Extend mixed models to multi-layer neural networks for genomic prediction including intermediate omics data

T. Zhao[1,2], J. Zeng[3] and H. Cheng[1*]

[1] Department of Animal Science, University of California Davis, 95616, Davis, US; [2] Integrative Genetics and Genomics Graduate Group, University of California Davis, 95616, Davis, US; [3] Institute for Molecular Bioscience, The University of Queensland, 4072, Brisbane, Australia; [*]qtlcheng@ucdavis.edu

Abstract

With the growing amount and diversity of intermediate omics data complementary to genomics (e.g., gene expression), there is a need to develop methods to incorporate intermediate omics data into conventional genomic evaluation. We developed a new method named NN-LMM to model the multiple layers of regulation from genotypes to intermediate omics features, then to phenotypes, by extending conventional linear mixed models ("LMM") to multi-layer neural networks ("NN"). NN-LMM incorporates intermediate omics features by adding middle layers between genotypes and phenotypes. Linear mixed models (e.g., GBLUP, Bayesian Alphabet) can be used to sample marker effects or genetic values on intermediate omics features, and activation functions in neural networks can capture the nonlinear relationships between intermediate omics features and phenotypes. NN-LMM had significantly better prediction performance than the recently proposed single-step approach. Moreover, NN-LMM can handle various patterns of missing omics measures. NN-LMM has been implemented in an open-source package called "JWAS".

Introduction

The advances in high-throughput sequencing technology provide growing amount and diversity of multi-omics data complementary to genomics. The effects of genotypes on phenotypes can be mediated by multiple layers of omics features through mechanisms such as regulatory cascades from epigenome, to transcriptome, and to proteome (Ritchie *et al.*, 2015; Wu *et al.*, 2018). This multi-layer regulation works as a unified system to connect genome variations to the trait, and the relationships between different layers can be complex with interactions and nonlinear relationships (Kitano, 2002). A system of two linear models has been developed recently for genomic evaluation (Christensen *et al*., 2021; Weishaar *et al*., 2020), where one linear model describes how genotypes affect gene expression levels, and another describes how gene expression levels affect phenotypes. This system of two linear models was further extended for incomplete omics data based on single-step approach (Christensen *et al.*, 2021).

We developed a new method named NN-LMM (Zhao *et al.*, 2021a; Zhao *et al.*, 2021b) to model the multiple layers of regulation from genotypes to intermediate omics features, then to phenotypes, by extending conventional linear mixed models ("LMM") to multi-layer artificial neural networks ("NN"). NN-LMM incorporates intermediate omics features by adding middle layers between genotypes and phenotypes. Linear mixed models can be used to sample marker effects or genetic values on intermediate omics features, and nonlinear activation functions in neural networks are used to approximate the nonlinear relationships between intermediate omics features and phenotypes. Compared to other methods, NN-LMM allows various patterns of missing omics data, for example, individuals can have different missing omics features, and the assumption of nonlinearity between intermediate omics features and phenotypes may be more biologically realistic.

Materials & Methods

The framework of NN-LMM incorporating intermediate omics data such as gene expression levels is shown in Figure 1(a). Genotypes affect the gene expression levels, then gene expression levels regulate the phenotypes. Linear mixed models can be applied to sample marker effects or genetic values on gene expression levels, and the non-linear activation function in neural networks will be used to capture the complex nonlinear relationships between gene expression levels and phenotypes. For an individual, the gene expression levels of the first two genes are 0.9 and 0.1, respectively, and the gene expression of the last gene is missing to be sampled. Individuals can have different missing gene expression levels. A detailed framework of NN-LMM is shown in Figure 1(b), and Markov chain Monte Carlo (MCMC) approaches are used to infer unknowns.

