Национальный исследовательский университет ИТМО
(Университет ИТМО)

На правах рукописи

Носкова Екатерина Эдуардовна

Методы построения моделей демографических историй

Специальность 1.2.2 —
«Математическое моделирование, численные методы и комплексы программ
(технические науки)»

Диссертация на соискание учёной степени
кандидата технических наук

Научный руководитель:
канд. техн. наук
Ульянцев Владимир Игоревич

Санкт-Петербург — 2023

ITMO University

As a manuscript

Noskova Ekaterina Eduardovna

Methods for inferring demographic history models

Specialty 1.2.2 —
Mathematical modeling, numerical methods and software packages (Engineering)

A thesis submitted in fulfillment of the requirements for the degree of
PhD in Engineering

Scientific advisor:
Doctor of Philosophy
Ulyantsev Vladimir Igorevich

Saint Petersburg — 2023

Содержание

Реферат 9

Synopsis 31

Введение 51

Глава 1. Обзор предметной области 59
1.1. Демографическая история популяций 60
1.2. Методы вывода демографической истории популяций по генетическим данным 66
1.3. Методы моделирования демографической истории популяций 75
1.3.1. Модели первого класса 75
1.3.2. Модели второго класса 81
1.3.3. Методы сравнения моделей с разным числом параметров 84
1.4. Методы и программные комплексы для вычисления правдоподобия генетических данных при условии заданной демографической истории 86
1.4.1. Основные понятия биологии и генетики 86
1.4.2. Используемые статистики генетических данных 89
1.4.3. Математические модели эволюции, методы дифференциального исчисления, численные методы и программные комплексы для вычисления правдоподобия . 94
1.5. Методы оптимизации для настройки параметров модели демографической истории популяций по генетическим данным . 103
1.6. Методы перебора моделей демографической истории 109
Выводы по главе 1 112

Глава 2. Расширенный класс моделей демографической истории популяций и методы настройки параметров моделей по генетическим данным 114
2.1. Расширенный класс моделей демографической истории популяций 114
2.2. Метод на основе комбинации генетического алгоритма и локального поиска для настройки параметров моделей демографической истории популяций по генетическим данным 120
2.2.1. Разработка метода на основе комбинации генетического алгоритма и локального поиска 121
2.2.2. Реализация разработанного метода, основанного на комбинации генетического алгоритма и локального поиска 130

2.2.3. Настройка гиперпараметров разработанного генетического алгоритма 135
2.3. Метод на основе комбинации байесовской оптимизации и локального поиска для настройки параметров модели демографической истории популяций по генетическим данным 142
2.3.1. Разработка метода на основе комбинации байесовской оптимизации и локального поиска 143
2.3.2. Реализация разработанного метода, основанного на комбинации байесовской оптимизации и локального поиска 149
2.3.3. Настройка гиперпараметров байесовской оптимизации и разработка ансамблевого метода 154
2.4. Экспериментальные исследования разработанного метода настройки параметров моделей, основанного на комбинации генетического алгоритма и локального поиска для данных одной, двух и трех популяций 160
2.4.1. Сравнение с существующими методами настройки параметров на симулированных данных одной, двух и трех популяций 161
2.4.2. Сравнение с существующими методами настройки параметров моделей на данных популяций кошачьей лягушки 170
2.4.3. Сравнение с существующими методами настройки параметров моделей на данных двух популяций американской пумы 172
2.4.4. Сравнение с существующими методами настройки параметров моделей на данных одной популяции огородной капусты 175
2.4.5. Сравнение методов вычисления правдоподобия на симулированных данных двух популяций орангутанга . . 179
2.4.6. Вывод демографической истории трех популяций современного человека 184
2.5. Экспериментальные исследования разработанного метода настройки параметров моделей, основанного на комбинации байесовской оптимизации и локального поиска для данных четырех и пяти популяций 187
2.5.1. Сравнение с разработанным генетическим алгоритмом на симулированных и реальных данных 188
2.5.2. Сравнение с существующим методом настройки параметров моделей на реальных данных четырех и пяти популяций современного человека 191
Выводы по главе 2 195

