文章编号：1000-8608(2020)01-0094-06

基于全基因组选择的长牡蛎肥满度分布参数预测方法

董青原[1,2]，曹隽喆[1,2]，张国范[3]，李　莉[3]，刘　圣[3]，顾　宏*[1,2]

（1.大连理工大学 工业装备智能控制与优化教育部重点实验室，辽宁 大连　116024；
2.大连理工大学 控制科学与工程学院，辽宁 大连　116024；
3.中国科学院海洋研究所 实验海洋生物学重点实验室，山东 青岛　266071）

摘要： 全基因组选择是一种用于改良动植物育种群体中数量性状的方法，通过使用覆盖整个基因组的分子标记信息对复杂性状进行预测，从而帮助筛选出更适合培育的亲本. 基于长牡蛎的单核苷酸多态性(SNP)位点信息，提出了一种预测长牡蛎肥满度分布参数的全基因组选择的新方法. 首先，采用一种基于不同评价准则的二次特征选择方法，挑选与肥满度相关性较高的SNP位点；其次，利用所挑选的SNP位点信息构建具有正则化项的高斯通用加性模型对每个长牡蛎样本肥满度分布参数进行预测；最后，在长牡蛎数据上将所提方法和一些现有方法进行了验证比较. 实验结果表明，所提方法具有更好的拟合精度和更低的均方误差，并能对样本性状稳定性进行有效的评估.

关键词： 全基因组选择；单核苷酸多态性；二次特征选择；高斯通用加性模型；长牡蛎

中图分类号： Q-31；TP301.6　　**文献标识码：** A　　**doi：** 10.7511/dllgxb202001013

0　引　言

长牡蛎(*Crassostrea gigas*)是我国黄渤海区的主养贝类之一，含有丰富的营养物质. 然而，由于我国牡蛎养殖产业整体效益长期在低水平徘徊[1]，对牡蛎进行选择育种是当今养殖业亟待解决的问题. 长牡蛎选择育种最重要的经济指标就是条件指数(condition index，CI)，即肥满度[2]. 作为牡蛎的一种数量性状，肥满度的定义为软体干质量与干壳质量的比，即牡蛎的软体部在贝壳中的充盈程度.

利用覆盖整个基因组的高密度遗传标记来预测对应性状和育种值的方法称为全基因组选择(genomic selection，GS)[3]. 其中最常用的分子标记为单核苷酸多态性(single nucleotide polymorphism，SNP)位点. Meuwissen 等[3]在模拟数据集中提出了全基因组最佳线性无偏预测模型(genomic best linear unbiased prediction，GBLUP)，并较准确地预测出了育种值. Chen 等[4]利用图结构来约束稀疏线性回归模型，并在人类微生物组数据集中得到了较好的预测效果. Maenhout 等[5]利用带核函数的支持向量回归法(support vector regression，SVR)构建了SNP位点之间的非线性关系，并将其应用于谷粒的杂交育种中得到了较高的预测准确度. Wang 等[6]利用神经网络预测出了奶牛的脂肪产量、产奶量和蛋白质产量3个响应变量. 但是这些方法均仅关注性状的预测精度，并未考虑样本个体性状表达方差的差异性. 然而适合育种的个体，不仅在性状值的平均水平上达到要求，而且其表现的波动范围应该尽量小，以保证所选择的优质培育个体子代均稳定地遗传表现，这就需要考察每个个体的目标性状分布情况，然而已有的研究方法大多无法处理. 因此，为了对优质的长牡蛎进行育种选择，对每个长牡蛎个体肥满度的分布参数进行预测成为亟待解决的问题.

本文首先利用二次特征选择的方式逐步提取

收稿日期： 2019-09-08；　**修回日期：** 2019-11-25.
基金项目： 国家自然科学基金资助项目(61502074).
作者简介： 董青原(1993-)，女，硕士生，E-mail：dqingyuan@163.com；张国范(1954-)，男，研究员；李莉(1975-)，女，研究员；顾　宏*(1961-)，男，教授，博士生导师，E-mail：guhong@dlut.edu.cn.

与长牡蛎肥满度相关的 SNP 位点；接着，利用这些位点信息构建一种高斯通用加性模型(Gaussian generalized additive models, GGAM)[7]，根据 SNP 位点信息同时预测出每只长牡蛎的肥满度均值和方差，为长牡蛎选择性育种挑选肥满度高且表现稳定的个体进行培育提供更全面的参考信息.

