Joint Prediction of Multiple Quantitative Traits Using a Bayesian Multivariate Antedependence Model

J. Jiang, Q. Zhang, and J.F. Liu

China Agricultural University, Beijing, China

ABSTRACT: Whole-genome prediction (WGP) is widely used in livestock breeding. Among various statistical methods for WGP, two independent strategies, i.e., joint prediction of multiple traits and the antedependence model, show their respective advantages. To take advantage of both the strategies, we propose a Bayesian multivariate antedependence-based method for joint prediction of multiple quantitative traits by modeling a linear relationship of effect vector between each pair of adjacent markers. Using simulation and the 16th QTL-MAS workshop dataset, we demonstrate that our proposed WGP method is more accurate than corresponding traditional counterparts (Bayes A and multi-trait Bayes A). Our method can be readily extended to deal with categorical traits and missing phenotypes, offering a feasible way to jointly predict multiple complex traits in plant and livestock breeding.

Keywords: whole-genome prediction; multiple quantitative traits; Bayesian multivariate antedependence model

INTRODUCTION

Logically, a more perfect strategy for enhancing whole-genome prediction (WGP) is to take relations between multiple traits as well as between single nucleotide polymorphism (SNP) effects into consideration simultaneously for achieving a high prediction accuracy from the viewpoint of statistics. So far, it is still a gap between multiple-trait joint prediction approaches and SNP-correlated models, e.g., antedependence-based WGP models. Accordingly, aiming at developing an improved prediction methodology, we are in an attempt to construct a multiple-trait antedependence-based WGP model to bridge such a gap aforementioned under the framework of Bayesian algorithm. Specifically, we proposed a novel multivariate method with two different types of prediction models via setting the antedependence parameter as either a matrix or a scalar, respectively. Theoretically, our proposed models can relax the conventional assumption of independence of marker effects while simultaneously take advantage of correlation between traits by modeling a linear relationship of effect vector between each pair of adjacent markers. Using simulation and empirical data analysis, we compared our proposed method to the classical approaches including Bayes A by Habier et al. (Habier, Fernando et al. 2011) and multi-trait Bayes A by Jia and Jannink (Jia and Jannink 2012) to further validate the performance of our proposed method. Our study clearly demonstrated that the proposed multiple-trait antedependence-based WGP method was more accurate than corresponding traditional counterparts for genomic prediction, offering a feasible way to jointly predict complex traits in human genetic epidemiology as well as livestock breeding.

MATERIALS AND METHODS

Bayesian multivariate antedependence model. Considering different types of correlation between SNP markers, we developed two different forms of first-order multivariate antedependence models by modeling the antedependence parameter (Zimmerman and Nunez-Anton 2010) as a matrix as well as a scalar named herein as “matrix model” and “scalar model,” respectively. Assuming N individuals with genotypes at p SNP loci, the matrix model for joint prediction of m traits is expressed as:

\[ y_i | \mu, \alpha, \Sigma \sim N(\mu + \sum_{j=1}^{m} Z_{ij} \alpha_j, \Sigma) \quad i = 1, ..., n \]

\[ \alpha_j = \begin{cases} \delta_j & j = 1 \\ T_{j-1} \alpha_{j-1} + \delta_j & j = 2, ..., p \end{cases} \]

\[ T_{j-1} | \mathbf{M}, \Sigma_{Vj}, \Sigma_{Vj} \sim \text{MatrixNormal}(\mathbf{M}, \Sigma_{Vj}, \Sigma_{Vj}) \quad j = 2, ..., p \]

\[ \delta_j \sim N(0, \Sigma_{Vj}) \quad j = 1, ..., p \]

\[ \Sigma_{Vj} \sim W^{-1}(\text{scale} = 1, d.f. = m) \]

\[ \Sigma_{Vj} \sim W^{-1}(\text{scale} = 1, d.f. = m) \]

where \( y_i \) is an m-element phenotypic vector for individual \( i \) (\( i = 1, ..., n \)); \( \mu \) is the vector of overall population mean of m traits, \( \alpha_j \) is an m-element vector for the effects of the \( j \)-th molecular marker on all m traits (\( j = 1, ..., p \)); \( Z_{ij} \) is the SNP genotype code for individual \( i \) at marker \( j \), and \( \Sigma \) is an \( m \times m \) covariance matrix of residual effects. In this study, all markers are considered to be biallelic, and marker genotypes were coded as 0, 1 or 2 corresponding to the number of copies of an allele at each locus.

