质组成,不受环境因素的影响,因此,可以通过筛选分子标记来筛选个体中的优异基因,开展分子辅助育种(marker-assisted selection, MAS)^[11]。

MAS 主要是寻找一些控制数量性状的位点,即 QTL,利用 QTL 的信息对育种个体选育;同时,对控制数量性状的基因作单基因分解、精细定位,甚至是图位克隆,获得控制个体性状的主效基因^[12]。该技术在植物,特别是作物遗传育种领域取得了巨大成功。研究人员通过双亲本杂交构建了大量的作图和定位群体,定位了大量控制作物农艺性状的 QTL,育成了应用于生产实践的优异品种。但是 QTL 分析本身存在一些缺陷,(1)QTL 分析群体一般来源于双亲本的杂交,其遗传代表性具有一定的局限,推广的潜力有限,在一个群体中发现到某个效应显著的 QTL,在其他群体中往往检测不到,对作物育种实践的指导作用较弱。(2)QTL 分析是建立在许多假设的基础上,如无基因型和环境互作、无上

位性效应、无一因多效等,即基因或者遗传因子控制性状是无偏的,而实际情况往往不符合这些假设,基因或遗传因子之间存在复杂的调控关系。(3) QTL 分析针对的是数量性状的主效位点^[13-14]。事实上,很多作物的性状并不只由主效基因或者遗传因子控制,而是由大量微效基因协同作用控制,对于这些微效多基因控制的性状, QTL 分析就无法解析。近年来,随着分子标记检测成本的迅速下降,开发了一种新的分子辅助育种方法,即关联分析(association mapping, AM)。它不需要构建作图群体,简化了 QTL 定位分析的过程,也在一些作物育种领域取得了成功^[15-17]。但是 AM 仍没有解决控制性状位点间的互动问题,只是简单做一些线性组合的修正,预测和估计能力的提高程度有限,无法从根本上解决大多数性状是由微效多基因控制这一遗传育种实践问题。

针对微效多基因控制性状这一育种问题,应该根据基因位点作用的效应程度对这些基因位点进行赋值,然后在全基因组范围内估测育种性状值,选育品种个体,即基因组选择(genomic selection)^[18]。基因组选择最早出现在家畜的遗传育种过程,如牛、猪等^[19-23],近些年开始应用到作物和林木遗传育种领域^[24-25],如小麦、玉米、松树^[9]、油棕榈^[26]等。对基因组选择也发展了多种模型,如最适无偏估计(BLUP)、岭回归分析(ridge regression, RR)以及贝叶斯估计(Bayesian estimation, BE)^[27-29]等,用于不同情况的育种分析。控制性状的多基因位点间作用方式比较复杂,采用传统的线性估计和参数估计,预测效果比较有限,目前主要处于研究阶段^[30]。

人工神经网络(artificial neural network, ANN)也简称为神经网络(NN),是一种模仿人类神经网络行为特征,进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度,通过训练和学习过程,分析掌握事物内部潜在的规律,调整内部大量节点之间相互连接的关系,从而达到处理信息的目的。人工神经网络是由大量处理单元互联组成

的非线性、自适应信息处理系统, 在处理复杂的线性和非线性组合关系时具有巨大的优势, 目前已经广泛地应用于社会生产的各个领域。本研究将人工神经网络技术应用到作物基因组选择育种中, 现有模型, 建立高效作物基因组选择预测系统, 并与其他预测模型优化比较, 分析人工神经网络的基因组选择模型特点, 探讨其应用范围。

1 材料与方法

1.1 实验数据

国际玉米小麦改良中心(CIMMYT)在小麦全球育种计划中选育了 599 个优异品系或株系(http://cropwiki.irri.org/icis/index.php/TDM_GMS_Browse), 根据小麦品系的系谱信息计算小麦品系相关系数, 构建系谱遗传效应参数矩阵。利用 DArT 芯片技术(Triticarte Pty. Ltd., Canberra, Australia, <http://www.triticarte.com.au>)分型研究所有品系, 获得 599 个小麦品系的分型数据, 小麦品系的产量数据参见 Pérez 等^[29]文献。

