Machine Learning in Population Genetics and Phylogenetics

Section 1: Reviews
Section 2: Population Genetics
Section 3: Phylogenetics

Disclaimer*:* This is a non-exhaustive list of papers I referenced and/or like that use machine learning (or discuss machine learning) in population and phylogenetics. There are many, many more papers on these topics. The summaries are very vague and short, and are only meant to help you identify papers that might be of interest to you. Several of these papers are preprints, meaning they have not gone through peer review.

1. Reviews

Schrider, D.R. and Kern, A.D., 2018. Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, 34(4), pp.301-312. doi: https://doi.org/10.1016/j.tig.2017.12.005

Summary: This paper offers a nice overview of supervised machine learning approaches and discusses some of the earliest applications of these approaches in population genetics.

Korfmann, K., Gaggiotti, O.E. and Fumagalli, M., 2023. Deep learning in population genetics. *Genome Biology and Evolution*, 15(2), p.evad008. doi: https://doi.org/10.1093/gbe/evad008

Summary: This is a very recent review of deep learning approaches in population genetics and includes nice discussions of current challenges and future directions.

Fountain-Jones, N.M., Smith, M.L. and Austerlitz, F., 2021. Machine learning in molecular ecology. *Molecular Ecology Resources*, 21(8), pp.2589-2597. doi: https://doi.org/10.1111/1755-0998.13532

Summary: This paper is an introduction to a special issue in Machine Learning in *Molecular Ecology Resources* and should offer a nice road-map to some of the papers in that special issue.

2. Population Genetics and Phylogeography

2.1. Decision tree approaches

Smith, M.L. and Carstens, B.C., 2020. Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*, 74(2), pp.216-229. doi: https://doi.org/10.1111/evo.13878

Summary: This paper introduces *delimitR*. *delimitR* uses Random Forests and the Site Frequency Spectrum to compare models that differ both in the number of populations and the presence or absence of gene flow between populations.

Schrider, D.R., Ayroles, J., Matute, D.R. and Kern, A.D., 2018. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genetics*, 14(4), p.e1007341. doi: https://doi.org/10.1371/journal.pgen.1007341

Summary: This paper uses Extra Trees classifiers and a set of hand-crafted summary statistics to classify genomic windows as introgressed between two closely-related populations or species.

2.2. Fully Connected Neural Networks

Sheehan, S. and Song, Y.S., 2016. Deep learning for population genetic inference. *PLoS Computational Biology*, 12(3), p.e1004845. doi: https://doi.org/10.1371/journal.pcbi.1004845

Summary: This paper uses a deep neural network and hand-crafted summary statistics to infer population sizes through time and detect selection.

2.3. Convolutional Neural Networks

Flagel, L., Brandvain, Y. and Schrider, D.R., 2019. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular Biology and Evolution*, 36(2), pp.220-238. doi: https://doi.org/10.1093/molbev/msy224

Summary: This paper introduced Convolutional Neural Networks (CNNs) as a potentially useful supervised machine learning approach in population genetics. They offer an overview of CNNs and apply them to several problems in population genetics.

Fonseca, E.M., Colli, G.R., Werneck, F.P. and Carstens, B.C., 2021. Phylogeographic model selection using convolutional neural networks. *Molecular Ecology Resources*, 21(8), pp.2661-2675. doi: https://doi.org/10.1111/1755-0998.13427

Summary: This paper uses CNNs to compare phylogeographic models.

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S. and Fumagalli, M., 2019. ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC bioinformatics*, 20(337), pp.1-12. doi: https://doi.org/10.1186/s12859-019-2927-x

Summary: This paper uses a CNN to detect selection.

2.4. Recurrent Neural Networks

Adrion, J.R., Galloway, J.G. and Kern, A.D., 2020. Predicting the landscape of recombination using deep learning. *Molecular biology and evolution*, *37*(6), pp.1790-1808. doi: https://doi.org/10.1093/molbev/msaa038

Summary: This paper uses Recurrent Neural Networks to estimate recombination rates across the genome.

Hejase, H.A., Mo, Z., Campagna, L. and Siepel, A., 2022. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Molecular Biology and Evolution*, 39(1), p.msab332. doi: https://doi.org/10.1093/molbev/msab332

Summary: This paper uses Recurrent Neural Networks to detect selection (and estimate selection coefficients) from the ancestral recombination graph.

2.5. Generative Models

Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H.H., Mathieson, I. and Mathieson, S., 2021. Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, 21(8), pp.2689-2705. doi: https://doi.org/10.1111/1755-0998.13386

Summary: This paper implements a GAN to infer demographic parameters under several models. This allows for a heuristic exploration of parameter space.