Глава 3. Метод автоматического перебора расширенных моделей с разным числом параметров и настройки параметров по генетическим данным одной, двух и трех популяций . . . 196
3.1. Метод автоматического перебора моделей расширенного класса 196
3.1.1. Разработка метода автоматического перебора моделей расширенного класса . . . 196
3.1.2. Реализация разработанного метода автоматического перебора моделей расширенного класса . . . 202
3.2. Экспериментальные исследования разработанного метода автоматического перебора моделей расширенного класса . . . 204
3.2.1. Вывод демографической истории трех популяций современного человека . . . 204
3.2.2. Вывод демографической истории популяций кошачьей лягушки . . . 206
3.2.3. Вывод демографической истории двух и трех популяций голубой акулы . . . 208
Выводы по главе 3 . . . 213

Глава 4. Программный комплекс GADMA для вывода демографической истории популяций по генетическим данным и расширение библиотек *stdpopsim* и *demes* . . . 215
4.1. Программный комплекс GADMA для вывода демографической истории популяций по генетическим данным . . . 215
4.1.1. Структура программного комплекса GADMA . . . 216
4.1.2. Входные данные и интерфейс запуска . . . 219
4.1.3. Выходные данные . . . 222
4.1.4. Разработка и сопровождение программного комплекса 224
4.2. Расширение библиотек *stdpopsim* и *demes* для проведения экспериментальных исследований и представления результатов . 226
4.2.1. Расширение библиотеки *stdpopsim* для симулирования генетических данных . . . 226
4.2.2. Расширение библиотеки *demes* для текстового и визуального представления демографических историй . . . 228
Выводы по главе 4 . . . 229

Заключение . . . 230

Список литературы . . . 232

Список иллюстраций . . . 245

Список таблиц . . . 253

Приложение А. Благодарности . . . 255

Приложение Б. Награды автора, полученные во время работы над диссертацией 256

Реферат

Общая характеристика работы

Актуальность темы исследования. Модели метрических деревьев с функциями на ребрах применяются для анализа и прогнозирования различных явлений реального мира, например, процессов, представимых в виде динамических систем с переменной структурой [9, 10]. Под *метрическими деревьями* (metric trees) понимают граф, являющийся деревом, где каждому ребру поставлен в соответствие интервал. В общем виде, метрические графы с функциями на ребрах нашли широкое применение, например, в виде квантовых графов [11], которые используются в физике при изучении квантового хаоса [12], волноводов [13] и фотонных кристаллов [14].

Построение модели представляет собой набор действий, направленных на выбор конфигурации, определение параметров модели и настройку их значений с целью достижения высокого соответствия результатов моделирования данным натурного эксперимента. На различных этапах построения модели часто требуются экспертные данные или предположения об исследуемом объекте. Эти данные могут быть неточными, ограниченными или неизвестными, что может негативно сказаться на точности и адекватности модели. Методы автоматизированного построения позволяют уменьшить вероятность человеческих ошибок при выборе модели и настройке ее параметров.

При работе с моделями метрических деревьев с функциями на ребрах прибегают к участию предметных специалистов. В первую очередь экспертные данные используются для определения свойств функций на ребрах дерева. Эта информация позволяет установить *конфигурацию* модели, где каждая функция на дереве принадлежит заданному семейству и характеризуется функциональными параметрами, доступными для настройки. В условиях отсутствия экспертных данных или для минимизации влияния специалиста на получение результата приходится рассматривать множество всех возможных моделей, отличающихся типами функций и функциональными параметрами. Например, при построении моделей демографических историй для каждой популяции в качестве динамики изменения численности обычно рассматриваются кусочно-заданные функции, состоящие из функций трех наиболее популярных типов: константная, линейная и экспоненциальная. Такой перебор конфигураций приводит к увеличению временных затрат при построении модели, тем больших, чем больше допустимых типов функций. Дополнительно, требуется следить за сложностью модели, числом ее параметров и переобучением.

Методы для настройки параметров моделей также могут быть ограничены в степени автоматизации и требовать экспертных данных. Например, при использовании методов локального поиска требуются вовлечение специалиста для определения начальных значений параметров, и эффективность настройки зависит от этого выбора.

