

МЕТОДЫ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ ВРЕМЕНИ ДО ПОСЛЕДНЕГО ОБЩЕГО ПРЕДКА ПО ГЕНОМУ

Е.В. Хомутов¹, К. Арзыматов²

*Национальный исследовательский университет «Высшая школа экономики»,
ул. Мясницкая, 20, 101000, г. Москва, Россия*

¹ehomutov@hse.ru

²karzymatov@hse.ru

Геномные данные являются богатым источником информации об истории популяции. В частности, для активно рекомбинирующих видов время до последнего общего предка (LCA) между двумя хромосомами может быть разным в разных хромосомных локусах. Оценка локального времени LCA важна для многих задач: ее можно использовать для определения генов, подвергающихся отбору или для определения эффективных изменений численности популяции. Современный метод PSMC для определения локального времени LCA и эффективного размера популяции основан на Скрытой Марковской модели. В этой работе мы предлагаем новый метод с использованием глубокого обучения для определения локального времени LCA в масштабе всего генома. Мы демонстрируем, что наш метод точен как по местному времени LCA, так и, как следствие, по распределению локального времени, что, в свою очередь, приводит к корректному вычислению траектории эффективного размера популяции. В будущем наш подход может быть обобщен для сложных демографических сценариев.

Ключевые слова: Глубокое обучение; популяционная геномика; LCA; эффективный размер популяции; демография; геном; хромосома.

DEEP LEARNING METHODS FOR DETERMINING THE TIME TO THE LAST COMMON ANCESTOR BY GENOME

E. V. Khomutov^a, K. Arzymatov^b

*National Research University "Higher School of Economics",
Myasnitskaya str., 20, 101000, Moscow, Russia*

^aehomutov@hse.ru

^bkarzymatov@hse.ru

Genomic data is a rich source of information about the history of the population. In particular, for actively recombining species, the time to the last common ancestor (LCA) between two chromosomes may be different at different chromosomal loci. The estimation of the local LCA time is important for many tasks: it can be used to determine genes undergoing selection, or to determine effective changes in population size. The modern PSMC method for determining the local LCA time and effective population size is based on a Hidden Markov model. In this paper, we propose a new deep learning platform for local

determination of LCA time at the scale of the entire genome. We demonstrate that our method is accurate both in local LCA time and, as a consequence, in LOCAL time distribution, which in turn leads to an effective population trajectory. In the future, our approach can be generalized to complex demographic scenarios.

Keywords: Deep learning; population genomics; LCA; effective population size; demography; genome; chromosome.

Введение

Популяционная генетика - междисциплинарная область исследований, которая опирается на математические и вычислительные методы для определения демографии и структуры популяции во времени на основе геномных данных. Данные секвенирования генома богаты такой информацией, но из-за длины генома и сложной связи между разными участками генома возникают сложности анализа этих данных.

Демографический вывод, который предполагает оценку изменений численности популяции во времени, является одной из ключевых задач популяционной генетики. Например, было показано [1], что все неафриканские популяции прошли через аналогичное узкое место примерно между 30 тыс. лет назад (20 тыс. лет назад) и 100 тыс. лет назад. Африканское население не проявляет признаков такого узкого места. Этот факт подтверждает гипотезу об африканском происхождении современной человеческой популяции.

Современные методы демографического вывода, используемые для млекопитающих и некоторых других животных, основаны на Скрытой Марковской модели (НММ). Такие НММ декодируют генетические варианты вдоль генома при последовательном Марковском слиянии (SMC) [2,3]. SMC — это популяционная модель, которая аппроксимирует слияние с рекомбинацией [4]. При объединении с рекомбинацией генеалогия набора хромосом представляет собой ориентированный ациклический граф с дополнительной структурой данных, отражающей пространственную структуру вдоль этих хромосом. Ограничение этой генеалогии одним генетическим сайтом приводит к древовидной структуре. Эти деревья могут меняться между соседними участками гена из-за рекомбинаций при наследовании. В приближении SMC генеалогические деревья изменяются вдоль генома в соответствии с Марковским процессом. В соответствии с НММ эти генеалогические деревья являются скрытыми состояниями, а генетические варианты - выбросами. Различные методы основаны на этой идее, включая PSMC [1], diCal [5], MSMC [6], SMC++ [7][8], MSMC-im [9], ngsPSMC (неопубликованный). Эти методы различаются структурой входных данных.