1　数据来源及预处理

1.1　数据来源

本文选用我国黄渤海区的主要养殖牡蛎种类——长牡蛎作为研究对象，数据来源为中国科学院海洋研究所在山东青岛胶南海域采集的野生长牡蛎样本. 7 月龄时将其分成单体后，于扇贝笼中等密度养殖，每层 10 只牡蛎，以消除环境效应可能带来的影响. 18 月龄时对牡蛎进行取样，测量壳高、壳长、壳宽、湿重、软体重等生长数据. 并使用 AffymetrixSNP 芯片[8]得到了 288 个牡蛎样本全基因组 SNP 位点信息，每个样本包含 189 203 个初始 SNP 位点.

1.2　数据预处理

根据生物信息学中通用的数据清洗方法[9]，删除其空值占比大于 10%的 SNP 位点和肥满度为空值的牡蛎样本，得到了 275 个样本的 SNP 数据，每个样本保留 165 825 个 SNP 位点. 将样本基因分型后，将空值用该位点在所有样本个体的突变均值进行填补.

图 1 为长牡蛎样本肥满度 I_c 的频率直方图以及其估计分布曲线. 可以看出长牡蛎肥满度的分布近似正态分布，因此，只需要根据长牡蛎样本的基因特征预测其肥满度的均值与方差这两个参数.

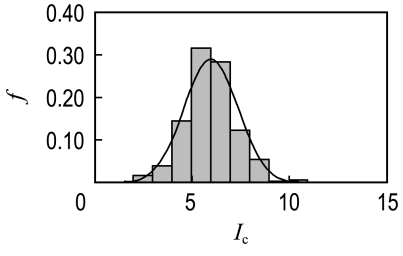

图 1　长牡蛎肥满度的频率直方图及拟合概率密度曲线

Fig. 1　Frequency histogram of CI values of *Crassostrea gigas* and corresponding fitted probability density curve

2　基于全基因组选择的预测方法

本文将长牡蛎全部 SNP 位点对应等位基因的基因型作为样本特征，肥满度条件指数的分布参数作为预测的目标变量，来构建预测模型.

2.1　二次特征选择方法

由于长牡蛎全基因组的 SNP 向量维度超过 16 万，而只有极少数 SNP 位点影响长牡蛎个体肥满度的表现差异，本文经过二次特征选择来逐步筛选位点. 先基于 Matrix eQTL 方法[10]，利用假设检验对 SNP 特征进行初步降维，然后根据对拟合模型复杂度的考量，利用脊回归(ridge regression)[11]与赤池信息准则(Akaike information criterion, AIC)[12]相结合的方法进行了二次特征选择，尽量用较少的参数来表现模型，使得在保证较高拟合精度的前提下，尽可能不遗漏相关度较高的基因位点.

设由 n 个长牡蛎样本组成的数据集为 $D=\{(\boldsymbol{x}_i, y_i)\mid i=1,2,\cdots,n\}$，对于其中第 i 个样本，其特征向量为 $\boldsymbol{x}_i=(1 \quad x_{i1} \quad x_{i2} \quad \cdots \quad x_{im})^{\mathrm{T}}\in \boldsymbol{R}^{m+1}$，这里 $x_{ij}\in\{0,1,2\}$ 为第 i 个样本的第 j 个 SNP 位点的基因型，$y_i\in\mathbf{R}$ 为该样本的肥满度，向量中的 1 是针对通用加性模型填充的常数项. 在第一次特征选择环节，面对超高维的 SNP 位点特征，本文采用了 Matrix eQTL 方法在处理 eQTL 定位问题时的策略，通过构建基于分块的一元线性模型进行初步的特征选择. 将基因型数据与长牡蛎的肥满度数据标准化，并对相关系数 r_j 进行 t 检验，使其处理大规模数据时快速地剔除了大量与肥满度性状关联度较低的冗余位点.

每个长牡蛎的特征向量 $\boldsymbol{x}_i$ 为解释向量，将 $\boldsymbol{x}_i$ 与该长牡蛎的肥满度 y_i 为目标值构建如下一元线性模型：

$$y_i=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i \tag{1}$$

式中：$\boldsymbol{\beta}$ 为系数向量，ε_i 为独立的高斯白噪声.