1.2 统计模型

1.2.1 混合模型(Mixed Model) 利用该模型进行全基因关联分析筛选, 模型为:

$$y = u + \beta x + \varepsilon \quad (1)$$

其中 y 是品系的表型性状, u 是固定表型值, x 是品系个体的基因型, β 是品系的遗传效应值, ε 是模型误差。利用广义最小二乘法进行迭代, 对分子标记的遗传效应 β 进行估计, 去除低效应或没有效应的分子标记数据, 在尽可能保持方差变异的情况下, 对数据进行降维处理^[31]。

1.2.2 岭回归 BLUP 模型(Ridge Regression BLUP, RR-BLUP) 在基因组选择中, 广义的线性回归模型可以用如下矩阵形式表示。

$$y = X\beta + \varepsilon \quad (2)$$

其中 $y = \{y_i\}$, $X = \{x_{ij}\}$, $\beta = \{\beta_i\}$, $\varepsilon = \{\varepsilon_i\}$ 。由于需要估计的参数较多, 而样本量较小, 采用最小二乘估计时就会出现较大误差。当自变量间存在多重共线性时, 可以采用岭回归模型估计, 即:

$$[X'X + \lambda I]\hat{\beta} = X'y, \text{ 即 } \hat{\beta} = [X'X + \lambda I]^{-1}X'y \quad (3)$$

岭回归模型将通过 λ , 对遗传效应较低参数进行优化, 加速参数估计速度。

岭回归分析和产量性状预测过程采用吉布斯抽样方法进行, 每次随机选择 499 个品系的遗传表型数据和产量性状数据进行岭回归分析, 然后用岭回归模型对剩余的 100 个品系进行产量性状预测, 计算岭回归模型预测值与实际表型值间的相关系数,

对预测效果进行相关性评价。上述过程重复 1 000 次, 计算预测育种值与实际育种值间相关系数的均值和方差, 分析过程全部在 R 平台的程序包 rrBLUP 下完成。

1.3 贝叶斯线性回归模型(Bayesian Linear Regression, BLR)

岭回归模型假设各个分子标记的效应值完全一样, 即 λI , 但这与实际不符。因此, 研究人员提出贝叶斯估计, 利用先验概率校准线基因组选择的性回归模型。在岭回归模型的基础上, 结合贝叶斯 LASSO 方法(Least Absolute Shrinkage and Selection Operator, LASSO)对分子标记的遗传效应进行修正, 修正后的模型为:

$$y = 1 * \mu + X_F \beta_F + Z u + X_R \beta_R + X_L \beta_L + \varepsilon \quad (4)$$

其中, y 是表型响应值; μ 是截距值; β_F 、 β_R 和 β_L 分别是分子标记的固定遗传效应、岭回归(RR)估计遗传效应和 LASSO 估计遗传效应; X_F 、 X_R 和 X_L 分别为相应的关联矩阵; $Z u$ 是根据品种系谱信息建立的遗传效应矩阵, 去除该参数即为不含系谱信息的 BLR 模型; ε 是整个模型的残差。该模型可以综合利用岭回归和 LASSO 对分子标记的遗传效应进行修正, 同时采用系谱信息对表型值进行修正, 极大地提高模型预测的准确性。

利用 R 平台的程序包 BLR (<http://cran.r-project.org/web/packages/BLR/index.html>)进行贝叶斯回归估计, 模型训练过程采用吉布斯抽样方法, 每次随机选择 499 个品系, 结合品系的系谱信息, 建立贝叶斯回归模型, 然后利用参数优化的贝叶斯回归模型对剩余的 100 个品系预测产量性状, 计算贝叶斯回归模型预测值与实际表型值间的相关系数, 评价预测效果相关性。重复上述过程 1 000 次, 计算预测育种值与实际育种值间相关系数的均值和方差。另外, 贝叶斯回归也可以不用各个品系的系谱信息进行初始化, 因此, 本研究去除小麦品种的系谱信息, 重复上述的模型分析过程, 比较两者之间的预测效率。

1.4 人工神经网络模型(Artificial Neural Network, ANN)