Глубокое обучение также используется в популяционной генетике для изучения демографических, популяционных и эволюционных параметров, хотя эти методы широко не используются. Подробный обзор методов глубокого обучения в популяционной генетике представлен в [10], поэтому здесь мы обсудим лишь несколько важных примеров. Мондал и др. [11] использовал глубокое обучение для создания сводной статистики генома для приближенных байесовских вычислений (ABC) [12] для изучения архаичной интрогрессии в Азии и Океании. В [13] глубокое обучение используется для совместного определения демографии и естественного отбора. В [14] для определения горячих точек рекомбинации использовались сменные нейронные сети и моделирование "на лету". Первый метод получения подробной истории популяции с помощью глубокого обучения предложен в [15] и доказывает, что подход, основанный на нейронных сетях, может быть мощным инструментом в популяционной генетике. Хотя на местах необходимо приложить больше усилий для изучения преимуществ и ограничений глубокого обучения, которые будут использоваться для анализа реальных данных.

В этой работе мы предлагаем новый подход с использованием глубокого обучения для прогнозирования времени до общего предка вдоль диплоидного генома аналогично PSMC [1].

1. Методология исследования

В популяционной генетике существуют две важные проблемы для приложений глубокого обучения. Во-первых, это длина генома. Например, геном человека составляет около 3,2 миллиарда нуклеотидов, что на несколько порядков больше, чем любой текст в задачах обработки естественного языка (NLP). Во-вторых, это отсутствие помеченных реальных данных. Эта проблема решается путем моделирования обучающих наборов данных с использованием эффективных тренажеров, доступных в полевых условиях.

Для теоретических изысканий в популяционной геномике мы не можем использовать реальные геномные данные, так как в этом случае мы не знаем истинные таргеты. Вместо реальных данных в популяционной геномике чаще используются смоделированные данные – в данной статье используются данные, полученные с помощью модуля [16]. После обработки выхода генератора имеем две последовательности – бинарный геном (где 0 – означает отсутствие мутации в позиции, 1 - наличие) и дискретное время до общего предка в каждой позиции (мы разбиваем прошлое на временные интервалы с экспоненциально растущей длиной). Такая обработка позволяет нам поставить задачу в терминах машинного

обучения – мы хотим решать задачу классификации с учителем. В качестве архитектуры нейронной сети мы выбрали рекуррентную нейронную сеть с 4 двунаправленной GRU [17]. Двунаправленные сети позволяют объединять информацию, полученную слева и справа от каждой позиции в последовательности, с не разделяемым набором весов для каждого из двух (слева направо и справа налево). Входные данные каждого уровня сети GRU являются выходными данными предыдущего уровня, и он также обрабатывает эти входные данные двунаправленным образом. Полученная модель схематично представлена на рисунке 1.

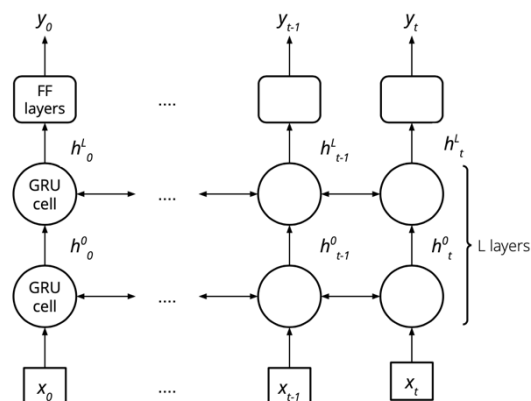


Рисунок 1 – схема архитектуры нейронной сети

2. Результаты и их обсуждение

Мы представляем результаты локального прогнозирования времени LCA модели, обученной по набору данных 1 на "постоянной" хромосоме (рисунок 2 левая панель), и модели, обученной по набору данных 2 на хромосоме "бутылочного горлышка". Тепловая карта представляет вероятности каждого временного класса LCA для каждой геномной позиции. Качественная оценка показывает, что в обоих случаях наша нейронная сеть довольно хорошо фиксирует время LCA.

