Deep learning based methods for estimating distribution of coalescence rates from genome-wide data

Deep learning based methods for estimating distribution of coalescence rates from genome-wide data

Evgeniy Khomutov

International Laboratory of Statistical and Computational Genomics, National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow, 101000, Russia

E-mail: evkhomutov@miem.hse.ru

Kenenbek Arzymatov

Laboratory of Methods for Big Data Analysis, National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow, 101000, Russia

E-mail: karzymatov@hse.ru

Vladimir Shchur

International Laboratory of Statistical and Computational Genomics, National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow, 101000, Russia

E-mail: vshchur@hse.ru

November 2020

Abstract. Demographic and population structure inference is one of the most important problems in genomics. Population parameters such as effective population sizes, population split times and migration rates are of high interest both themselves and for many applications, e.g. for genome-wide association studies. Hidden Markov Model (HMM) based methods, such as PSMC, MSMC, coalHMM etc., proved to be powerful and useful for estimation of these parameters in many population genetics studies. At the same time, machine and deep learning have began to be used in natural science widely. In particular, deep learning based approaches have already substituted hidden Markov models in many areas, such as speech recognition or user input prediction. We develop a deep learning (DL) approach for local coalescent time estimation from one whole diploid genome. Our DL models are trained on simulated datasets. Importantly, demographic and population parameters can be inferred based on the distribution of coalescent times. We expect that our approach will be useful under complex population scenarios, which cannot be studied with existing HMM based methods. Our work is also a crucial step in developing a deep learning framework which would allow to create population genomics methods for different genomic data representations.

Keywords: Recurrent neural networks, RNN, deep learning, machine learning, population genomics, genomics, population structure.

1. Introduction

Genomic data is a rich source of information about population structure and history. A lot of relations between individuals and whole populations are encrypted there. For example, one can see the out-of-Africa expansion in human history from a single non-African individual [1]. This event is represented as a bottleneck in the population history, in other words the population size during the corresponding period of time appears to be small. There are other interesting characteristics, or parameters, of population histories: population split times, migration rates, admixture times and proportions. All these parameters affect the distribution of times to the most recent common ancestor of individuals representing populations, or shortly coalescent times. Due to recombinations, coalescent times change across a genome. And given that genome is a long sequence (e.g. human genome is approximately $3x10^9$bp) even one diploid genome can provide a reliable estimation of the coalescent time distribution. So, inference of coalescent times from real data (e.g. from a single diploid genome or multiple genomes) is a crucial step underlying the inference of population parameters and population history.

Hidden Markov Model (HMM) based methods proved to be powerful for demographic and population analysis from genome-wide sequences. The benefit of using HMMs is that it allows to infer population history based on the local structure of genome-wide sequences and allow to model recombinations. This is an important difference from site frequency spectrum (SFS) based methods such as dadi [7] and momi [8]. A number of methods (e.g. PSMC, diCal, MSMC, MSMC2, SMC++, ASMC [1–3]) based on HMM were implemented. The population models underlying these methods is SMC [4] and its corrected version SMC' [5]. Both of these models consider genealogies as a set of coalescent trees across the genomes, with neighbour trees connected by recombination which can be viewed as a subtree prune-and-regraft process. Pruning corresponds to breaking an ancient lineage by recombination, while regrafting corresponds to recoalescence of a prunned lineage back on the coalescent tree.

In the coalescent HMM, hidden variables are coalescent times which change along the genome due to recombinations. Observations are homozygous and heterozysous sites in the genome sequences. Heterozygous site arises as a result of a mutation in the sample history.

For coalescent HMM, transition and emission matrices are parametrised by population parameters [1]. Most of the methods assume a single population model with effective population size varying over time. (e.g. PSMC, SMC++). Other methods require phasing (e.g. MSMC, coalHMM, MSMC-im [6]). Some of the methods work with special data representations (ASMC for sparse data, unpublished method ngsPSMC for genotype likelihoods [11].

Extending HMM-based methods for wider class of population models, and for new types of input data is always rather challenging and requires substantial efforts in mathematical model derivation, software implementation and method verification. Our goal is to develop a machine learning framework which would allow to make inference under complex population models (structured populations with admixture and continuous migration) and would allow straightforward extension (by training a model on a simulated dataset) for different input data types (e.g. multiple genomes, phased/unphased genomes, called variants or genotype likelihoods). Currently we study performance of different neural network architectures, their precision and scalability in the simplest case of a single diploid genome and a single population scenario. We present the model which predicts local coalescent times across the genome.

2. Method

2.1. Training dataset

We train our models on simulated datasets, because it is not feasible to obtain empirical data with the ground truth local coalescent times. We developed a random generator of population scenarios. Overall, it samples times of demographic events (changes of effective population size), population sizes and some other population parameters from prior distributions. For each generated population scenario, we run *msprime* [10] to simulate chromosomes along with known local coalescent trees.