人工神经网络的广义模型为:

$$\begin{aligned} y(x) &= f[w_0 + \sum_{j=1}^J w_j \cdot f(w_{0j} + \sum_{i=1}^n w_{ij} x_i)] \\ &= f[w_0 + \sum_{j=1}^J w_j \cdot f(w_{0j} + W_j^T X)] \end{aligned} \quad (5)$$

其中, $y(x)$ 是表型性状, 是性状在各个神经元的初始

值, J 是分子标记数目, w_{ij} 是各个神经元的权值系数, x_i 个体的基因型。

人工神经网络的初始连接参数采用随机分配的方法确定, 然后采用训练学习的方法不断优化人工神经网络各个神经元的权值系数和相应的基准值 (w_{ij} 、 w_0 和 w_{0j}), 使观测值与样本值间逐渐吻合的过程, 可以通过监测两者之间的误差评估训练结果, 误差计算模型为:

$$E = \frac{1}{2} \sum_{l=1}^L (y_l' - y_l)^2 \quad (6)$$

其中, $l = 1, 2, \dots, L$ 是训练样本, y_l' 是样本 l 的估计值, y_l 是样本 l 的真实值。通过不断优化人工神经网络参数, 使得 E 逐渐变小, 直到达到人工神经网络的训练要求。训练过程采用共轭梯度反馈算法加速网络的搜索速度, 大大提高人工神经网络训练的效率。

利用 Matlab 的人工神经网络工具包(The Math Works, Inc., R2008b, Neural Network Toolbox Version 6.0.1)建立双隐层人工神经网络(图 1), 隐层神经元转移函数选用 LOGSIG 和 PURELIN。采用吉布斯抽样方法, 每次随机选择 499 个品系的遗传数据和产量数据训练人工神经网络, 重复训练过程 30 次, 根据训练过程中预测数据和实际育种值间的相关性, 选择训练效果较好的 5 个人工神经网络, 对剩余的 100 个品系进行产量性状预测, 计算人工神经网络方法的预测值与实际表型值间的相关系数, 对预测效果进行相关性评价。重复上述过程 1 000 次, 计算预测育种值与实际育种值间相关系数的均值和方差。

通过吉布斯抽样计算预测育种值与实际育种值间的相关系数, 比较岭回归 BLUP 模型、贝叶斯线性回归模型、贝叶斯线性回归结合系谱信息模型和人工神经网络方法预测效果。

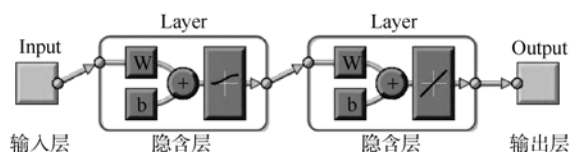


图 1 人工神经网络设计与结构
Fig. 1 Design and structure of neural network

2 结果与分析

2.1 小麦全基因组关联分析

通过对 1 279 个标记的全基因组关联分析, 发现大多数分子标记(533)对表型的遗传效应为 0, 即对小麦的产量性状没有遗传效应。因此, 在进行全

基因组选择之前, 需要对这些分子标记进行优化, 去除对产量性状没有遗传效应或者遗传效应较低的分子标记。通过检测分子标记对产量性状贡献程度, 去除大量对小麦产量性状遗传效应不显著的分子标记, 最终保留 86 个具有显著遗传效应的标记(图 2), 用于后续的基因组选择研究。

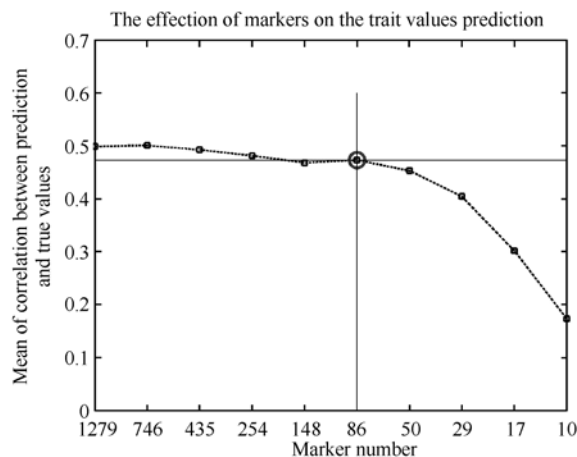


图 2 小麦全基因组分子标记对产量性状预测的影响
Fig. 2 Genetic effects of markers among the whole wheat genome to yield