Далее мы вычислили кумулятивное распределение времени до общего предка (сумма по всем позициям генома) из смоделированных геномов и из прогноза нейронной сети. Мы исследовали предсказание моделей dataset 1 и dataset 2 как для геномов "const", так и для геномов "bottleneck". Результаты представлены на рисунке 3. Модель, обученная на наборе данных 1, хорошо работает с геномом "const", а модель, обученная на наборе данных 2, хорошо работает с геномом "bottleneck". Предсказания модели набора данных 1 для генома "узкого места" и модели набора данных 2 для генома "const" менее точны.

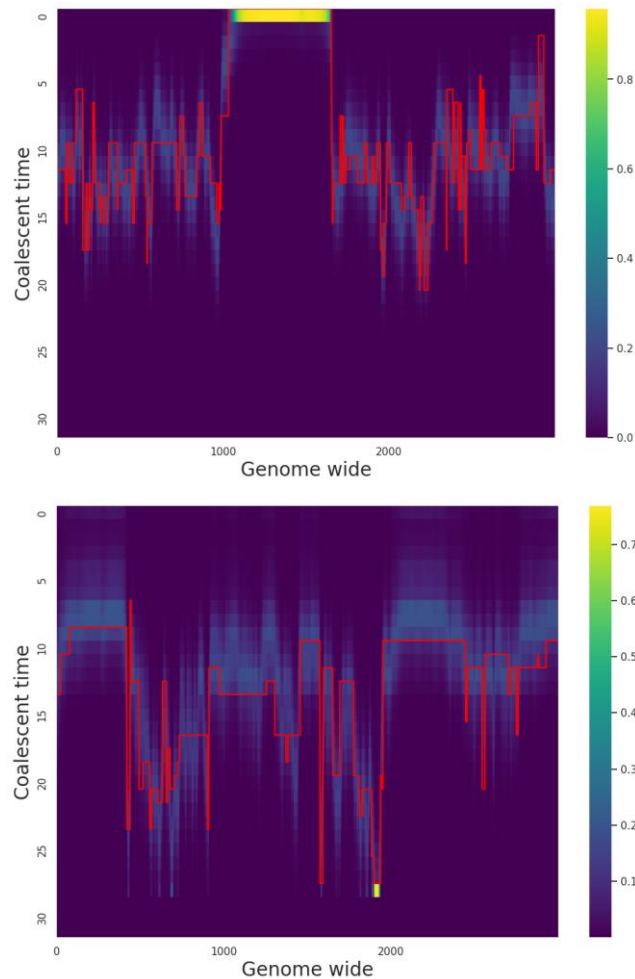


Рисунок 2 - Время до общего предка вдоль хромосомы. Тепловые карты показывают вероятности того, что время до общего предка находится на временном интервале вдоль хромосомы. Левая панель демонстрирует время до общего предка, выведенное из хромосомы, представляющей популяцию постоянного размера, нейронной сетью, обученной на наборе данных постоянного размера. Правая панель представляет время до общего предка вдоль хромосомы из популяции с узким местом, нейронная сеть была обучена на наборе данных со случайными демографическими данными. Красная линия представляет истинное базовое время до общего предка (известное из моделирования).