第 j 个特征与目标值的相关系数如式(2)所示，其中 $\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，为第 j 个特征的样本均值：

$$r_j=\frac{\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}=$$
$$\sum_{i=1}^{n}x_{ij}y_i;\quad j=1,2,\cdots,m \tag{2}$$

对 r_j 构建 t 统计量，进行假设检验：

$$t=\sqrt{n-2}\frac{r_j}{\sqrt{1-r_j^2}} \tag{3}$$

经过初步筛选，从 165 825 个初始位点中筛选出 1 754 个 P 值满足显著性水平要求即 $P<0.01$[13] 的 SNP 位点.

在第二次特征选择环节，本文利用 AIC 与对参数有约束的脊回归相结合的方式来权衡估计模型的复杂度和数据拟合的优良性. AIC 可表示为下式所示的形式：

$$C_{\mathrm{Ai}}=2k+n\ln\frac{R_{\mathrm{ss}}}{n} \tag{4}$$

式中：k 为参数个数，即线性模型中所用长牡蛎等位基因的个数；n 为长牡蛎样本个数；R_{ss} 为残差平方和，其表达式如下：

$$R_{\mathrm{ss}}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{5}$$

在脊回归所构成的模型集合中，通过前向选择方式选取 AIC 为最小值的模型，最终当 AIC 为 −646.467 2 时达到最低点. 此时，得到 211 个作为解释变量的 SNP 位点，即 $\boldsymbol{X}_1=(\boldsymbol{x}_1^{(1)}\quad\boldsymbol{x}_2^{(1)}\quad\cdots\quad\boldsymbol{x}_n^{(1)})^{\mathrm{T}}$，低于 275 个样本数量.

由此可见，经过二次特征提取原来的大 P 小 N 问题得到了解决，这使得在数据样本数量一定的情况下极大程度上缩小了特征空间维度，使得数据在空间中表现足够紧密，为保证预测模型的有效性与稳定性提供了良好的数据基础.

2.2 基于 GGAM 的肥满度分布参数预测模型

获得了有效的 SNP 位点之后，接下来就要利用这些位点作为特征构建有效的预测模型. 由于在长牡蛎的育种选择中，总是期望选择肉质更加饱满，并且其高肥满度的特征表现更稳定的个体作为优良品种加以培育. 为此，本文使用 GGAM[7] 对样本数量性状的均值与方差同时进行预测，并进一步与 Bagging 集成算法[14] 相结合以提高模型对方差值的预测效果.

对于方差的估计，考虑到要保证方差模型的复杂度尽可能降低，本文又进一步采用 AIC 来选择用来构建 GGAM 拟合方差的 SNP 位点. 利用前向选择的 AIC 选出方差对应的 SNP 位点特征变量，即 $\boldsymbol{X}_2=(\boldsymbol{x}_1^{(2)}\quad\boldsymbol{x}_2^{(2)}\quad\cdots\quad\boldsymbol{x}_n^{(2)})^{\mathrm{T}}$. 最终挑选出 176 个位点特征用作方差的评估时，AIC 为 −1 443.11 时达到最低点.

由于长牡蛎的 I_{c} 值即肥满度服从正态分布，对于任意样本 $\boldsymbol{x}_i$，其对应的 I_{c} 为 y_i，则每个长牡蛎样本肥满度的分布函数如下：

$$f(y_i\mid\mu_i,\sigma_i)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(\frac{-(y_i-\mu_i)^2}{2\sigma_i^2}\right) \tag{6}$$

将其均值 μ_i 与方差 σ_i 均参数化，其链接函数如下：

$$\mu_i(\boldsymbol{\beta}_1)=\boldsymbol{x}_i^{(1)\mathrm{T}}\boldsymbol{\beta}_1 \tag{7}$$

$$\sigma_i(\boldsymbol{\beta}_2)=\exp(\boldsymbol{x}_i^{(2)\mathrm{T}}\boldsymbol{\beta}_2) \tag{8}$$

其中 $\boldsymbol{\beta}_1$ 为预测均值的系数向量，$\boldsymbol{\beta}_2$ 为预测方差的系数向量，将 n 个样本的概率密度函数相乘可得到以下的似然函数：

$$L=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(\frac{-(y_i-\boldsymbol{x}_i^{(1)\mathrm{T}}\boldsymbol{\beta}_1)^2}{2(\exp(\boldsymbol{x}_i^{(2)\mathrm{T}}\boldsymbol{\beta}_2))^2}\right) \tag{9}$$