2.2 RR-BLUP 对小麦产量性状的预测

通过岭回归将线性模型的参数确定, 然后对利用训练优化的 RR-BLUP 模型预测小麦的产量性状, 将预测的育种值与小麦的实际产量值相比(表 1)。预测的育种值与实际育种相关性最高达到了 0.8076, 而且相关系数的均值为 0.6422, 标准差为 0.0593, 这说明 RR-BLUP 模型的预测能力较强, 且预测能力稳定。

表 1 4 种模型预测的产量值与实际产量值间相关系数
Table 1 Correlation coefficients between predicted and true yield values among four models

方法 Method	均值 Mean	标准差 SD	最大值 Max.	最小值 Min.
RR-BLUP	0.6422 C	0.0593	0.8076	0.4083
BLR-P	0.6573 B	0.0590	0.8069	0.4443
BLR	0.6294 D	0.0621	0.7790	0.3779
ANN	0.6636 A	0.0654	0.8379	0.3731

均值后不同字母表示多重比较差异显著 ($P < 0.01$)。

Means followed by a different letter are significantly different at $P < 0.01$.

2.3 BLR 对小麦产量性状的预测

BLR 在系谱信息的基础上, 进行遗传育种值估计效果较好, 均值达到 0.6573 (BLR-P), 略好于 RR-BLUP 的预测效果, 最高值达到 0.8069, 与玉米中报道的极值(0.79)相近。当系谱信息缺失时, 预测

能力较 RR-BLUP 相比一般, 均值达到 0.6294(BLR)。这说明在进行作物遗传育种值估计时, 系谱信息会影响预测值的准确性。在这组小麦的基因型和产量数据中, 由于标记数据较少, 系谱信息对预测结果的影响作用较小。另外, 在这组小麦数据中, 由于分子标记数据较少, 而且标记间联系也较少, 在不考虑系谱信息情况下, RR-BLUP 可能更适合用于遗传育种值的预测。但是从预测值与实际值的相关系数分布来看, BLR 在高相关系数区域出现的次数明显偏多, 而 RR-BLUP 略少(图 3), 这比较适合用于无法进行模型预测效率的育种过程。

2.4 ANN 对小麦产量性状的预测

建立双层 BP 人工神经网络(图 2), 对每组数据训练 30 次, 形成 30 个训练效果良好, 具有预测能力的人工神经网络, 从中根据训练过程选取预测效果优异的 5 个, 构建人工神经网络预测群体, 能大大提高人工神经网络的预测能力(图 4)。吉布斯抽样 1 000 次结果显示, 预测育种值与实际育种之间相关系数达 0.6636, 较 RR-BLUP、BLR 和 BLR-P 模型分

别高 0.0214、0.0342 和 0.0063(表 1), 说明人工神经网络群体预测能力略优于 RR-BLUP 和 BLR, 与基于系谱信息的 BLR 预测能力相当。而且相关系数的标准差为 0.0654, 较 RR-BLUP、BLR 和 BLR-P 稍大一些, 说明人工神经网络能够预测变异幅度稍大(图 3), 容易寻找到较传统的统计遗传学方法预测效果更优的人工神经网络(遗传预测产量值与实际产量值间相关性高达 0.8379, 图 5)。

3 讨论

3.1 基因组选择育种的优势

为了寻找控制性状的主效位点, 通过杂交等方式构建合适的遗传群体, 定位控制性状的主效基因位点, 即 QTL 定位。此后又开发了关联分析方法, 在自然群体中, 定位一些控制动植物性状的位点。这 2 种遗传育种方法都只能发现一些遗传作用效应较大的位点, 特别是 QTL 定位, 一般都是定位遗传效应在 10%或者 5%以上^[12]。但是大多数动植物的表型性状并不是由极少数的主效基因控制, 而是由大量