Specially, as shown in the model (1), each marker’s effect is assumed to have a linear relationship with that of the preceding adjacent marker based on the relative physical location along the chromosome, i.e., \( \alpha_j = T_{j-1} \alpha_{j-1} + \delta_j \). \( T_{j-1} \) is a marker specific \( m \times m \) matrix called antedependence parameter (Zimmerman and Nunez-Anton 2010), following a matrix normal distribution with mean matrix \( \mathbf{M} \) (\( m \times m \)), among-row covariance \( \Sigma_{Vj} \), and among-column covariance \( \Sigma_{Vj} \). It is worth noting that \( \Sigma_{Vj} \) is the among-row covariance of the prior distribution of \( T_{j-1} \) as...
well as the covariance of the prior distribution of $\delta_j$, which can facilitate the implementation of Gibbs sampling (Minka 2001). Further, we specified for $V_j$ an inverse Wishart prior distribution with an identity scale matrix and $m$ degrees of freedom. Accordingly, we assigned an inverse Wishart prior distribution with scale matrix $\Psi_{\Sigma}$ and $v_{\Sigma}$ degrees of freedom for matrix $\Sigma$ and an inverse Wishart prior distribution with scale matrix $\Psi_V$ and $v_V$ degrees of freedom for $V_j$ ($1, \ldots, p$). Both $\Psi_{\Sigma}$ and $\Psi_V$ were further assumed to follow a Wishart distribution with an identity scale matrix and $m$ degrees of freedom. We specified the hyper-parameters in model (1) as $\{v_{\Sigma} = 4.2, v_V = 4.2, M = 0_{n,m}\}$. The matrix model was solved through Gibbs sampling similar to the matrix model. The antedependence parameter $\alpha_j$ was solved for each trait. Normal error deviates were added to achieve heritabilities of 0.5 for trait 1 and 0.1 for trait 2. The covariance of errors between traits was set to zero. We considered the above simulation scenario as the default validation data set. In addition, we also simulated the scenario with 300 QTLs. Each of the simulation scenarios was repeated 30 times. For both the simulation scenarios, generation 5,001 was used as the training population and generation 5,002 as the validation population. Both the training and the validation population had a sample size of 500. Each individual in the training population had phenotypes for both simulated traits. For each dataset, we used the $5^\text{th}$ and $10^\text{th}$ filtered SNP to predict genomic breeding values, respectively, to explore the effect of decreasing LD levels of adjacent markers on the prediction accuracy of the methods. We used a paired t-test to test if there was a significant difference of prediction accuracy between our proposed models and the conventional models across the 30 replicates.

**Analysis of the 16\textsuperscript{th} QTL-MAS workshop dataset.** In addition to simulated datasets, we also analyzed the 16\textsuperscript{th} QTL-MAS workshop dataset (http://qtl-mas-2012.kassiopeagroup.com/en/index.php). The simulated dataset contained 10,000 SNP genotypes for each individual. Three milk production quantitative traits were simulated and genetically correlated. Three thousand females from generations 1, 2, and 3 had observations for all the three traits, which were used as the training population, and 1,020 individuals from generation 4 were used as the validation population. The prediction accuracies obtained by the different multi-trait methods were compared with officially reported results. The highest prediction accuracies officially reported were 0.794, 0.853, and 0.828 for traits 1, 2, and 3, respectively, which were used as the results of best single-trait methods.