Таким образом, при моделировании явлений реального мира в виде метрического дерева с функциями на ребрах *актуальна* разработка специализированных моделей и методов для автоматического построения и настройки моделей с целью минимизации влияния экспертных данных на результат моделирования, что рассматривается в данной диссертации на примере задачи вывода демографических историй по генетическим данным.

Популяция — это группа особей одного вида, живущих на одной территории. *Демографическая история популяций* — это исторический процесс их развития и эволюции, который включает в себя такие явления, как изменения численности популяций, разделения популяций, миграция и отбор. Демографические истории используются для датирования исторических событий, не оставивших письменных свидетельств [15, 16], а также играют важную роль в области консервативной генетики [5] и даже в медицине [17].

Различные статистические и алгоритмические методы позволяют строить модели демографических историй в виде метрических деревьев с функциями на ребрах и настраивать их непрерывные параметры по генетическим данным. В случае демографических историй, метрическое дерево является деревом, которое определяет разделение популяций, а функции на ребрах — динамиками изменения численности популяций. В качестве динамик рассматривают кусочно-заданные функции, состоящие из функций трех наиболее популярных типов: константная, линейная и экспоненциальная. При построении моделей требуется определить число временных интервалов, а также тип динамики для каждой кусочно-заданной функции.

Вовлечение специалиста также происходит и на этапе настройки параметров моделей демографической истории популяций, для чего используются комбинация методов численного моделирования и методов оптимизации. Методы численного моделирования используются для вычисления функции правдоподобия, которая позволяет оценить степень соответствия модели генетическим данным. Для поиска параметров, обеспечивающих максимальное значение правдоподобия, используются методы локальной оптимизации. Именно эти методы ограничены в степени автоматизации: они требуют экспертных данных для определения начальных значений параметров, а их эффективность зависит от этого выбора.

Задача вывода демографической истории популяций дополнительно усложняется необходимостью реализации пользователем программного кода модели и алгоритма вывода ее параметров. Методы численного моделирования, используемые существующими решениями, имеют разные возможности и стабильность работы, и пользователь может применить несколько из них для сравнения результатов. Однако при применении различных программных решений одновременно, пользователь сталкивается с необходимостью задавать одни и те же модели с использованием разных интерфейсов.

Таким образом, развитие методов автоматического построения и настройки метрических деревьев с функциями на ребрах приведет к минимизации

влияния экспертных данных, и, следовательно, к повышению качества моделирования явлений реального мира по данным натурного эксперимента.

Степень разработки проблемы. Модели графов исследуются и применяются для решения широкого круга задач. В работах А.М. Райгородского [18, 19] приведены описания и примеры применения моделей случайных графов. Графовые вероятностные модели, такие как байесовские сети, обширно представлены в работах И. Бена-Гала [20] для моделирования индустриальных систем [21], классификации [22] или идентификации сайтов связывания транскрипционных факторов [23]. Л. Кларк и Д. Прегибон [24] описали примеры применения моделей, основанных на деревьях, к которым относятся, например, решающие деревья [25].

Теория метрических графов была сформирована работами В.Г. Болтянского [26], П.С. Солтана [26, 27] и А. Дресса [28]. Свойства метрических деревьев и метрических пространств, порожденных ими, были изучены А. Дрессом [28], Б. Бунеманом [29] и Д. Олдосом [10, 30, 31]. В работах А.С. Матвеева и С.И. Матвеева [32–34] метрические графы были применены при построении координационных моделей для интеллектуальной навигации.

Разработкой моделей, приближающих неявные функции, также активно занимаются многие ученые. Наиболее широкое применение, описанное в работах Л. Фармейра [35] и Р. Сни [36], эти модели получили для решения задач регрессии. При использовании моделей кусочно-заданных функций обычно фиксируют общий вид формирующих функций, например, строят кусочно-постоянные [37, 38], кусочно-линейные [39] или кусочно-экспоненциальные [40] модели. Число точек смены функции, а также их положение являются неизвестными характеристиками моделей кусочно-заданных функций. В работах [41, 42] рассмотрены методы автоматического построения таких моделей для решения задачи кусочно-заданной регрессии, где число точек смены функции определяется с использованием байесовского информационного критерия (BIC) и информационного критерия Акаике (AIC) [43] соответственно.