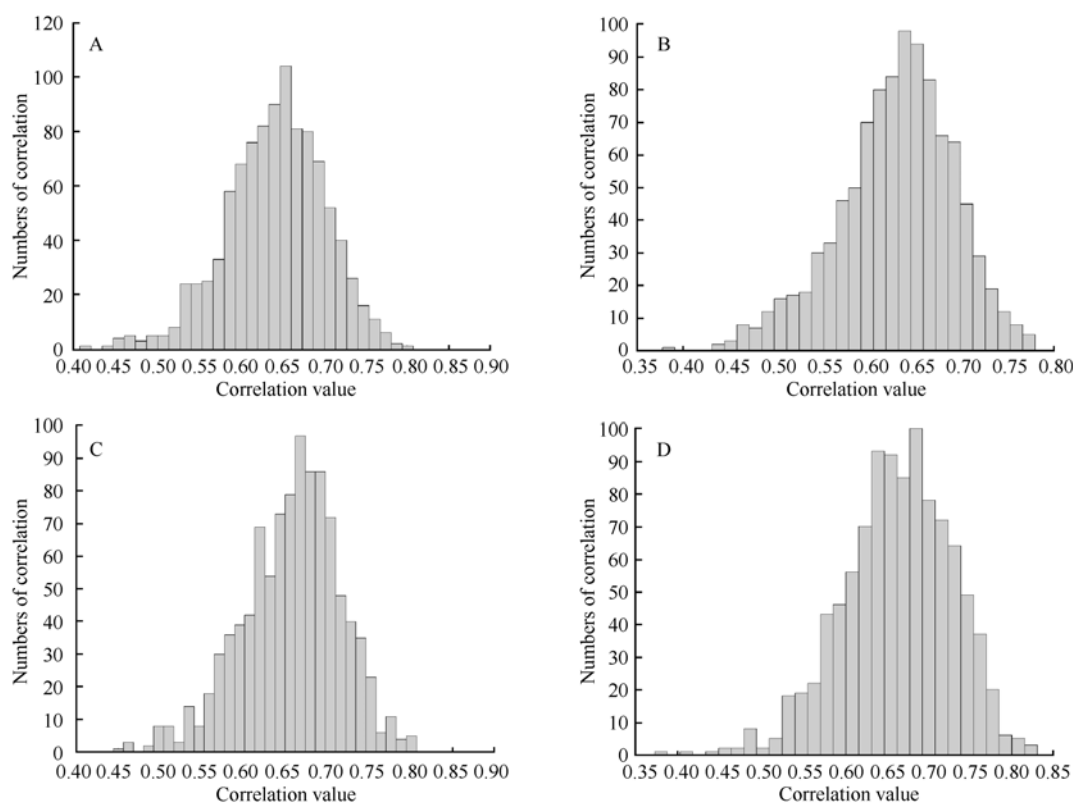


图 3 4 种基因组选择模型的预测产量值与实际产量性状的相关性分布

Fig. 3 Distribution of correlations between predictions from four models and true yield values

A: RR-BLUP 预测结果与实际值间的相关系数; B: 不含系谱信息的 BLR 预测结果与实际值间的相关系数; C: 基于系谱信息的 BLR 预测结果与实际值间的相关系数; D: ANN 预测结果与实际值间的相关系数.

A: Distribution of correlation between RR-BLUP prediction and true values; B: Distribution of correlation between BLR prediction and true values; C: Distribution of correlation between BLR prediction with pedigree information and true values; D: Distribution of correlation between ANN prediction and true values.