Currently we have the following process for generating population scenarios (each described by a picewise constant effective population size history).

- Sample the number N of effective population size changes from the uniform distribution $U[1; 20]$.
- Sample time T_{max} of the deepest (in the past) change of the effective population size from distribution.
- We sample times between population size changes from exponential distribution on the logarithmic time scale: first we sample t_i from $exp(1)$ and then re-scale them using the following formula:

$$T_i(t_i) = \alpha \exp(\beta\ t_i),$$

  where α is a scaling parameter and $\beta = \frac{T_{max}}{max_i(t_i)}$. This is the time scale used in PSMC.
- For each time interval we sample effective population size independently from Beta distribution $B(2; 5)$

For each demographic scenario we generate a diploid genome (two haplotypes) together with local coalescent times with msprime coalescent simulator.

We choose to solve the classification problem (estimating probabilities that local coalescent time falls within a certain time range) instead of regression problem

(predicting a single most likely value of coalescent time). This strategy is similar to the HMM posterior decoding, and provides us with the tractable uncertainty of our estimates.

2.2. Data preparation

2.2.1. Data splitting The input dataset d represents a set of SNPs $d = \{d_1, d_2, \ldots, d_n\}$, where n is an amount of available genome sequences, and each sequence d_i has a length of 30 millions numbers of either 0 or 1. Due to limitations of available memory in GPU, every sequence d_i is divided into segments of length seg_len, i.e $d_i = \{d_{i1}, d_{i2}, \ldots, d_{iL}\}$, where L – is a number of segments. As a result, training dataset consists of $d = \{d_{11}, d_{12}, \ldots, d_{1L}, d_{21}, \ldots, d_{2L}, \ldots, d_{n1}, \ldots, d_{nL}\}$ segments. It's worth noting that the main limitation of the approach in this paper is that neural network will be trained on each segment independently of other surrounding segments from the same genome.

2.2.2. Neural network architecture Our first attempt was to build recurrent NN models (GRU, LSTM) [13, 14] because it is natural to use them for sequential data. However, they failed to capture a reasonable signal in the data. So, we added a CNN layer [12] that extracts low-level features. Also, several convolutional layers can extract more advanced features based on combinations from previous layers. As a result, a combination of CNN and RNN layers give a satisfactory output.

3. Results

In order to evaluate the results, we visualise the ground truth and predicted coalescent times along the genomes (see Fig 1). Each panel represent a genome segment. Y-axis corresponds to the time intervals. The inferred probabilities of the local coalescent time being at a certain time interval, are encoded as a heatmap. The ground truth coalescent times are shown with the black line.

In general, our neural network predicts the MRCA time for most positions. Though, there are some artefacts like "oscillations" around the true constant values (see Fig 1 c and d). More studies should be done to resolve the issue. Also, quantitative evaluations of the prediction quality should be done.

4. Summary

In this work we validated the deep learning approach to the population analysis from whole genome sequences. We present a deep neural network which can reasonably well predict the local coalescent times across the genome. We considered only simplest demographic scenarios. In future we will improve the accuracy of our method and we

