Inferring effective population sizes of bacterial populations while accounting for unknown recombination and selection: a deep learning approach

Jean Cury[1,4], Théophile Sanchez[1], Erik Bray[1], Jazeps Medina-Tretmanis[2], Maria Avila-Arcos[3], Emilia Huerta-Sanchez[2], Guillaume Charpiat[1], and Flora Jay[1]

[1] Université Paris-Saclay, CNRS, INRIA, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
[2] Center for Computational Molecular Biology, Brown University, Providence, RI, USA
[3] International Laboratory for Human Genome Research, Universidad Nacional Autónoma de México (UNAM), Querétaro, México
[4] SEED, U1284, INSERM, Université de Paris, Paris, France.

Abstract. Inferring population size through time is a long-standing problem in population genetics. It consists, essentially, in reconstructing the demography of a population in the past, based on a sample in the present of the population. Many types of methods have been developed for decades, but it is only recently that deep learning based methods started to emerge. It has been shown, however, that in the case of bacterial populations, classical methods do not work, because the underlying assumption of these methods were not satisfied. Here, we design and evaluate how an end-to-end deep learning approach that accounts for unknown recombination and selection events performs on bacterial populations. We also propose various improvements to this framework, such as implementing uncertainty estimation.

Keywords: population genetics · deep learning · Bacteria

1 Introduction

The evolutionary history of a population is imprinted in present-day individuals' DNA. These remains of information are used to reconstruct various evolutionary signals. They can be used to infer, for instance, demographic changes, selection, or migration events. The genetic diversity observed in a sample of sequenced individuals from a population is at the basis of population genetic inferences. Diverse methods exist; some require simulations in addition to real data, while others require only real data. In bacterial population genetics, most of the work on demographic inference was done using skyline plots [6], but it has been shown that this method was not suitable for bacterial populations [9].

In the last years, methods based on deep learning emerged [11]. The first ones started using summary statistics as an intermediate representation of the input data[13], whereas more recent ones directly used the raw genetic diversity as input data[4, 10]. These latter approaches were coined end-to-end deep learning approaches. Although in their infancy, they achieved performances comparable to long-standing methods. In this work, we present an end-to-end deep learning approach for the inference of bacterial population's past demography, based on a network we previously designed for a similar task (yet without selective nor bacterial recombination processes) [10]. Our general approach is based on dnadna [11], a framework we developed that facilitates the use, reuse, and sharing of population genetics neural networks. We also highlight how this package is helpful for the community. Finally, we propose various improvements to current settings, notably in terms of uncertainty quantification, which is valuable when inferring an intrinsically stochastic process, and even more so when models train on simulations rather than real datasets.

2 Methods

2.1 Simulations

The task of inferring effective population size through time is a regression task in a context of supervised learning. Thus, it requires a training procedure on a labelled dataset. There are no real datasets for which we know the exact population size through time, thus we rely on a simulated dataset to train and test our network. The simulations were done using SLiM, a forward-in-time simulator that we adapted to match bacterial populations more closely [2]. Notably, we made possible bacterial recombination, which is similar to gene conversion, but between different bacteria. It corresponds to horizontal gene transfer of homologous DNA. In combination with msprime[7], it allows simulating a wide range of various scenario while starting at the mutation-drift equilibrium.

We parametrized the simulation after the bacterial species Streptococcus agalactiae. We simulated three types of demography: bottleneck without recovery (a sudden reduction in population size), constant size, immediate expansion. For each type of demography, half of the datasets were simulated under neutrality and half in presence of an allele under positive selection. We randomly draw most of the simulation parameters using a generalized Halton sequence, which allows a better coverage of the parameter space compared to drawing from a uniform distribution in n-dimensional space. Varying parameters were: the mutation rate, the recombination rate (as a ratio of the mutation rate), the initial population size, the time of population size change (bottleneck or expansion), the strength of this variation, the time of apparition of the allele under selection and the strength of the selection. We kept other parameters constant across simulations, such as the chromosome size (2.065 Mb), the number of generations (21900) and the mean size of the recombination track length (122 kb). We define a scenario as being a fixed set of parameters used for the simulation, and each scenario has 100 replicates (so 100 simulations with the same parameters). We simulated

11000 scenarios, among which 5000 with the bottleneck or expansion settings, and 1000 for the constant size scenario. Among each, half contains selection, and the other half is without. At the end of the simulations (i.e., corresponding to the present), we sampled 600 individuals.