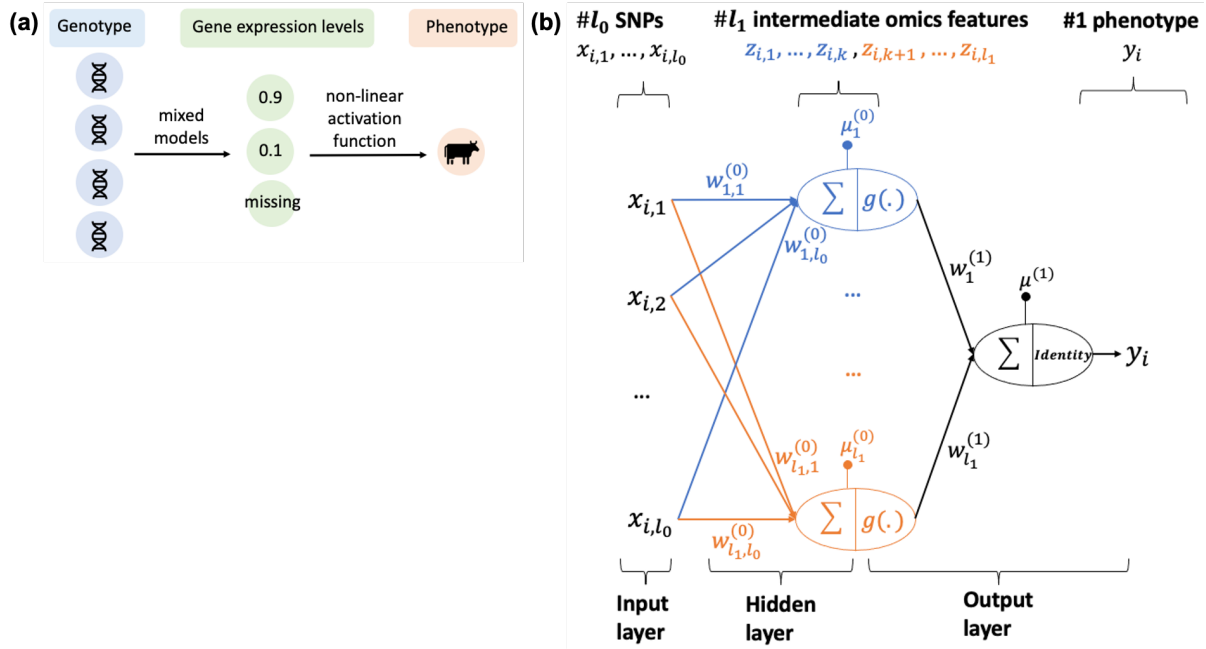

Figure 1. The framework of NN-LMM incorporating intermediate omics data.

From middle layer (intermediate omics features) to output layer (phenotypes): non-linear activation function. Given all intermediate omics features, the phenotype of individual *i* is modeled as:

$$y_i = \mu^{(1)} + \sum_{j=1}^{l_1} w_j^{(1)} g\left(z_{i,j}\right) + e_i \quad (1)$$

where y_i is the phenotype, $\mu^{(1)}$ is the overall mean with flat prior, $z_{i,j}$ is the *j*-th intermediate omics feature, $g(.)$ is the activation function in neural networks, $w_j^{(1)}$ is the effect of $g(z_{i,j})$ on y_i with a normal prior, and e_i is the random residual with normal prior.

From input layer (genotypes) to middle layer (intermediate omics features): mixed models. Given all intermediate omics features, for individual *i*, the relationship between the *j*-th intermediate omics feature and genotypes can be written as a single-trait mixed model (e.g., Bayesian Alphabet) as:

$$z_{i,j} = \mu_j^{(0)} + \sum_{m=1}^{l_0} x_{i,m} w_{j,m}^{(0)} + \epsilon_{i,j} \quad (2)$$

where $z_{i,j}$ is the *j*-th intermediate omics feature, $\mu_j^{(0)}$ is its overall mean with flat prior, $x_{i,m}$ is the genotype covariate at locus *m* (coded as 0, 1, 2), $w_{j,m}^{(0)}$ is the marker effects of locus *m* on *j*-th intermediate omics feature, which can be sampled by various linear mixed models, ε_{ij} is the random residual with normal prior.

Sample missing omics data by Hamiltonian Monte Carlo. Each missing omics feature of individual *i* will be treated as an unobserved intermediate trait to be sampled by Hamiltonian Monte Carlo (HMC) from its full conditional distributions. The introduction to the concepts underlying HMC can be found in Betancourt (2018).

Data Analysis. To compare the prediction performance of NN-LMM to the single-step approach in Christensen *et al.* (2021), a linear activation function was used in NN-LMM. GBLUP was used to sample genetic values on intermediate omics features (i.e., NN-GBLUP). Simulated data from Christensen *et al.* (2021) were used. Note that in Christensen *et al.* (2021), a polygenic effect whose covariance matrix is defined by the pedigree or/and genotypes is also included in Equation (1). This part is ignored here for simplicity, and was subtracted from the simulated phenotypes. Two patterns of missing omics data were considered: missing omics pattern (i): all omics data are completely missing for some individuals; missing omics pattern (ii): for each omics feature, some random individuals have no omics data. The single-step approach only works in the scenario (i), while NN-LMM allows both scenarios. Different proportions of missing omics data in the training dataset were considered, where 0% denotes the scenario where all omics features are measured on all individuals. 20 replicates were used for each scenario. We randomly sampled 5% individuals from the simulated data in Christensen *et al.* (2021) to have a subset of 1,055 individuals. The genotypic data consisted of 15,000 SNP markers observed for all individuals, and the intermediate omics data consisted of 1,200 omics features. The heritability of each omics feature was 0.61, and the heritability of the phenotypic trait was 0.337. More details about the simulation process are in Christensen *et al.* (2021).
We also simulated nonlinear relationships between intermediate omics features and phenotypes, where the logistic non-linear transformation was applied to the omics data as in Equation (1). In this case, the same heritability and variance components were applied.

Results
When all omics features were measured on all individuals, NN-GBLUP had similar prediction accuracies as the system of two mixed model equations in Christensen *et al.* (2021) (correlation r=0.999). When some omics data were missing, NN-LMM had equivalent or better prediction performance, as shown in Figure 2. Overall, the prediction accuracy decreased when the proportion of missing omics data increased. For missing omics pattern (i), when a small proportion of individuals had no omics data, NN-GBLUP (red solid line) had similar prediction performance as the single-step approach in Christensen *et al.* (2021) (blue solid line). However, when a large proportion of individuals had no omics data (e.g., >80%), NN-GBLUP had significantly higher prediction accuracies (pairwise t-test P-value < 0.005). When >90% individuals had no omics data, the single-step approach performed even worse than the baseline (black dashed line), which was a conventional GBLUP model where no omics data were used. For missing omics pattern (ii), when some random individuals had no omics measures for each omics feature, the prediction accuracy of NN-GBLUP (red dashed line) decreased with larger proportion of missing omics data, and eventually close to the baseline, whereas the single-step approach did not work for this scenario.
When the underlying relationships between omics features and phenotype was nonlinear, results verified that using the nonlinear sigmoid activation function in NN-LMM was significantly better than using the linear activation function.

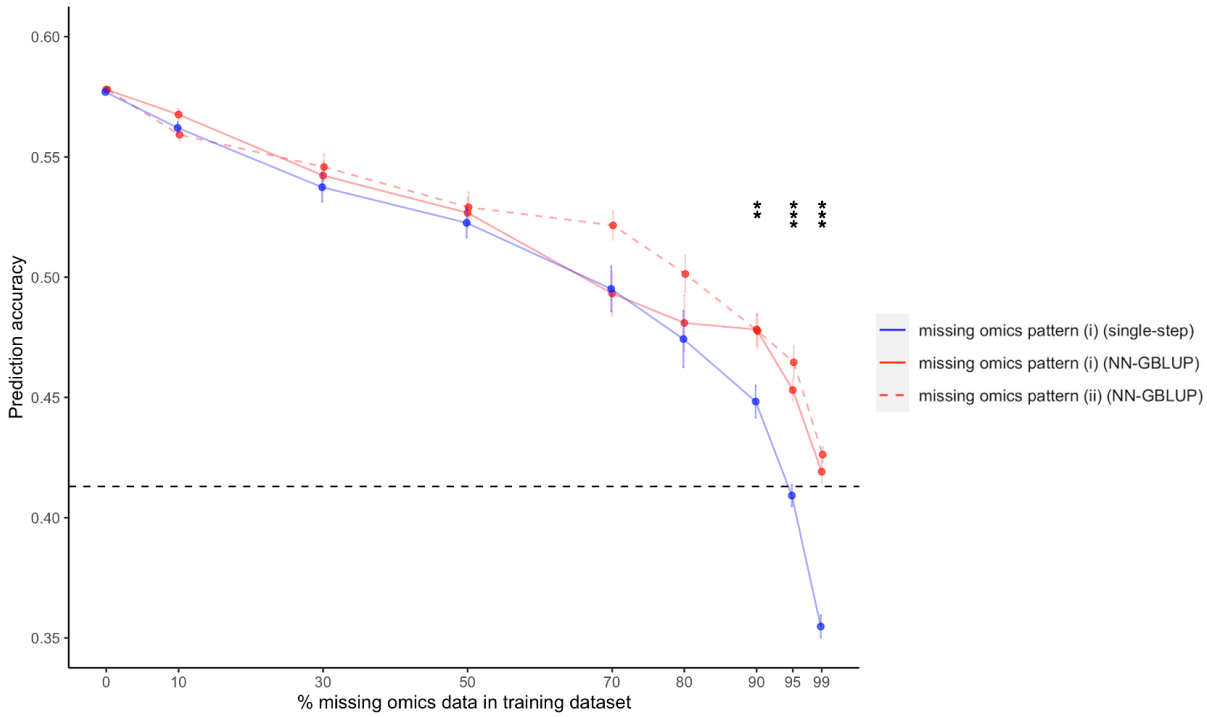

Figure 2. Prediction accuracies of NN-GBLUP with the linear activation function and the single-step approach in Christensen et al. (2021). The missing omics pattern (i): all omics features were not measured on some individuals, and missing omics pattern (ii): for each omics feature, some random individuals had no omics measures. The horizontal black dashed linear is the conventional GBLUP model used as the baseline. Each dot represents the mean of prediction accuracies from 20 replications with its standard error. The asterisk symbol indicated a significantly better performance of NN-GBLUP.

Discussion

To sample marker effects on omics features, a naive multi-threaded parallelism has been implemented to employ multiple single-trait mixed models in parallel at each MCMC iteration. Parallel computing strategies such as those in Zhao *et al.* (2020) will be further studied.

References

Betancourt M. (2018) arXiv preprint. https://arxiv.org/abs/1701.02434
Christensen O.F., Börner V., Varona L., and Legarra A. (2021) Genetics 219(2):iyab130. https://doi.org/10.1093/genetics/iyab130
Kitano H. (2002) Nature 420(6912):206-210. https://doi.org/10.1038/nature01254
Ritchie M.D., Holzinger E.R., Li R., Pendergrass S.A., and Kim D. (2015) Nat. Rev. Genet. 16(2):85–97. https://doi.org/10.1038/nrg3868
Weishaar R., Wellmann R., Camarinha-Silva A., Rodehutscord M., and Bennewitz J. (2020) J. Anim. Breed. Genet. 137(1):14–22. https://doi.org/10.1111/jbg.12447
Wu Y., Zeng J., Zhang F., Zhu Z., *et al*. (2018) Nat. Commun. 9(1):1-14. https://doi.org/10.1038/s41467-018-03371-0
Zhao T., Fernando R., Garrick D., and Cheng H. (2020) Genet. Sel. Evol. 52(1):1-11. https://doi.org/10.1186/s12711-020-00533-x
Zhao T., Fernando R., and Cheng H. (2021a) G3. 11(10). https://doi.org/10.1093/g3journal/jkab228
Zhao T., Zeng J., and Cheng H. (2021b) BioRxiv preprint https://doi.org/10.1101/2021.12.10.472186