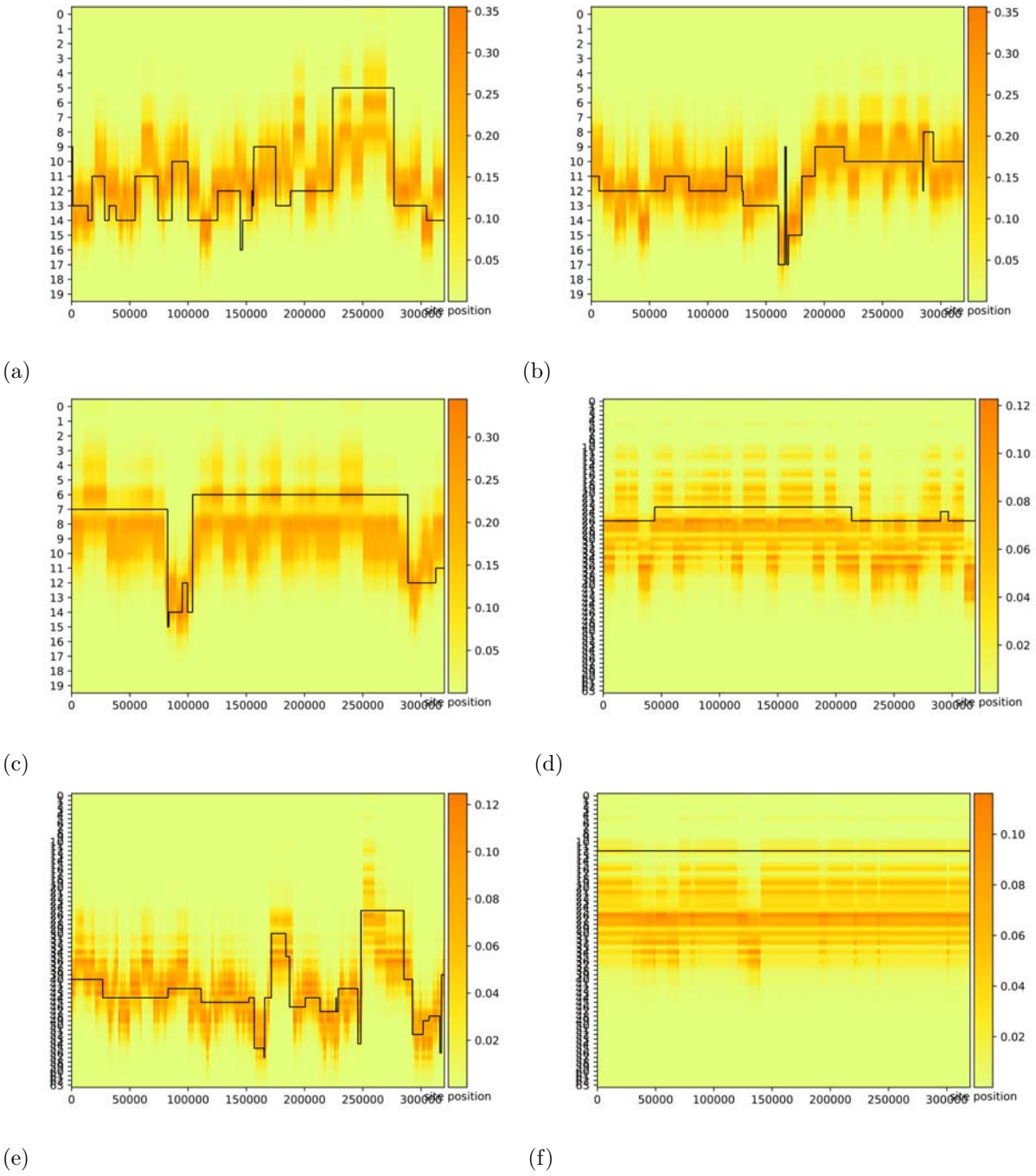

Figure 1: The heatmaps show the distribution of local coalescent times along the genome inferred by deep learning model. The black lines show the true coalescent time known from the simulations.

will consider complicated demographic scenarios with admixture of multiple populations. Such scenarios are much closer to real ones. So, this approach should greatly improve the accuracy of our method compared to HMM-based methods which usually assume a single population model.

5. Aknowledgments

We would like to thank Viktoria Vasileva and Vladislav Tolkach for their help with the project. This research was supported in part through computational resources of HPC facilities at NRU HSE. The work of E. Khomutov and V. Shchur is supported by the Russian Science Foundation under grant 20-71-00143.

References

[1] Heng L and Durbin R 2011 Inference of human population history from individual whole-genome sequences *Nature* **475** 493
[2] Schiffels S and Durbin R 2014 Inferring human population size and separation history from multiple genome sequences, *Nature genetics* **46**
[3] Tataru P,Nirody J A and Song Y S 2014 diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals, *Bioinformatics* **30** 3430–1
[4] McVean G and Cardin N J 2005 Approximating the coalescent with recombination . Trans. R. Soc.
[5] Marjoram P, Wall J D 2006 Fast "coalescent" simulation *BMC Genet* **7** 16
[6] Wang K, Mathieson I, O'Connell J and Schiffels S 2020 Tracking human population structure through time from whole genome sequences. *PLOS Genetics* **16**
[7] Gutenkunst RN, Hernandez RD, Williamson SH and Bustamante CD 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data *PLOS Genetics* **5**
[8] Kamm J, Terhorst J, Durbin R, Song Y S 2020 Efficiently Inferring the Demographic History of Many Populations With Allele Count Data, *Journal of the American Statistical Association* **115:531** 1472-87
[9] Eraslan G, Avsec Ž, Gagneur J et al. 2019 Deep learning: new computational modelling techniques for genomics *Nature genetics* **20** 389-403
[10] Kelleher J, Etheridge A M and McVean G 2016 Efficient Coalescent Simulation and Genealogical Analysis for LargeSample Sizes, PLoS Comput Biol **12**
[11] https://github.com/ANGSD/ngsPSMC
[12] Kalchbrenner K, Grefenstette E and Blunsom Ph 2014 A Convolutional Neural Network for Modelling Sentences *Preprint* arXiv:1404.2188
[13] Chung J,Gulcehre C, Cho K and Bengio Y 2014 Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling *Preprint* arXiv:1412.3555,
[14] Sutskever I, Vinyals O and Quoc L V 2014 Sequence to Sequence Learning with Neural Networks *Advances in Neural Information Processing Systems*