В то время как в первом случае это неудивительно (демографический сценарий фиксирован, и из него случайным образом извлекаются только хромосомы), во втором случае это менее объяснимо (где демографические сценарии выбираются случайным образом). Эти примеры показывают, что прогноз может быть чувствительным к набору обучающих данных, и необходимы дальнейшие исследования, чтобы понять, как создать оптимальный набор обучающих данных. На самом деле, существует три потенциальных источника сигнала для нейронной сети. Первый — это плот-

ность гетерозигот в геномной области: чем больше время LCA, тем больше гетерозигот присутствует в области. Во-вторых, более длинные области соответствуют более коротким временам LCA. Третий источник информации – это закономерности переходов между различными временами LCA. В терминах Скрытой Марковской модели первая точка определяется матрицей излучения, в то время как вторая и третья определяются матрицей перехода. Более того, они содержат информацию о дополнительной структуре населения, которая станет предметом нашей будущей работы.

Исследование выполнено в рамках Программы фундаментальных исследований НИУ ВШЭ.

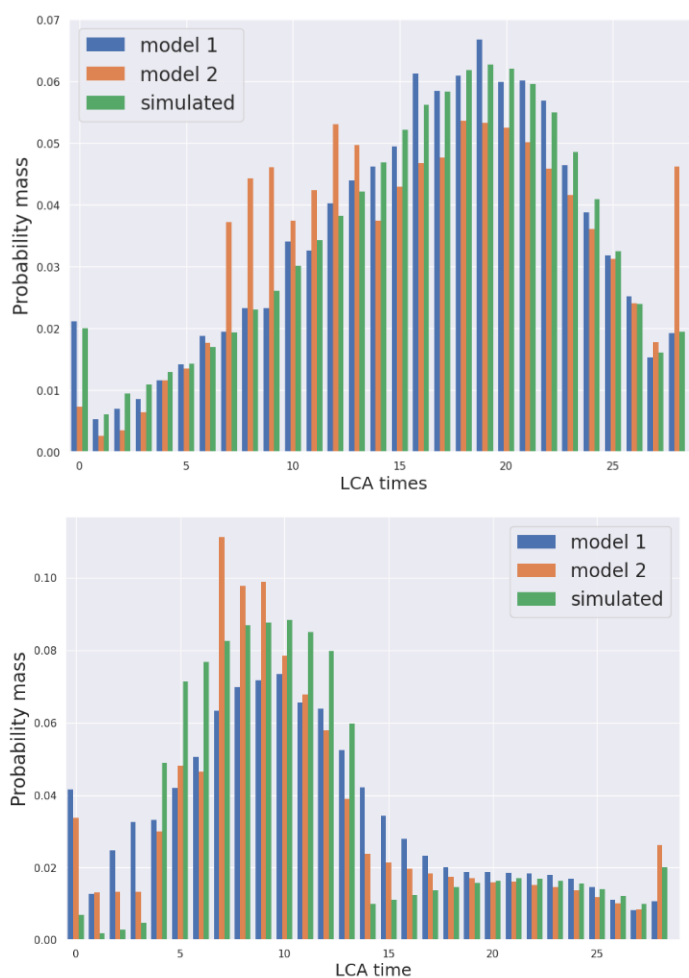


Рисунок 3 – Кумулятивные распределения времен LCA, рассчитанные на основе смоделированного генома (“смоделированный”), оцененные моделью, обученной на наборе данных 1 (“модель 1”), и моделью, обученной на наборе данных 2 (“модель 2”)

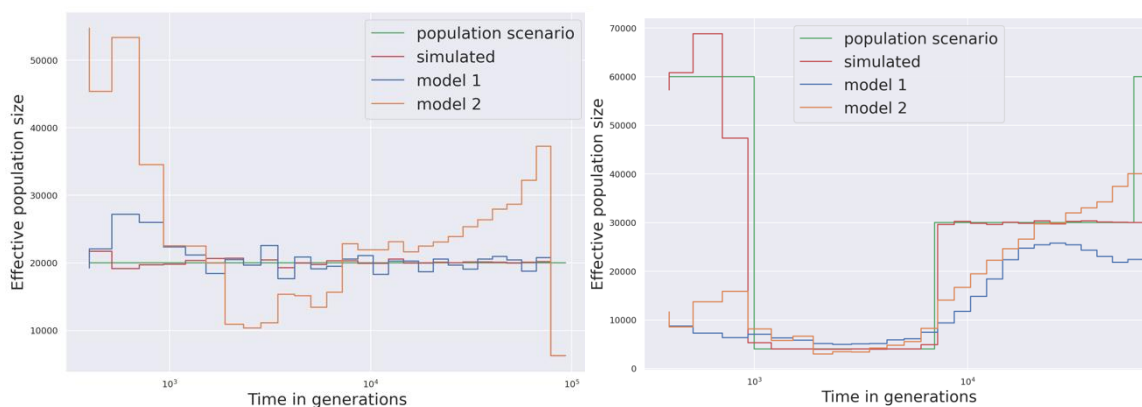


Рисунок 4 – Эффективная численность популяции, оцененная как обратная скорости слияния, рассчитанной на основе смоделированного генома (“смоделированный”), оцененная моделью, обученной на наборе данных 1 (“модель 1”), и моделью, обученной на наборе данных 2 (“модель 2”). Также показана эффективная численность населения в моделируемом демографическом сценарии (“демографический сценарий”).

Библиографические ссылки

1. Li H. and Durbin R. "Inference of human population history from individual whole-genome sequences," *Nature* **475**, 2011. P. 493–496.
2. G. A. T. McVean and N. J. Cardin, "Approximating the coalescent with recombination," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 2005. P. 1387–1393.
3. P. Marjoram and J. D. Wall, "Fast "coalescent" simulation," *BMC Genetics*. 2006.
4. R. R. Hudson, "Gene genealogies and the coalescent process," *Oxford surveys in evolutionary*, 1990. P. 1–44.
5. S. Sheehan, K. Harris, and Y. S. Song, "Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach," *Genetics*, 2013. P. 647–662.
6. S. Schiffels and R. Durbin, "Inferring human population size and separation history from multiple genome sequences," *Nature Genetics* **46**, 2014. P. 919–925.
7. J. Terhorst, J. A. Kamm, and Y. S. Song, "Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*," *Nature Genetics* **49**, 2017. P. 303–309.
8. P. F. Palamara, J. Terhorst, Y. S. Song, and A. L. Price, "High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability," *Nature Genetics* **50**, 2018. P. 1311–1317.
9. K. Wang, I. Mathieson, J. O’Connell, and S. Schiffels, "Tracking human population structure through time from whole genome sequences," *PLOS Genetics* **16**, 2020. P. 1–24.
10. G. Eraslan, Z. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nat. Rev. Genet.* **20**, 2019. P. 389–403.
11. M. Mondal, J. Bertranpetit, and O. Lao, "Approximate bayesian computation with deep learning supports a third archaic introgression in asia and oceania," *Nature Communications* **10**, 2019.
12. S. Tavar'e, D. J. Balding, R. C. Griffiths, and P. Donnelly, "Coalescence Times From DNA Sequence Data," *Genetics* **145**, 1997. P. 505–518.

13. S. Sheehan and Y. S. Song, "Deep learning for population genetic inference," *PLOS Computational Biology* {\bf 12}, 2016. P. 1–28.
14. J. Chan, V. Perrone, J. P. Spence, P. A. Jenkins, S. Mathieson, and Y. S. Song, "A likelihood-free inference framework for population genetic data using exchangeable neural networks," *Advances in neural information processing systems* {\bf 31}, 2018. P. 8594–8605.
15. T. Sanchez, J. Cury, G. Charpiat, and F. Jay, "Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation," *Molecular Ecology Resources* {\bf 21}, 2021. P. 2645–2660.
16. Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, Ariella L Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W Kretzschmar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S Pope, Consuelo D Quinto-Cortés, Murillo F Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W Wohns, Yan Wong, Simon Gravel, Andrew D Kern, Jere Koskela, Peter L Ralph and Jerome Kelleher (2022), Efficient ancestry and mutation simulation with *[16] 1.0*, *Genetics*, Vol. 220, Issue 3. doi.org/10.1371/journal.pcbi.1004842.
17. K. Cho B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv*. 2014. 1409.1259.