对似然函数取负对数转换为凸函数，D 为长牡蛎数据集：

$$\begin{aligned}l(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\mid D)&=-\log L(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\mid D)=\\&\sum_{i\in D}\left(\log\sqrt{2\pi}+\log\sigma_i(\boldsymbol{\beta}_2)+\frac{(\mu_i(\boldsymbol{\beta}_1)-y_i)^2}{2\sigma_i(\boldsymbol{\beta}_2)^2}\right)=\\&\sum_{i\in D}l_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\mid D)\end{aligned} \tag{10}$$

并对 $l_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\mid D)$ 关于 $\boldsymbol{\beta}_1$、$\boldsymbol{\beta}_2$ 分别求偏导并置零得到方程组：

$$\frac{\partial l_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\mid D)}{\partial\boldsymbol{\beta}_1}=\frac{\mu_i(\boldsymbol{\beta}_1)-y_i}{\sigma_i(\boldsymbol{\beta}_2)^2}\boldsymbol{x}_i^{(1)}=0 \tag{11}$$

$$\frac{\partial l_i(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\mid D)}{\partial\boldsymbol{\beta}_2}=\left[1-\frac{(\mu_i(\boldsymbol{\beta}_1)-y_i)^2}{\sigma_i(\boldsymbol{\beta}_2)^2}\right]\boldsymbol{x}_i^{(2)}=0 \tag{12}$$

由式(10)可知似然函数为凸函数，则其解为全局最大值，求得式(11)、(12)如下所示：

$$\hat{\boldsymbol{\beta}}_1=(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{y} \tag{13}$$

$$\hat{\boldsymbol{\beta}}_2=(\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{z} \tag{14}$$

式中：$\boldsymbol{X}_1=(\boldsymbol{x}_1^{(1)}\quad\boldsymbol{x}_2^{(1)}\quad\cdots\quad\boldsymbol{x}_n^{(1)})^{\mathrm{T}}$ 为 $\hat{\boldsymbol{\beta}}_1$ 的计划矩阵；$\boldsymbol{X}_2=(\boldsymbol{x}_1^{(2)}\quad\boldsymbol{x}_2^{(2)}\quad\cdots\quad\boldsymbol{x}_n^{(2)})^{\mathrm{T}}$ 为 $\hat{\boldsymbol{\beta}}_2$ 的计划矩阵；$\boldsymbol{y}$ 为长牡蛎肥满度的输出矩阵；$\boldsymbol{z}$ 为肥满度观测值与其期望估计值间偏差的绝对值的对数构成的向量，如下式所示：

$$\begin{aligned}\boldsymbol{z}=&(\ln(|y_1-\hat{\mu}_1|)\quad\ln(|y_2-\hat{\mu}_2|)\quad\cdots\\&\ln(|y_n-\hat{\mu}_n|))^{\mathrm{T}}\end{aligned} \tag{15}$$

此外，本文在利用 GGAM 预测方差时，针对长牡蛎样本较少的情况，采用 Bagging 提升方法扩充训练数据集. 本文利用重采样的方法训练了 100 个基学习器，经简单平均法得到最终结果. 为消除训练选择的误差，利用留一法(leave-one-out cross-validation)[15] 进行实验验证.

在 Bagging 的重采样过程中，由于每次采用有放回采样的方式必然会导致式(14)中 $\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2$ 并不是一个满秩矩阵，其逆矩阵无法求得. 本文通过

添加惩罚项的方式，在 $\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2$ 上添加合适的正则化项，如式(16)所示，使其可逆，以解决该问题：

$$\boldsymbol{R}=\begin{pmatrix}0 & 0 & \cdots & 0\\ 0 & a & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & a\end{pmatrix} \tag{16}$$

添加正则化项之后 $\hat{\boldsymbol{\beta}}_2$ 的求解公式如下式所示：

$$\hat{\boldsymbol{\beta}}_2=(\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2+\boldsymbol{R})^{-1}\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{z} \tag{17}$$

经测试，a 取值为 0.6 时效果最优. 长牡蛎肥满度分布参数预测流程如图 2 所示.