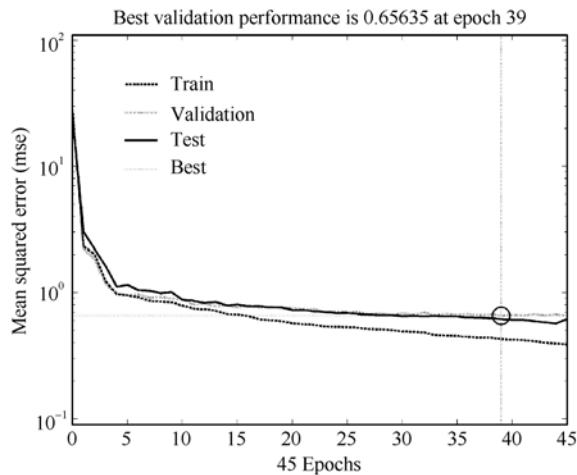


图 4 人工神经网络训练结果
Fig. 4 Training results of the neural network

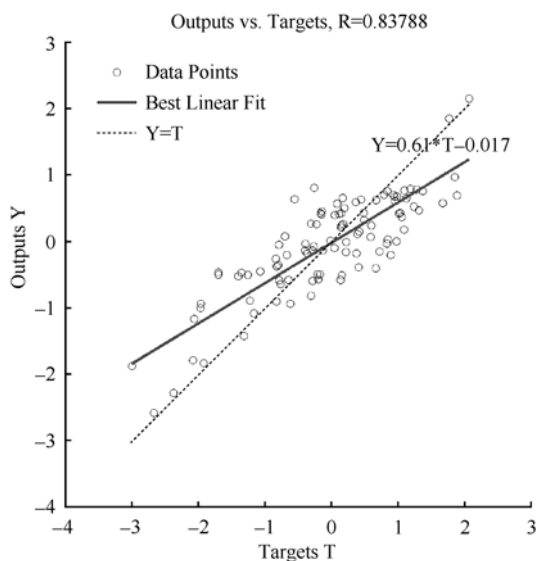


图 5 人工神经网络的预测结果
Fig. 5 Prediction results from artificial neural network

的微效基因控制。面对动植物基因组内的大量关系复杂的微效控制基因,传统的 QTL 定位或者关联分析检测都遇到了瓶颈^[32-33]。

基因组选择是指在基因组范围内,考虑全部基因或者遗传位点对表型性状的影响。它一般通过构建“训练群体”,利用其个体的基因型和表型值,优化模型。然后根据“候选群体”的基因型,利用优化的模型获得遗传估计育种值(genomic estimated breeding values, GEBV)^[32-33]。GEBV 虽然无法解释基因间的调控关系,但在育种中具有重要作用,是动植物育种中理想的选择标准。在作物遗传育种过程中,根据个体的基因型获得遗传估计育种值将会大大加快优异作物品种的选育。

3.2 RR-BLUP 和 BLR 模型的优缺点

在几种基因组选择育种模型中,RR-BLUP 并不

像传统的全基因组关联分析方法那样,假设所有的分子标记具有显著的遗传效应,或者不具有遗传效应,它是假设所有的分子标记都具有一定的、相同的遗传效应,然后通过回归过程将遗传效应极低或者没有遗传效应的分子标记系数紧缩到 0,从而达到全基因组选择的效果^[33]。因此在选择的同时,它还可以迅速评估每个分子标记的遗传效应。它主要适用于分子标记较少,或者大量微效分子标记的群体训练,如本研究中的小麦产量预测。在实际生产中,全基因组范围内,有些分子标记分布在 QTL 区域,具有较强的遗传效应,而另外一些分子标记却分布在没有遗传效应的区域,这使得每个分子标记的遗传效应并不一样,这就大大限制了 RR-BLUP 模型的使用。贝叶斯估计的模型将有助于解决全基因组分子标记遗传效应差异问题。针对分子标记的遗传效应进行特异的调整,排除大量无关的、或者低效应分子标记干扰,适合应用于分子标记密度较大的全基因组选择育种研究^[29,32]。在本研究中,86 个分子标记经过一次全基因关联分析的选择表明,都是具有一定遗传效应的分子标记,因此,针对这样的群体样本训练,RR-BLUP 模型的效果要略优于 BLR。如果考虑样本品系间的关系,BLR 模型的预测效果就会提升很多,预测效果会略优于简单的 RR-BLUP 模型。