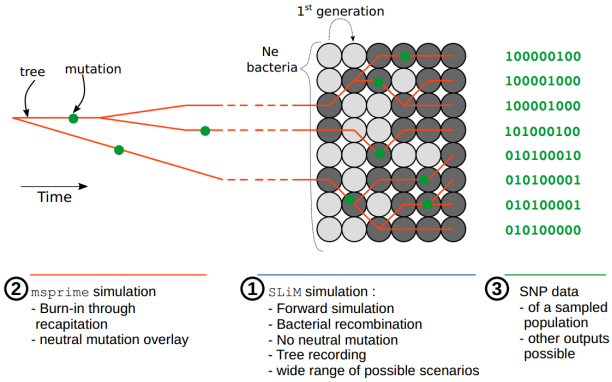


Fig. 1: Schematic of the simulation pipeline. 1/ Forward simulation of bacterial population. 2/ Recapitation with msprime (ie. generation of the ancient part of the coalescence trees) 3/ A SNP matrix outputted by the simulator. Figure adapted from [2]

2.2 Input data

Real data or simulated data are represented in the form of a matrix of SNPs (single nucleotide polymorphisms) of dimension N x S, with N individuals and S SNPs. Each simulation has a different number of SNPs. Scenarios for which any of the replicate had less than 400 SNPs were discarded. At the end, 8629 scenarios were kept, out of which 6041 constituted the training set, 1295 the validation set, and 1293 the test set. While training, we subsampled 20 of the 600 individuals of each matrix each time a new batch was built.

2.3 Deep Neural Network architecture

We use the SPIDNA architecture that was previously developed in the lab [10]. This architecture is invariant to the permutation of individuals (the lines of the SNP matrix), and to the input dimension. It is based on 7 blocks that perform equivariant operations and progressively reduce the data dimension. Each block contains (i) 50 1D-convolutional filters treating each 'individual' (line) equivalently, (ii) an invariant operation (mean) that aggregates individual features into global ones, (iii) the concatenation of individual and global features, which will be mixed in the next block. At each block a subset of 21 invariant features contributes directly to the prediction through a fully connected layer.

2.4 dnadna

dnadna is a python package developed recently in the lab [11] that allows researchers not proficient in deep learning or even coding to use such methods on population genetic datasets. At the same time, advanced users can create their own network and share it easily so that the community can reuse it for prediction or training on another task. A plugin system allows anyone to easily create new components (network, loss, transformations) for the training procedure.

