Inferring Demographic History from Allele Frequency Spectra with Supervised Machine Learning

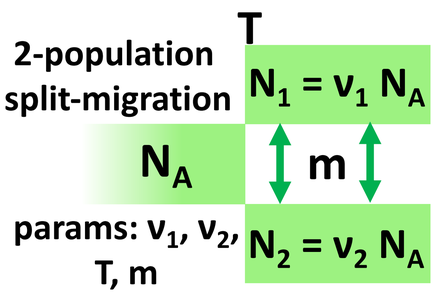
Linh N. Tran[1,2], Connie K. Sun[2], Ryan N. Gutenkunst[2]

[1] Genetics Graduate Interdisciplinary Program, [2] Department of Molecular and Cellular Biology, University of Arizona; lnt@email.arizona.edu; http://gutengroup.mcb.arizona.edu

Summary

Previously, our group had developed the software *dadi*[1] for inferring demographic history from DNA sequence data represented as allele frequency spectra (AFS) but computational expense challenges remain and significantly limit scalability. The major pipeline bottleneck lies in *dadi*'s optimization process, so we aimed to improve it with supervised machine learning (ML). We used *dadi* to simulate 10,000 AFS under a 2-population model (split-migration), then used them to train the scikit-learn[2] Random Forest and Multi-layer Perceptron regressors (RFR & MLPR) for parameter estimation. We tested our trained ML models with simulated AFS and found that their predictive accuracy varied for different ML models, demographic parameters, and different levels of variance in the AFS data. We also compared the computational efficiency between the original *dadi* method and a *dadi*+ML hybrid approach and found that our trained ML models can give good pre-optimization parameter estimations that can help reduce the optimization runtime by half without sacrificing accuracy.

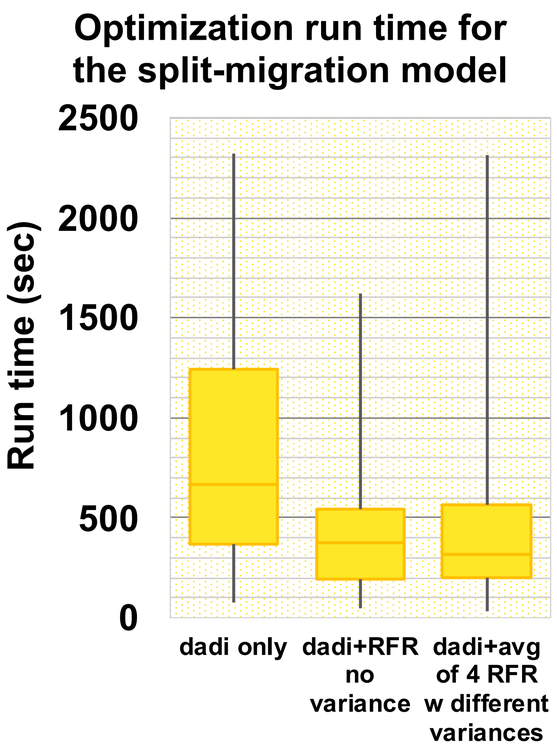
Demographic model & Variance

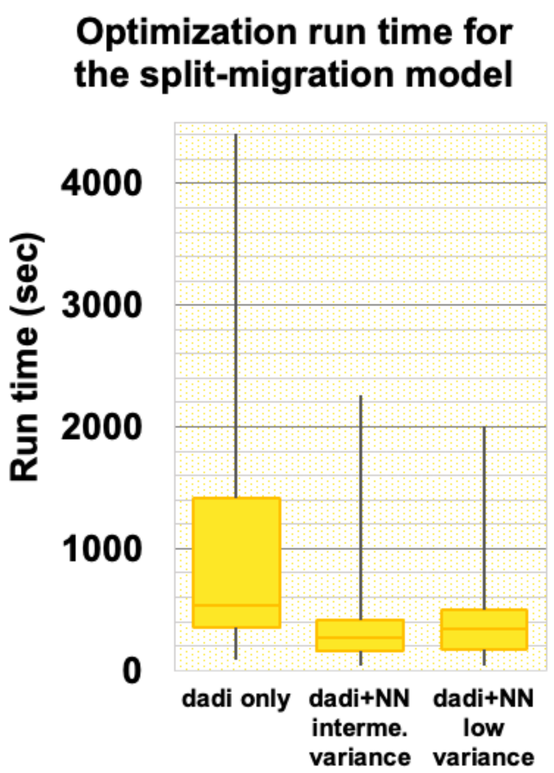


The split-migration model has 4 model parameters: ν_1 and ν_2 for population sizes, T for divergence time, and m for migration rate between populations. For ML training data, we first simulated 2,5000 expected AFS under this model from a realistic range of parameter values (no variance). We then Poisson-sampled from those spectra to generate 3 more sets of 2,5000 AFS, each set with different levels of variance. For testing data, we randomly selected 100 sets of model parameters and simulated 4 sets of 100 AFS each with different levels of variance in the same manner. We included different variances in order to observe how variance affects the RFR and MLPR learning and predictive performance.

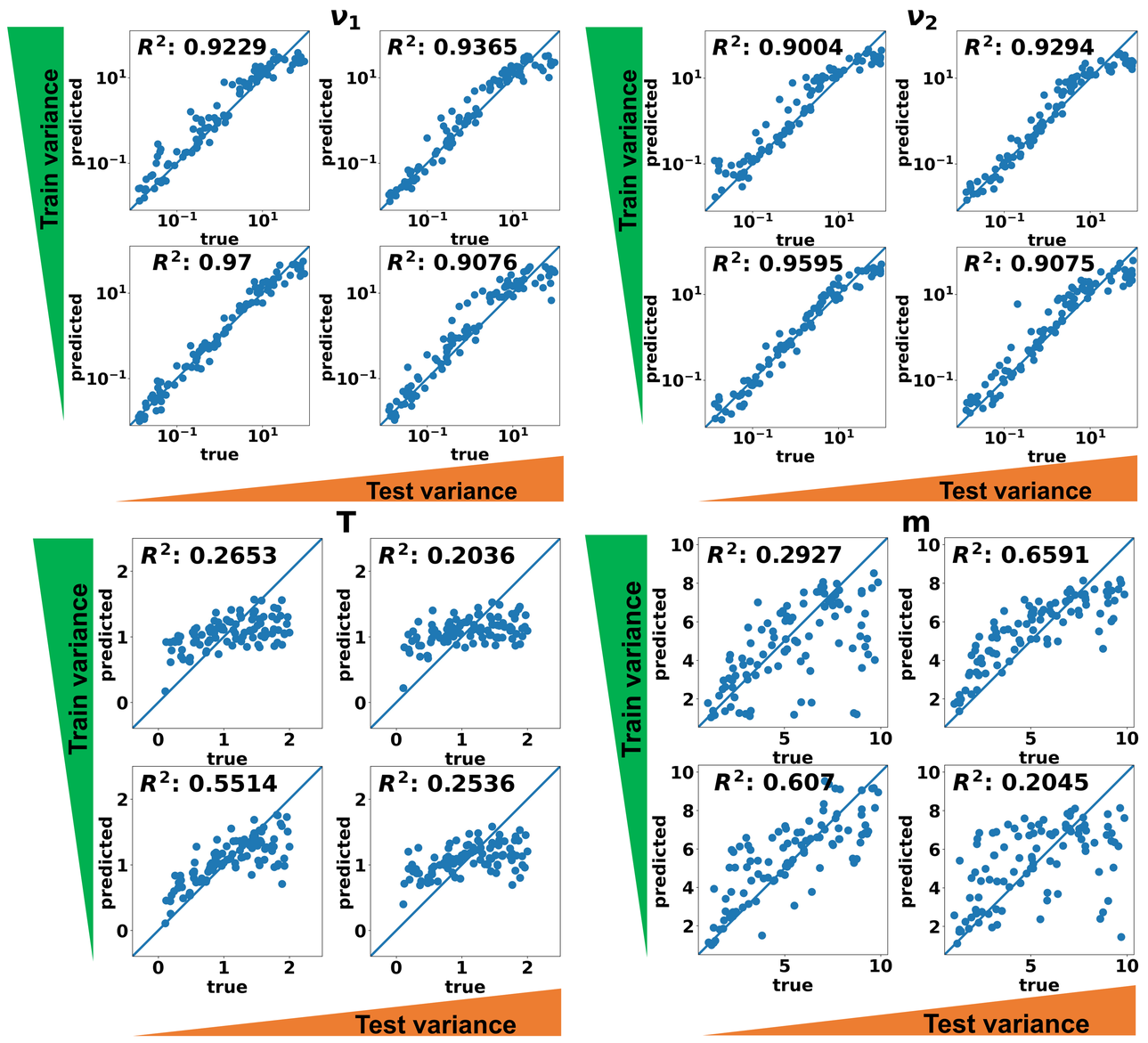
Benchmarking

To test whether a *dadi*+ML hybrid optimization approach could decrease the computational expense of the original dadi pipeline, we compared the run time for one round of optimization between optimizing on arbitrary starting parameters (*dadi* only) and optimizing on parameters predicted from AFS data by different trained RFR and MLPR models. We found that the hybrid approach inferred params with comparable to better accuracy (not shown) with about half the computing time.





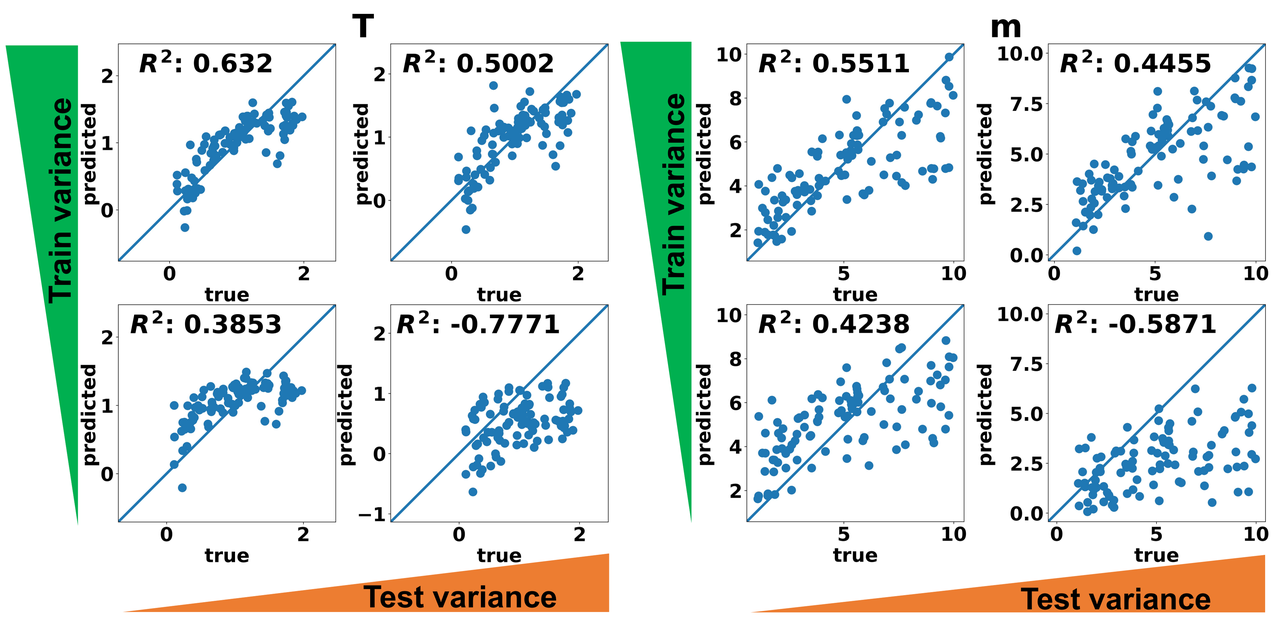
Random Forest Regressor



Here we plot the true against the predicted values for each parameter with R^2 scores quantifying prediction accuracy. Overall, the trained RFRs predicted ν_1 and ν_2 (top row) accurately but struggled to predict T and m (bottom row). We also observe that variance does not significantly affect prediction performance since the RFR performed similarly across variance cases.

Multilayer Perceptron Regressor

We also trained and tested an MLPR with 1 hidden layer of 2000 nodes paired with the Adam optimizer. The MLPR predicted ν_1 and ν_2 similarly well compared to the RFR (not shown) with a slight improvement in T and m predictions. MLPR appeared to be more sensitive to variance, with MLPR trained on data with variance outperforming MLPR trained on data without variance.



Acknowledgements

This research was supported by the National Institute of General Medical Sciences of the NIH (R01GM127348 to RNG).

References

1. Gutenkunst, R.N. et al., 2009. PLoS Genet, 5: e1000695.
2. Pedregosa, F. et al., 2011. J. of ML Research, 12: 2825–2830