2.6. Image Segmentation

Ray, D., Flagel L., and Schrider, D.R. 2023. IntroUNET: identifying introgressed alleles via semantic segmentation. *bioRxiv.* doi: https://doi.org/10.1101/2023.02.07.527435

Summary: This paper uses image segmentation to identify introgressed haplotypes.

2.7. Domain Adaptation

Mo, Z. and Siepel, A., 2023. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. *bioRxiv*. doi: https://doi.org/10.1101/2023.03.01.529396

Summary: This paper implements domain-adaptive neural networks in ReLERNN (Adrion et al., 2020) and SIA (Hejase et al., 2022). They are able to design networks that are more robust to simulation misspecification by implementing a gradient reversal layer.

3. Phylogenetics

3.1. Inferring tree topologies (or concordance)

Suvorov, A., Hochuli, J. and Schrider, D.R., 2020. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic biology*, 69(2), pp.221-233. doi: https://doi.org/10.1093/sysbio/syz060

Summary: This paper uses a Convolutional Neural Network to infer quartet topologies from sequence alignments.

Zou, Z., Zhang, H., Guan, Y. and Zhang, J., 2020. Deep residual neural networks resolve quartet molecular phylogenies. *Molecular Biology and Evolution*, 37(5), pp.1495-1507. doi: https://doi.org/10.1093/molbev/msz307

Summary: This paper implements a similar approach to Suvorov et al., (2020) to infer quartet topologies from sequence alignments.

Solis-Lemus, C., Yang, S. and Zepeda-Nunez, L., 2022. Accurate phylogenetic inference with a symmetry-preserving neural network model. *arXiv* preprint arXiv:2201.04663. doi: https://doi.org/10.48550/arXiv.2201.04663

Summary: This paper improves upon the approaches of Suvorov and Zou to infer quartet topologies from sequence alignments.

Zaharias, P., Grosshauser, M. and Warnow, T., 2022. Re-evaluating deep neural networks for phylogeny estimation: the issue of taxon sampling. *Journal of Computational Biology*, 29(1), pp.74-89. doi: https://doi.org/10.1089/cmb.2021.0383

Summary: This paper evaluates the effectiveness of the approaches introduced by Zou et al. (2020) and Suvorov et al. (2020) and suggest that the limitation to inferring quartet topologies prevents these methods from being competitive.

Smith, M.L. and Hahn, M.W., 2022. Phylogenetic inference using Generative Adversarial Networks. *bioRxiv*. doi: https://doi.org/10.1101/2022.12.09.519505

Summary: This paper implements a Generative Adversarial Network (GAN) to heuristically explore tree space and infer topologies.

Rosenzweig, B., Kern, A. and Hahn, M., 2022. Accurate Detection of Incomplete Lineage Sorting via Supervised Machine Learning. *bioRxiv*. doi: https://doi.org/10.1101/2022.11.09.515828

Summary: This paper implements several networks, including a Fully Connected Neural Network, to infer the proportion of concordant gene trees at a branch of interest. This facilitates correcting for gene tree estimation error when external branch lengths are long.

3.2. Inferring branch lengths

Suvorov, A. and Schrider, D.R., 2022. Reliable estimation of tree branch lengths using deep neural networks. *bioRxiv*. doi: https://doi.org/10.1101/2022.11.07.515518

Summary: This paper uses a CNN to infer branch lengths for a quartet tree.

3.3. Inferring phylodynamic and macroevolutionary parameters

Voznica, J., Zhukova, A., Boskova, V., Saulnier, E., Lemoine, F., Moslonka-Lefebvre, M. and Gascuel, O., 2022. Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Communications*, 13(1), p.3896. doi: https://doi.org/10.1038/s41467-022-31511-0

Summary: This paper introduces a clever encoding of gene tree topologies that enables the use of CNNs on gene tree inputs. They use this and other approaches to estimate parameters of birth-death models.

Lambert, S., Voznica, J. and Morlon, H., 2022. Deep Learning from Phylogenies for Diversification Analyses. *bioRxiv*. doi: https://doi.org/10.1101/2022.09.27.509667

Summary: This paper applies approaches similar to those of Voznica et al., (2022) to estimate speciation and extinction rates.

Lajaaiti, I., Lambert, S., Voznica, J., Morlon, H. and Hartig, F., 2023. A Comparison of Deep Learning Architectures for Inferring Parameters of Diversification Models from Extant Phylogenies. bioRxiv. doi: https://doi.org/10.1101/2023.03.03.530992

Summary: This paper compares several approaches, including Graphical Neural Networks, to estimate speciation and extinction rates.

Thompson, A., Liebeskind, B., Scully, E.J. and Landis, M., 2023. Deep learning approaches to viral phylogeography are fast and as robust as likelihood methods to model misspecification. *bioRxiv*. doi: https://doi.org/10.1101/2023.02.08.527714

Summary: They compare likelihood and machine learning methods to estimate parameters of birth-death models when there is model misspecification.