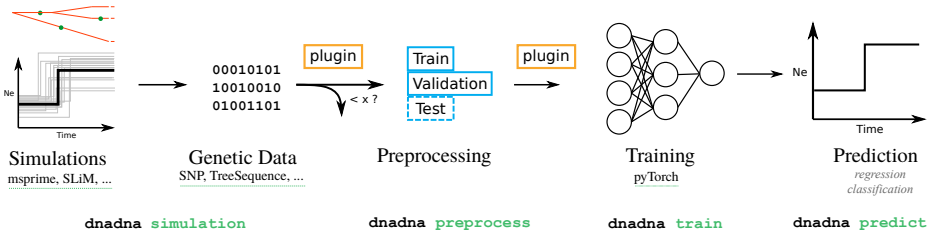


Fig. 2: Pipeline of the different command implemented in dnadna

2.5 Uncertainty estimation

Uncertainty in deep learning and machine learning in general can be characterized in two types of uncertainty. First, aleatoric uncertainty corresponds to the randomness that affects the data itself. Because evolutionary processes are random, trying to capture such uncertainty is essential in population genetics. Second, epistemic uncertainty corresponds to the fact that the training set does not cover the entirety of the space of possible inputs. This is highly relevant in population genetics as we use simulations and real data may lie outside the space defined by the simulations (reality gap).To capture aleatoric uncertainty, we use instead of the classic MSE loss, the negative log likelihood loss for each parameter [8]. This loss takes as input 2 parameters (instead of 1 for the MSE), the mean and the variance of the value to be predicted. Hence, the networks is trained to output the parameters defining a Gaussian posterior for each targeted parameter. This has been investigated in the context of recombination hotspot inference [1]. Concerning the epistemic uncertainty, various methods have been proposed. We use the deep ensemble method, which consists in training a set of networks (typically 5) independently. These are considered as uniformely-weighted Gaussian mixture model for which predictions are combined [8].

3 Results

We use SPIDNA to infer the population size through time. Although there were only two different population sizes in the simulations (before and after the bottleneck/expansion), we decided to predict 21 time steps as in the original SPIDNA

paper [10]. Similarly to other tools commonly used in the field [3, 12, 14], this flexible modelling has the advantage of being agnostic to the type of population size trajectory.

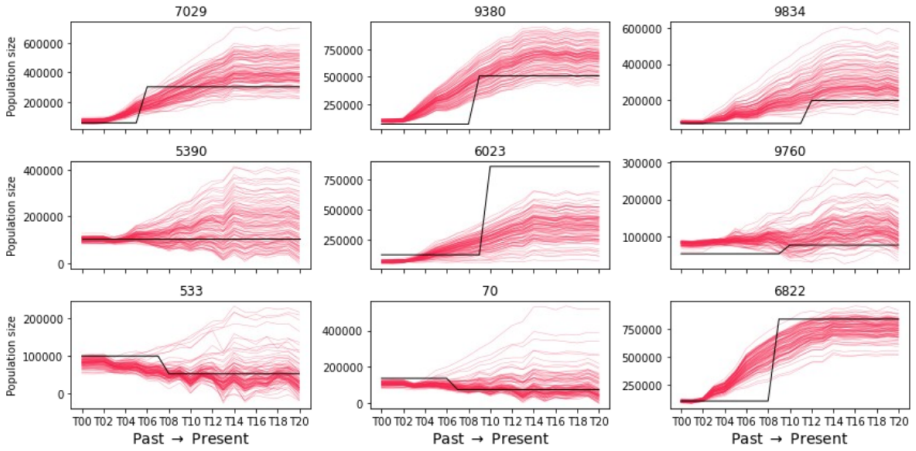


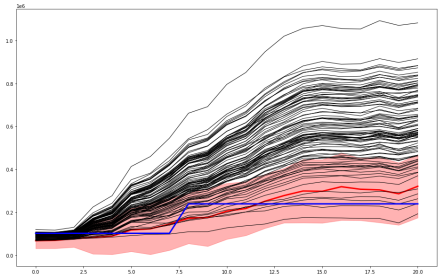
Fig. 3: Prediction of the population size with SPIDNA on the validation set for 9 randomly picked scenarios. The black line is the true population size used for the simulation, and in red is the prediction for the 100 replicates.

Fig. 3 presents the estimated histories for a subset of nine scenarios (the MSE over the validation dataset is 0.63). We observe that the trend of the predicted (red) and true (black) histories are often similar, although the predicted changes are smooth rather than sudden. This confirms that the method captures informative signal in the data indicative of past history, but that it may be difficult to associate it to an exact time of change unless sudden changes are enforced as done in model-based inference (with e.g. *dadi* [5]).