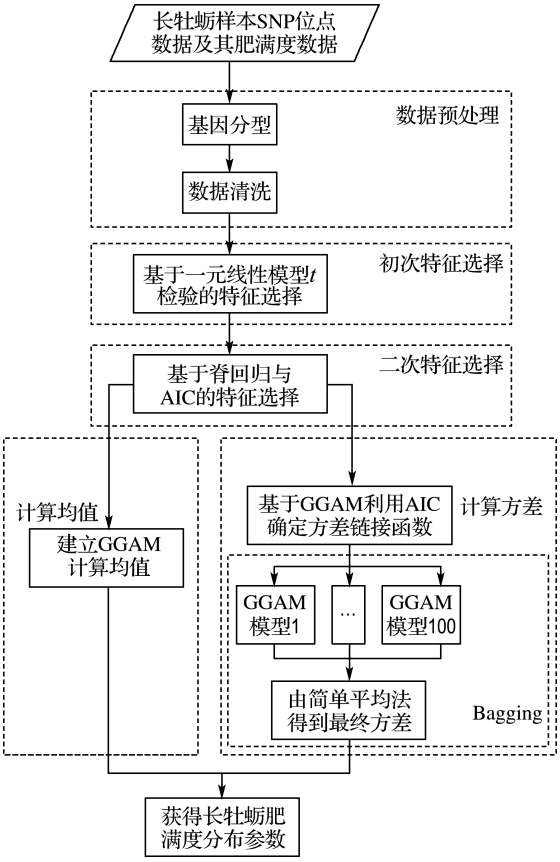

图 2　长牡蛎肥满度分布参数预测流程

Fig. 2　Flow chart of distribution parameters prediction in *Crassostrea gigas*'s condition index

3　实验结果与分析

3.1　二次降维的拟合精度对比

基于 1.2 节预处理后得到的长牡蛎样本，对本文方法进行了检验. 结果如图 3 所示，经过初次特征选择，从初始的 165 825 个 SNP 位点中筛选得到 1 754 个位点，构建脊回归模型得到拟合优度 R^2 为 0.736. 二次特征选择后，得到 211 个位点特征，再次构建脊回归模型得到拟合优度为 0.784. 经过二次特征选择后，不仅在拟合优度上有小幅度提升，并且特征空间维度 d 还被大大降低，使得特征向量维度小于样本个数，从而解决了大 P 小 N 问题.

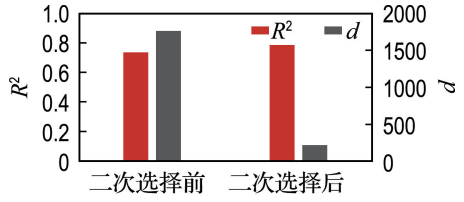

图 3　二次特征选择前后脊回归拟合优度及特征维度对比

Fig. 3　Comparison of R^2 in ridge regression and dimensions before and after two-stepwise feature selection

3.2　利用 GGAM 的分布预测

基于二次特征选择得到的 SNP 位点，进一步检验本文所建立的 GGAM 的预测效果. 为保证得到稳定的模型，本文依然使用留一法对每个样本进行逐一预测，最后得到每个样本的均值和方差，其中均值与真实值的拟合精度能达到 0.994 1，真实值落入置信区间的百分比能稳定到 70% 以上. 图 4 给出了预测曲线，其中横轴为样本按其真实肥满度由小到大的顺序依次排列，纵轴为对应的肥满度. 其中红色线为样本真值连线，蓝色线为预测样本均值($\hat{\mu}_i$)连线，灰色线为经过平滑之后的置信区间($\hat{\mu}_i\pm3\hat{\sigma}_i$). 如图 4 所示，红色曲线与蓝色曲线基本重合，证明均值的拟合结果较好. 而灰色曲线基本包裹了红色曲线，也就是说，大部分样本的真值落入求得的置信区间. 统计结果表明本文的方差估计对于 70% 以上的样本是准确的.