3.3 ANN 模型的优势

在生物体内,其表型性状一般都是由大量基因组成的复杂调控网络所调控,这些基因间的调控关系只有极少数是线性或近似线性的调控关系,大多数都是非线性的复杂调控关系,这就意味着基于传统线性回归的模型,如 BLUP、RR-BLUP 和 BLR 等就很难真正地模拟出基因间的真实关系,进而限制这些模型在育种中的预测能力。而人工神经网络是由大量处理单元互联组成的非线性、自适应信息处理系统,非常适合应用于一些复杂系统的模拟,理论上能够无限接近生物系统的真实调控关系,从而提高模型的预测能力,提高作物的选育效率。

在作物育种过程中,虽然系谱信息对育种过程的指导作用显著,但是由于成本等原因,很多品种在育成过程中发生系谱记录不全、丢失等情况,造成育种品种的遗传作用估计不足等现象。人工神经网络是黑箱的处理模式,模型的假设比较少,对模型数据要求也少。如本研究涉及的小麦遗传育种数据,BLR 模型在有系谱信息的基础上,预测效果会有大幅度的提升,但是缺乏系谱信息时,预测效果

就会差很多。而 ANN 模型在系谱信息未知的情况下, 还能保证高精度的预测结果, 略优于利用系谱信息的 BLR。因此, 基于 ANN 的基因组选择非常适合用于系谱信息未知的作物遗传育种。

此外, 人工神经网络是自适应系统, 具有自主学习数据的能力, 学习能力跟很多因素相关, 如学习样本的群体, 网络的初始状态等。在同样训练样本的条件下, 由于网络的初值状态不同, 导致训练后的人工神经网络的预测能力也会出现差异, 这样就可以选择训练效果较好、预测能力优异的人工神经网络构建人工神经网络预测群体, 通过群体的预测结果会大大提高预测的精度。而传统的线性回归模型只是机械地优化回归参数, 同样的训练样本训练出的模型参数完全一样, 训练完的模型预测能力一样, 根本没有选择的余地, 在模型改进方面限制较大。

4 结论

建立了基于人工神经网络的作物全基因组选择模型, 对小麦产量性状的预测效果良好, 预测育种值与实际育种值间相关性很高。与传统线性回归模型 RR-BLUP、BLR 和 BLR-P 相比, 具有预测能力强、精度高等优点, 可以用于小麦以及其它作物的遗传育种, 具有重要的理论价值和应用前景。

致谢: 本文所用数据由国际玉米小麦改良中心(CIMMYT)收集和整理, 在此谨致谢意。

References

- [1] Henderson C. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 1975, 31: 423–447
- [2] Henderson C R. Applications of Linear Models in Animal Breeding. Guelph (ONT): University of Guelph, 1984
- [3] Cantet R J C, Smith C. Reduced animal model for marker assisted selection using best linear unbiased prediction. *Genet Selection Evol*, 1991, 23: 1–13
- [4] Panter D M, Allen F L. Using best linear unbiased predictions to enhance breeding for yield in soybean: I. Choosing parents. *Crop Sci*, 1995, 35: 397–405
- [5] Panter D M, Allen F L. Using best linear unbiased predictions to enhance breeding for yield in soybean: II. Selection of superior crosses from a limited number of yield trials. *Crop Sci*, 1995, 35: 405–410
- [6] Bernardo R. Best linear unbiased prediction of maize single-cross performance given erroneous inbred relationships. *Crop Sci*, 1996, 36: 862–866
- [7] Purba A R, Flori A, Baudouin L, Hamon S. Prediction of oil palm (*Elaeis guineensis* Jacq.) agronomic performances using the best linear unbiased predictor (BLUP). *Theor Appl Genet*, 2001, 102: 787–792
- [8] Bauer A M, Reetz T C, Léon J. Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. *Crop Sci*, 2006, 46: 2685–2691
- [9] Xie C, Carlson M, Murphy J. Predicting individual breeding values and making forward selections from open-pollinated progeny test trials for seed orchard establishment of interior lodgepole pine (*Pinus contorta* ssp. *latifolia*) in British Columbia. *New For*, 2007, 33: 125–138
- [10] Piepho H, Möhring J, Melchinger A, Büchse A. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 2008, 161: 209–228
- [11] Varshney R K, Graner A, Sorrells M E. Genomics-assisted breeding for crop improvement. *Trends Plant Sci*, 2005, 10: 621–630
- [12] Kearsey M J, Farquhar A G L. QTL analysis in plants; where are we now? *Heredity*, 1998, 80: 137–142
- [13] Wang J K(王建康), Wolfgang H P. Simulation approach and its applications in plant breeding. *Sci Agric Sin* (中国农业科学), 2007, 40(1): 1–12 (in Chinese with English abstract)
- [14] Wang J-K(王建康), Li H-H(李慧慧), Zhang X-C(张学才), Yin C-B(尹长斌), Li Y(黎裕), Ma Y-Z(马有志), Li X-H(李新海), Qiu L-J(邱丽娟), Wan J-M(万建民). Molecular design breeding in crops in China. *Acta Agron Sin* (作物学报), 2011, 37(2): 191–201 (in Chinese with English abstract)
- [15] Agrama H, Eizenga G, Yan W. Association mapping of yield and its components in rice cultivars. *Mol Breed*, 2007, 19: 341–356
- [16] Zhu C, Gore M, Buckler E S, Yu J. Status and prospects of association mapping in plants. *Plant Genome*, 2008, 1: 5–20
- [17] Zhao K, Aranzana M J, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M. An Arabidopsis example of association mapping in structured samples. *PLoS Genet*, 2007, 3: e4
- [18] Goddard M E, Hayes B J. Genomic selection. *J Anim Breed Genet*, 2007, 124: 323–330
- [19] De Roos A P W, Schrooten C, Mullaart E, Calus M P L, Veerkamp R F. Breeding value estimation for fat percentage using dense markers on *Bos taurus* autosome 14. *J Dairy Sci*, 2007, 90: 4821–4829
- [20] Long N, Gianola D, Rosa G J M, Weigel K A, Avendaño S. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J Anim Breed Genet*, 2007, 124: 377–389
- [21] Meuwissen T. Genomic selection: marker assisted selection on a genome wide scale. *J Anim Breed Genet*, 2007, 124: 321–322
- [22] Legarra A, Robert-Granié C, Manfredi E, Elsen J M. Performance of genomic selection in mice. *Genetics*, 2008, 180: 611–618
- [23] Luan T, Woolliams J A, Lien S, Kent M, Svendsen M, Meuwissen T H E. The accuracy of genomic selection in norwegian red cattle

- assessed by cross-validation. *Genetics*, 2009, 183: 1119–1126
- [24] Piyasatian N, Fernando R L, Dekkers J C. Genomic selection for marker-assisted improvement in line crosses. *Theor Appl Genet*, 2007, 115: 665–674
- [25] Li Y(黎裕), Wang J-K(王建康), Qiu L-J(邱丽娟), Ma Y-Z(马有志), Li X-H(李新海), Wan J-M(万建民). Crop molecular breeding in China: current status and perspectives. *Acta Agron Sin* (作物学报), 2010, 36(9): 1425–1430 (in Chinese with English abstract)
- [26] Wong C, Bernardo R. Genome wide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor Appl Genet*, 2008, 116: 815–824
- [27] de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes J M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 2009, 182: 375–385
- [28] Crossa J, Campos G D L, Pérez P, Gianola D, Burgueño J, Araus J L, Makumbi D, Singh R P, Dreisigacker S, Yan J, Arief V, Banziger M, Braun H J. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 2010, 186: 713–724
- [29] Pérez P, de los Campos G, Crossa J, Gianola D. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *Plant Genome*, 2010, 3: 106–116
- [30] He Z-H(何中虎), Xia X-C(夏先春), Chen X-M(陈新民), Zhuang Q-S(庄巧生). Molecular design breeding in crops in China. *Acta Agron Sin* (作物学报), 2011, 37(2): 202–215 (in Chinese with English abstract)
- [31] Kang H M, Sul J H, Service S K, Zaitlen N A, Kong S-Y, Freimer N B, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 2010, 42: 348–354
- [32] Jannink J L, Lorenz A J, Iwata H. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics*, 2010, 9: 166–177
- [33] Heffner E L, Sorrells M E, Jannink J-L. Genomic selection for crop improvement. *Crop Sci*, 2009, 49: 10–12