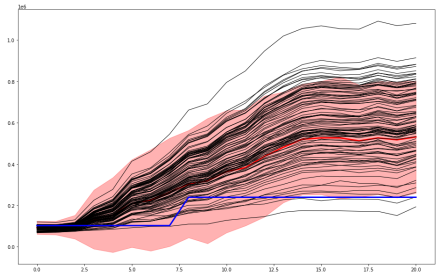
Nonetheless, the accuracy of the inference varies. When focusing on multiple replicates of a same scenario, we observe a variability of the predictions, usually high in recent times and lower in the ancient past. It is due to the underlying variability of the replicates generated by the stochastic evolutionary processes. The prediction of the variability could hence be an indication of the uncertainty of the method regarding the reconstruction of a given scenario. However, when predicting the history from real data, we only have one replicate (corresponding to a single circular chromosome for a sample of individuals).

Thus, we applied our new SPIDNA network that models the target parameters as Gaussian distributions. It enables to quantify an uncertainty per replicate rather than per scenario. We see in the figure 4 that the prediction of the uncertainty is smaller when the network makes a good prediction, and conversely, the uncertainty is larger when the prediction is farther off. In Figure 5, we see that on average, a better prediction of the population size at a given time step(lower

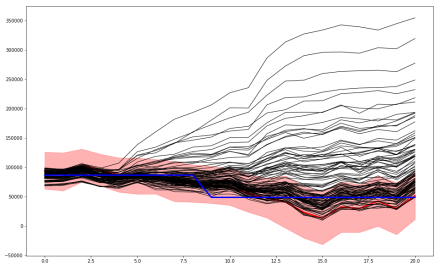
MSE) is associated with a smaller uncertainty (lower predicted std). This is especially true for the time steps where a change of population occurs (basically, after the step 4). For the first steps (most ancient times), the prediction is often accurate (also because the variability in prediction is relatively small here), as seen in figures 3 and 4.



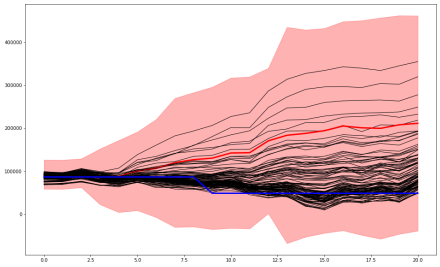
(a) Expansion : good prediction and small uncertainty



(b) Expansion : bad prediction and high uncertainty



(c) Bottleneck : good prediction and small uncertainty



(d) Bottleneck : bad prediction and high uncertainty

Fig. 4: Example of scenarios where the uncertainty estimation differs depending on the accuracy of the prediction. The top row corresponds to the same scenario during which an expansion occurred, while the bottom row corresponds to another scenario where a bottleneck occurred. Each black line is the prediction for the 100 replicates of the given scenario. The blue line is the true demography. The red line is the prediction of one of the replicates chosen to be close to the true value or not. The red shaded area is the standard deviation as predicted by the network (predicted mean +/- predicted std)

4 Conclusion

Overall, we proposed a set of tools to use deep learning methods for bacterial population genetics, from a bacterial simulator to the easy-to-use dnadna framework for users who are not proficient in coding and machine learning. We

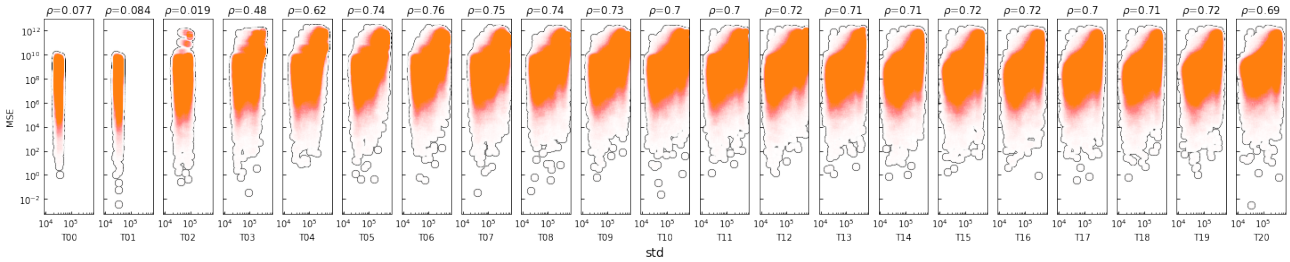


Fig. 5: Relation between the predicted error (mean square error) and the predicted standard deviation for each parameter (time steps). For a given plot, corresponding to a time step, the MSE of a given replicate is plotted against the predicted standard deviation for this replicate, across all scenarios. We computed the spearman ρ to assess whether worse MSE is correlated to higher uncertainty (i.e. larger predicted std). All p-values are below 10^{-10}.

demonstrated the practicality of this pipeline and confirmed that deep neural networks have great potential, despite the particular complexity of the task (inferring effective population sizes without knowing the recombination parameters or selective pressure). Furthermore, we enhanced the pipeline by providing uncertainty estimations, a key feature for a broader adoption of machine learning approaches by the microbiologist and population genetics communities. Another feature that could foster deep learning applications is the use of transfer learning, where users avoid retraining a model from scratch, and instead adjust an existing model to the specificity of their dataset and task. We are currently adding this feature within dnadna, and will present the assessment of its impact. Transfer learning is an important step to decrease the training time and energy cost of deep learning tools. Indeed, we must not ignore the impact of deep learning on energy consumption, and thus we must seek to decrease that cost. Another advantage of having both uncertainty estimation and transfer learning, is to deploy this type of method for ancient DNA. With such data, there is an additional source of randomness due to the higher rate of sequencing error, which makes uncertainty estimation even more essential. Moreover, the high cost of simulating ancient DNA (reproducing DNA degradation, low coverage, etc.) places us in the context of training on small sets, where transfer learning is helpful.

Acknowledgements TAU GPU platform (Titanic) and Kepler for computing resources. ASARD team (LISN). DIM One Health 2017 (number RPH17094JJP), Human Frontier Science Project (number RGY0075/2019), ANR-20-CE45-0010-01 RoDAPoG, ATIP-Avenir program from INSERM (R21042KS / RSE22002KSA) for funding.

References

1. Chan, J., Perrone, V., Spence, J., Jenkins, P., Mathieson, S., Song, Y.: A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.
2. Cury, J., Haller, B.C., Achaz, G., Jay, F.: Simulation of bacterial populations with SLiM. Peer Community Journal **2** (2022).
3. Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A.: Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution **29**(8), 1969–1973 (Aug 2012).
4. Flagel, L., Brandvain, Y., Schrider, D.R.: The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. Molecular Biology and Evolution **36**(2), 220–238 (Feb 2019).
5. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D.: Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genetics **5**(10), e1000695 (Oct 2009).
6. Ho, S.Y.W., Shapiro, B.: Skyline-plot methods for estimating demographic history from nucleotide sequences: INVITED TECHNICAL REVIEW. Molecular Ecology Resources **11**(3), 423–434 (May 2011).
7. Kelleher, J., Etheridge, A.M., McVean, G.: Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology **12**(5), e1004842 (May 2016).
8. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles
9. Lapierre, M., Blin, C., Lambert, A., Achaz, G., Rocha, E.P.C.: The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography. Molecular Biology and Evolution **33**(7), 1711–1725 (Jul 2016).
10. Sanchez, T., Cury, J., Charpiat, G., Jay, F.: Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Molecular Ecology Resources **21**(8), 2645–2660 (2020).
11. Sanchez, T., Madison Bray, E., Jobic, P., Guez, J., Letournel, A.C., Charpiat, G., Cury, J., Jay, F.: dnadna: deep neural architectures for DNA - A deep learning framework for population genetic inference
12. Schiffels, S., Durbin, R.: Inferring human population size and separation history from multiple genome sequences. Nature Genetics **46**(8), 919–925 (Aug 2014).
13. Sheehan, S., Song, Y.S.: Deep Learning for Population Genetic Inference. PLOS Computational Biology **12**(3), e1004845 (Mar 2016).
14. Terhorst, J., Kamm, J.A., Song, Y.S.: Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature Genetics **49**(2), 303–309 (Feb 2017).