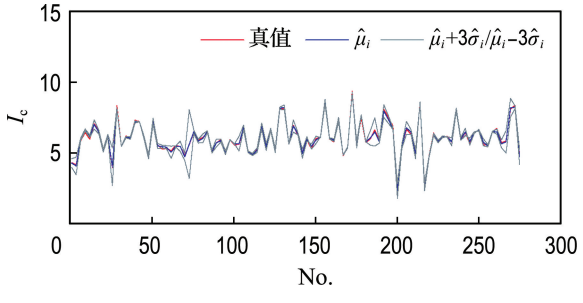

图 4　长牡蛎样本肥满度的真值与置信域($\hat{\mu}_i\pm3\hat{\sigma}_i$)预测曲线

Fig. 4　True value and confidence domain ($\hat{\mu}_i\pm3\hat{\sigma}_i$) prediction curve of every *Crassostrea gigas*'s CI value

4 与其他方法的实验结果对比

为了进一步检验本文方法的有效性，将所提出的 GGAM 方法对长牡蛎肥满度均值的拟合效果与 GBLUP[3]、snnR[6]、Glmgraph[4]、SBL[16] 等方法进行了对比，对比结果如表 1 所示. 从结果可以看到，本文方法取得了最好的拟合优度和最小的均方误差.

表 1 5 种方法的拟合优度及均方误差对比

Tab. 1 Comparison of R^2 and MSE for five methods

方法	R^2	均方误差
GGAM	0.994	0.011
SBL	0.978	0.043
GBLUP	0.940	0.114
snnR	0.866	0.225
Glmgraph	0.692	0.587

对上述方法的特点和实验结果综合分析可知，GBLUP 方法作为全基因组选择的常用方法，在采用本文的二次特征选择所得到的 SNP 位点时，能够得到较好的拟合优度与较低的均方误差. GBLUP 方法的优势在于预测样本遗传力时能够利用关系矩阵的变换进行降维计算，但在预测性状时则受其模型结构的制约无法降低特征维度. 而 Glmgraph 方法利用图结构来约束稀疏线性回归模型来解决大 P 小 N 问题，但其在长牡蛎数据集中的拟合效果一般. snnR 方法则通过搭建神经网络的方法，使得拟合精度有所提升，但随着隐层与节点数量的增多其耗时也越来越长. SBL 方法利用坐标下降算法，具有较高的拟合优度，但是均方误差比本文方法要大. 此外这几种方法均只能对目标性状值进行预测，无法对性状的波动进行评估.

本文构建的 GGAM 对牡蛎肥满度的预测精度达到了 0.994，相比其他 3 种方法有着绝对的优势，并且还能够对单个样本的方差进行评估. 在选择性育种时，样本个体数量性状表达的稳定性在育种选择中也是一个需要考量的重要指标，因此在选择数量性状高表达个体的同时也需要考察该样本个体数量性状的表达是否稳定，这也是本文为何要对每个样本的分布方差进行预测的原因. 本文方法在预测长牡蛎肥满度时，有 70% 以上的样本落入置信区间 $(\hat{\mu}_i \pm 3\hat{\sigma}_i)$ 中，这说明对于大部分长牡蛎样本条件指数的方差估计是正确的. 因此，本文利用基于二次特征选择的 GGAM 高精度地预测了每个牡蛎样本的均值与方差，为长牡蛎的育种选择提供了更全面的信息.

5 结 语

本文利用全基因组选择方法对长牡蛎肥满度的个体分布参数进行了预测. 首先，基于一元线性模型假设检验对 SNP 位点特征进行初次筛选，尽可能剔除冗余的基因位点；接着，利用脊回归和 AIC 相结合的方式进行二次特征选择，选取与长牡蛎肥满度最具关联性的 SNP 位点. 经过二次特征选择之后，原有的特征维度大大降低，避免了有限样本数量在高维空间中表现过于稀疏的问题. 最后，利用这些位点信息构建高斯通用加性模型将均值与方差参数化，实现样本与样本之间的异方差化，以预测样本目标性状的分布参数. 采用本文所提方法，能够获知影响长牡蛎肥满度均值和方差的基因特征，并在长牡蛎的育种筛选与养殖中提供更加全面的参考依据，进而筛选出肥满度高且方差小的个体进行繁育，提高种群中优质基因比例，使长牡蛎产品的肥满度差异达到最小.

参考文献：

[1] 宁 岳，郭 香，曾志南，等. 牡蛎育种研究进展[J]. 厦门大学学报(自然科学版)，2016，**55**(5)：624-636.
NING Yue, GUO Xiang, ZENG Zhinan, *et al*. Progress on oyster breeding [J]. **Journal of Xiamen University** (**Natural Science**), 2016, **55**(5): 624-636. (in Chinese)

[2] MERCADO-SILVA N. Condition index of the eastern oyster, *Crassostrea virginica* (Gmelin, 1791) in Sapelo Island Georgia-effects of site, position on bed and pea crab parasitism [J]. **Journal of Shellfish Research**, 2005, **24**(1): 121-126.

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

[4] CHEN Li, LIU Han, KOCHER J, *et al*. Glmgraph: An R package for variable selection and predictive modeling of structured genomic data [J]. **Bioinformatics**, 2015, **31**(24): 3991-3993.

[5] MAENHOUT S, DE BAETS B, HAESAERT G, *et al*. Support vector machine regression for the prediction of maize hybrid performance [J]. **Theoretical and Applied Genetics**, 2007, **115**(7): 1003-1013.

[6] WANG Y, MI X, ROSA G J M, *et al*. Technical note: An R package for fitting sparse neural networks with application in animal breeding [J]. **Journal of Animal Science**, 2018, **96**: 2016-2026.

[7] HASTIE T, TIBSHIRANI R. Generalized additive models [J]. **Journal of Statistical Science**, 1986, **1**(3): 297-310.

[8] QI Haigang, SONG Kai, LI Chunyan, *et al*. Construction and evaluation of a high-density SNP array for the Pacific oyster (*Crassostrea gigas*) [J]. **PLoS One**, 2017, **12**(3): e0174007.

[9] GUTIERREZ A P, BEAN T P, HOOPER C A, *et al*. A genome-wide association study for host resistance to ostreid herpesvirus in pacific oysters (*Crassostrea gigas*) [J]. **G3-Genes Genomes Genetics**, 2018, **8**(4): 1273-1280.

[10] SHABALIN A A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations [J]. **Bioinformatics**, 2012, **28**(10): 1353-1358.

[11] MCDONALD G C. Ridge regression [J]. **WIREs Computational Statistics**, 2010, **1**(1): 93-100.

[12] AKAIKE H. A new look at the statistical model identification [J]. **IEEE Transactions on Automatic Control**, 1974, **19**(6): 716-723.

[13] BIAU D J, JOLLES B M, PORCHER R, *et al*. *P* value and the theory of hypothesis testing: An explanation for new researchers [J]. **Clinical Orthopaedics and Related Research**, 2010, **468**(3): 885-892.

[14] BREIMAN L. Bagging predictors [J]. **Machine Learning**, 1996, **24**(2): 123-140.

[15] KEARNS M, RON D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation [J]. **Neural Computation**, 1999, **11**(6): 1427-1453.

[16] WANG M, XU S. A coordinate descent approach for sparse Bayesian learning in high dimensional QTL mapping and genome-wide association studies [J]. **Bioinformatics**, 2019, **35**(21): 4327-4335.

Method for predicting distribution parameters of condition index of *Crassostrea gigas* based on genomic selection

DONG Qingyuan[1,2], CAO Junzhe[1,2], ZHANG Guofan[3], LI Li[3], LIU Sheng[3], GU Hong[*1,2]

(1. Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Dalian 116024, China;
2. School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China;
3. Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China)

Abstract: Genomic selection (GS) is a method for improving quantitative traits in animal and plant breeding. By using genetic markers covering the whole genome of the species to predict complex traits, it can help screen out parents that are more suitable for breeding. Based on the single nucleotide polymorphism (SNP) locus information of *Crassostrea gigas*, a novel GS method for predicting distribution parameters of condition index of *Crassostrea gigas* is proposed. Firstly, a two-stepwise feature selection method based on different evaluation criteria is used to select SNP loci tightly bound to condition index. Secondly, the selected SNP loci are used to construct a generalized additive model under Gauss distribution with regularization terms for each sample to predict distribution parameters of condition index of *Crassostrea gigas*. Finally, the method is compared with other methods by employing the *Crassostrea gigas* data. The results show that the proposed method has better fitting accuracy and more accurate estimation variance. Meanwhile it can effectively evaluate the stability of sample traits.

Key words: genomic selection; single nucleotide polymorphism (SNP); two-stepwise feature selection; Gaussian generalized additive models; *Crassostrea gigas*