Модели метрических деревьев с функциями на графах являются комбинацией моделей метрических деревьев и функциональных моделей на ребрах. Квантовые графы, которые являются метрическими графами с дифференциальными операторами на ребрах, и их приложения подробно рассмотрены в работах Г. Берколайко [11, 44]. Метрические деревья с функциями на ребрах используются для моделирования демографических историй популяций в работах Р. Гутенкунста [45], Д. Камма [46], А. Рэгсдейла и С. Гравеля [47, 48]. Однако методы, представленные в этих работах, предполагают, что пользователь определяет и фиксирует общий вид кусочно-заданной функции на ребрах дерева, а также задает начальные значения параметров настройки параметров методами локальной оптимизации. В работах Д. Портика [49, 50] и Р. Гутенкунста [51] были представлены методы глобальной оптимизации для настройки параметров моделей демографических историй, которые минимизируют, однако все еще требуют во-

влечение пользователя. Общее применение методов численной оптимизации для решения задач представлено в классической работе Б.Т. Поляка [52], а описание современных методов глобальной оптимизации в работе [53].

На момент начала исследований автором (в 2017 году) не существовало метода автоматического построения и настройки моделей метрических деревьев с функциями на ребрах. К концу диссертационного исследования появилось первое альтернативное решение для метода автоматического перебора моделей на примере задачи вывода демографических историй [54]. Однако метод позволяет анализировать модели, определенные специфичным каталогом и только для вывода демографической истории *двух* популяций, а выбор наилучшей модели происходит в предположении независимости данных, что не всегда является корректным.

Целью настоящей диссертации является повышение качества[1] компьютерного моделирования явлений реального мира за счет автоматизации построения и настройки моделей метрических деревьев с функциями на ребрах.

Для решения цели в диссертации решаются следующие **задачи**:
- исследование текущего состояния предметной области, уточнение задачи и способов оценки результатов;
- формализация постановки задачи построения и настройки моделей метрического дерева с функциями на ребрах;
- разработка метода автоматической настройки моделей метрического дерева с функциями на ребрах на основе комбинации методов глобальной и локальной оптимизации;
- разработка метода автоматического перебора моделей метрического дерева с кусочно-заданными функциями на ребрах;
- проектирование и реализация программного комплекса, включающего разработанные модели и методы для вывода демографической истории популяций по генетическим данным;
- проведение экспериментальных исследований, подтверждающих эффективность разработанных моделей и методов, а также их применимость для вывода демографической истории популяций по генетическим данным, анализ результатов экспериментов.

Научная новизна диссертации состоит в том, что: (1) разработаны методы на основе комбинации методов глобальной и локальной оптимизации для настройки параметров заданной модели метрического дерева с функциями на ребрах; (2) разработан метод автоматического перебора моделей метрического

[1] Качество моделей в данной работе определяется степенью соответствия настроенной модели данным натурного эксперимента. В случае задачи вывода демографических историй популяций качество определяется значением функции правдоподобия, полученным численными методами за фиксированное время настройки модели.

дерева с кусочно-заданными функциями на ребрах, не требующий вовлечения эксперта на этапе выбора параметров рассматриваемых моделей.

Теоретическая значимость работы определяется расширением классической постановки задачи настройки модели метрического дерева с функциями на ребрах не только как задачи настройки параметров заданной модели, но и как задачи выбора самой модели путем автоматического перебора. Полученные методы моделирования и настройки применимы для произвольных моделей метрического дерева с функциями на ребрах. Более того, разработанные методы оптимизации могут быть использованы или адаптированы для задач поиска оптимальных параметров в других научных областях.

Практическую значимость работы определяют:

- расширение научно-практического инструментария специалистов-биоинформатиков методами и алгоритмами для вывода демографических историй популяций;
- открытый программный код разработанного программного комплекса GADMA, который доступен к переиспользованию по адресу `https://github.com/ctlab/GADMA`;
- применимость разработанных методов для анализа генетических данных;
- внедрение разработанного метода на основе генетического алгоритма в стороннее программное решение [54].

На защиту выносятся положения, обладающие научной новизной:

1. Метод моделирования и настройки параметров моделей метрических деревьев с функциями на ребрах по данным натурного эксперимента, содержащий модели с непрерывными функциональными параметрами, отличающийся тем, что с целью автоматической настройки без привлечения экспертных данных в нем используются модели с дискретными парамерами, определяющими семейства функций, а также методы глобальной оптимизации — генетический алгоритм и байесовская оптимизация, и реализующий его комплекс программ.
2. Метод автоматического перебора моделей метрических деревьев с функциями на ребрах с разным числом параметров и настройки этих параметров по данным натурного эксперимента, содержащий сравнение моделей с использованием информационного критерия Акаике, отличающийся тем, что с целью повышения уровня автоматизации и обеспечения возможности настраивать не только параметры модели, но и саму модель, он включает метод увеличения числа временных интервалов для кусочно-заданных функций на ребрах дерева, а также реализующий его комплекс программ.

Методы исследования. В работе использованы методы оптимизации, численные методы, методы теории вероятности и математической статистики, методы машинного обучения и методы проведения экспериментальных исследований.

Достоверность научных результатов обусловлена корректным использованием методов, обоснованием постановки задач, экспериментальными исследованиями, покрывающими разработанные технологии и алгоритмы. Демографические истории, полученные разработанными методами на проверяемых симулированных данных, согласуются с исходными историями, используемыми для моделирования. Результаты, полученные на реальных данных, согласуются с опубликованными ранее исследованиями [45, 55–59].

Соответствие паспорту специальности. Полученные научные результаты соответствуют следующим пунктам паспорта специальности 1.2.2 — «Математическое моделирование, численные методы и комплексы программ (технические науки)».

Пункт 2 паспорта специальности «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий». Были разработаны, обоснованы и протестированы методы настройки параметров моделей метрического дерева с функциями на ребрах, основанные на методах численной оптимизации.

Пункт 4 паспорта специальности «Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели». В диссертационном исследовании представлены методы для построения моделей метрического дерева с функциями на ребрах по данным натурного эксперимента с целью анализа явлений реального мира.

Апробация результатов работы

Основные результаты работы были представлены на следующих конференциях:

- Международный конгресс «VII съезд Вавиловского общества генетиков и селекционеров, посвященный 100-летию кафедры генетики СПбГУ, и ассоциированные симпозиумы», 2019, Санкт-Петербург, Россия;
- Moscow Conference on Computational Molecular Biology, 2019, Москва, Россия;
- Probabilistic Modeling in Genomics, 2019, Осуа, Франция;
- Probabilistic Modeling in Genomics, 2021, онлайн;
- Moscow Conference on Computational Molecular Biology, 2021, Москва, Россия;
- Вероятностные методы в анализе: пространства голоморфных функций, 2021, Сочи, Россия;

– LI Научная и учебно-методическая конференция Университета ИТМО, 2022, Университет ИТМО, Санкт-Петербург, Россия;
– Probabilistic Modeling in Genomics, 2022, Окфорд, Великобритация;
– XI Конгресс молодых ученых, 2022, Университет ИТМО, Санкт-Петербург, Россия;
– Conservation Genomics at the Population Level, 2022, Кембридж, Великобритания;
– Probabilistic Modeling in Genomics, 2023, Колд Спринг Харбор, США;
– XII Конгресс молодых ученых, 2023, Университет ИТМО, Санкт-Петербург, Россия;
– Society for Molecular Biology and Evolution Meeting (SMBE23), 2023, Феррара, Италия.

Награды

– Бронзовая награда в номинации 17th Human-Competitive Awards на онлайн конференции The Genetic and Evolutionary Computation Conference (GECCO) в 2020 году.
– Победитель конкурсной программы поддержки исследовательских проектов System Biology Fellowship от Сколковского института науки и технологий по проекту «Computational methods for unsupervised demographic inference of multiple populations from genomic data» в 2021 году. Число победителей — пять на всю страну в год.

Публикации

По результатам, представленным в диссертации, было опубликовано восемь статей в рецензируемых научных журналах, входящих в международные реферативные базы данных и системы цитирования Scopus и Web of Science.

Личный вклад автора

1. В публикации [1] