Inferring Demographic History from Allele Frequency Spectra with Multi-layer Perceptron Regressors

Linh N. Tran[1,2], Connie K. Sun[2], Mathews Sajan[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

Objectives

We aim to develop a novel demographic history inference method that

- Provides instantaneous parameter estimation from an input allele frequency spectrum (AFS)
- Is likelihood-free
- Makes efficient use of all simulated data
- Provides uncertainty quantification
- Is interoperable with existing methods

Methods

We use fully-connected feedforward neural networks (scikit-learn MLPR) that take an AFS as input and output the expected parameter(s) for a specified demographic model.

Training data: dadi-simulated AFS labeled with corresponding demographic model parameter values.

Test data: (1) dadi-simulated AFS with added noise by scaling with θ and Poisson-sampling from the expected AFS; (2) msprime-simulated AFS to incorporate linkage.

For hyper-parameter optimization, we implemented the hyperband algorithm with scikit-learn successive halving random search.

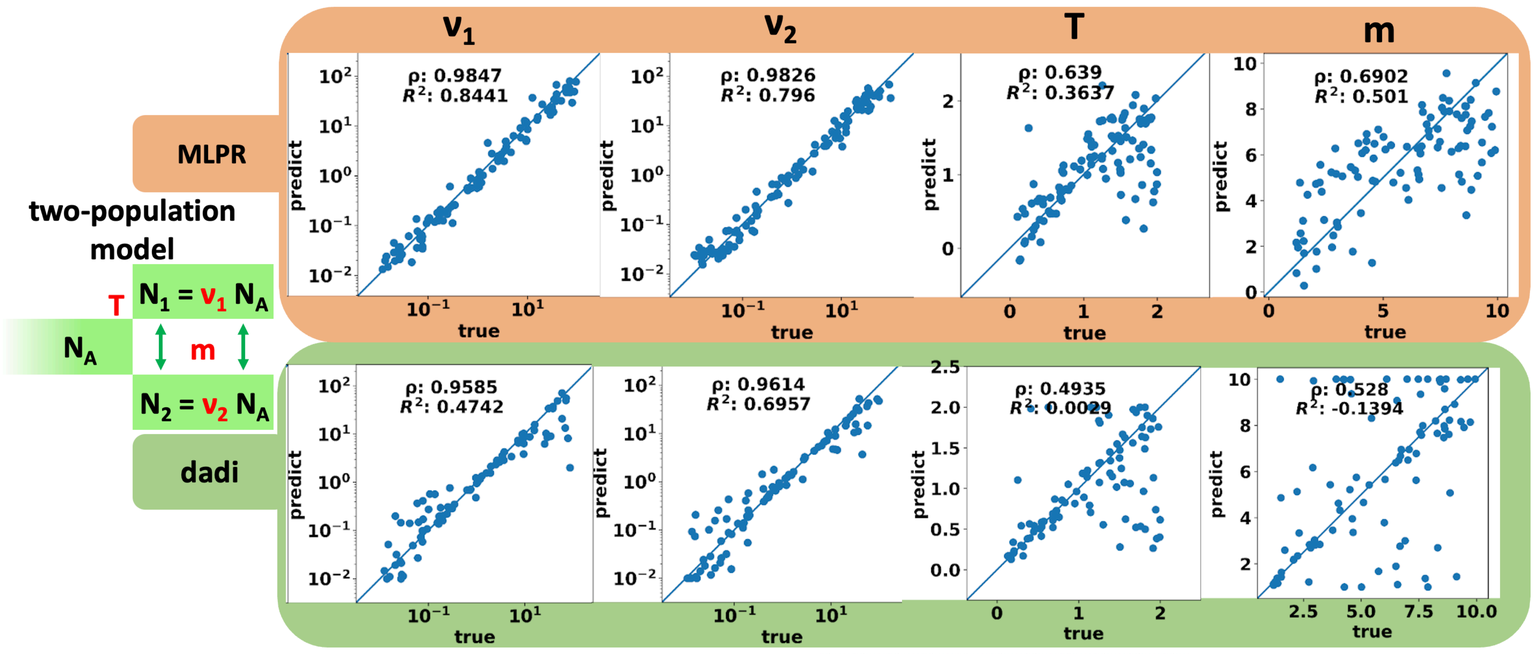
For uncertainty quantification, we use MAPIE, a scikit-learn-compatible package based on the jackknife+ method.

Results

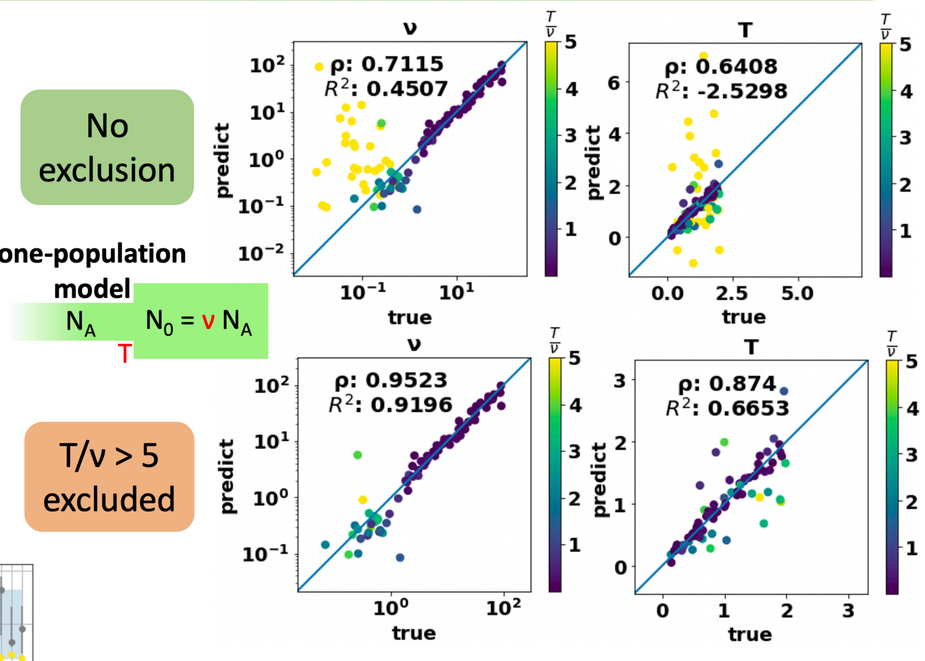
For tested one- and two-population demographic histories, the trained MLPRs can infer the population-size-change parameters (ν) very well (ρ≈0.98) and infer migration rate (m) and time (T) of event fairly well (ρ≈0.6-0.7).

They also perform similarly well when tested on AFS generated with msprime, under similar demography but including linkage, with r =10^{-8} -10^{-9} (data not shown).

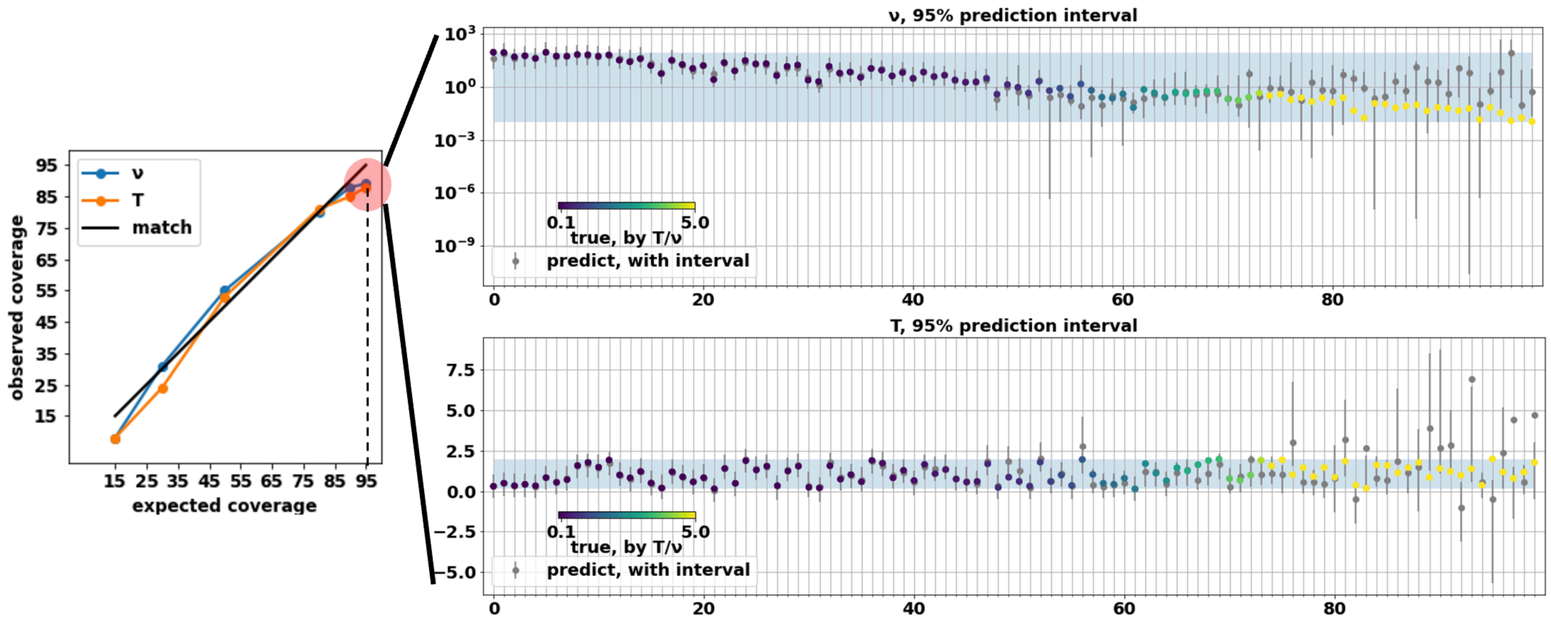
Importantly, trained MLPRs can estimate demographic parameters instantaneously from input AFS, with accuracy comparable to those inferred by the more time-consuming dadi likelihood optimization (shown here in the two-population model).



Certain input AFS can be uninformative for the specified demographic model. For example, in the one-population model shown, each yellow dot is an AFS generated with a true T and ν parameter pair where T/ν ≥ 5. This condition models a bottleneck event occurring a very long time ago (large T) in a contemporarily small population (small ν), whose signal is not detectable in the AFS. Consequentially, such input AFS are poorly predicted for both ν and T, and their removal significantly improves overall prediction accuracy.



In practice, it is not possible to specify the uninformative parameter ranges for every demographic model. Therefore, we implemented prediction interval estimation for each parameter. As shown below, our method can match the expected uncertainty coverage, and most estimations on uninformative AFS have appropriately wide intervals.



Discussion

The results shown here are for unfolded AFS with a sample size of 20 chromosomes per population and assumed correct ancestral state identification. We are currently developing more trained MLPRs to support other demographic scenarios, different sample sizes, as well as cases where the ancestral state is unknown (folded AFS) or mis-specified (by co-estimating a mis-identification parameter).

References

dadi: Gutenkunst, R.N. et al., 2009. PLoS Genet, 5: e1000695.

msprime: Kelleher, J. et al., 2016. *PLoS comp. biology* 12.5: e1004842.

MAPIE: https://github.com/scikit-learn-contrib/MAPIE

scikit-learn: Pedregosa, F. et al., 2011. J. of ML Research, 12: 2825–2830

hyperband: Li, L. et al., 2017. J. of ML Research 18.1: 6765-6816.

jackknife+: Barber, R. F. et al., 2021. The Annals of Statistics 49.1: 486-507.

Acknowledgements

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