**RESULTS**

**Results from the simulation.** As showed in Figure 1, the three multi-trait models, including our proposed ones, outperformed the single-trait Bayes A method for both the traits. We hereafter focused on comparison of prediction accuracy between multi-trait models. On the whole, our proposed methods (i.e., the scalar model and the matrix model) had higher prediction accuracies for both the simulated traits than the multi-trait Bayes A method. Specially, under the default scenario, the scalar model obtained a significantly higher prediction accuracy than the multi-trait Bayes A model for both the high-heritability trait 1 ($P<0.001$) and the low-heritability trait 2 ($P<0.05$) in each of the three LD levels, with the prediction accuracy increasing by 0.006-0.023. Under the default scenario, the matrix model also had higher prediction accuracies than the multi-trait Bayes A model for both trait 1 ($P<0.01$) and the low-heritability trait 2 ($P<0.05$) in each of the three LD levels and for trait 2 ($P<0.05$) in LD levels of $r^2 = 0.27$ and $r^2 = 0.33$, with the prediction accuracy increasing by 0.014 at most.

Under the 300-QTL scenario, however, there were no significant differences of prediction accuracy between each of our proposed models and the multi-trait Bayes A model, except those between the scalar model and the multi-trait Bayes A model in the lowest LD level ($r^2 = 0.22$). The prediction accuracies of traits 1 and 2 increased...
by 0.002 (P<0.01) and 0.005 (P<0.05) using the scalar model compared to the multi-trait Bayes A model, respectively, under the 300-QTL scenario with the lowest LD level.

In every LD level, each of the methods used in the study obtained higher prediction accuracy for both the high-heritability trait 1 and the low-heritability trait 2 under the 30 QTLs scenario than the 300-QTL scenario (Figure 1). Additionally, the prediction advantage of the multi-trait antedependence models over the multi-trait Bayes A model tended to be smaller as the number of underlying QTLs increased from 30 to 300.

**Results from the empirical data analysis.** As showed in Figure 2, multi-trait methods, including the multi-trait Bayes A model, the scalar model and the matrix model, showed considerable advantage over single-trait methods. The prediction accuracies of traits 1 and 3 increased by 0.040 and 0.032, respectively, using the conventional multi-trait Bayes A model compared to best single-trait methods, while the prediction of trait 2 was slightly lower. Further, using our proposed methods, i.e., the matrix model and the scalar model, the prediction accuracies were even higher than the multi-trait Bayes A model (Figure 2). Specifically, the prediction accuracies of traits 1, 2 and 3 increased by 0.028, 0.027 and 0.013, respectively, using the scalar model compared to the conventional multi-trait Bayes A model. The prediction accuracies of traits 1 and 3 increased by 0.005 and 0.011, respectively, using the matrix model compared to the multi-trait Bayes A model, while the prediction accuracy of trait 2 was near equal between the two models. In general, the scalar model obtained the best results with regard to the 16th QTL-MAS workshop dataset. The prediction accuracies for the three traits increased by as high as 8.6%, 2.8% and 5.5%, respectively, using the scalar model compared to best single-trait methods, which showed considerable superiority of the scalar model. As the three traits were generated to mimic milk production traits in this empirical dataset, the superiority of the scalar model may also be available in whole-genome prediction used for actual cattle breeding.

**DISCUSSION**

In general, as demonstrated by the analyses of both the simulated data and the empirical data, our proposed methods showed superiority over the conventional multi-trait Bayes A model which further exhibited advantage over single-trait method. This is consistent with our anticipation that the Bayesian multivariate antedependence models can benefit from both relaxing of the assumption of independence of marker effects and joint prediction of multiple traits. The multi-trait antedependence-based WGP method offers a feasible way in joint prediction of multiple complex traits in human genetic epidemiology as well as livestock breeding.

**LITERATURE CITED**

Habier, D., R. L. Fernando, et al. (2011). BMC Bioinformatics 12: 186.
Jia, Y. and J. L. Jannink (2012). Genetics 192(4): 1513-1522.
Minka, T. P. (2001). "Bayesian Linear Regression." Technical report, MIT Media Lab.
Zhang, Z., X. Ding, et al. (2012). 4th International Conference on Computer Modeling and Simulation (ICCMS 2012). Hong Kong, China, IACSIT Press. 22: 87-93.
Zimmerman, D. L. and V. A. Nunez-Anton (2010). Antedependence Models for Longitudinal Data. London/New York, Chapman & Hall.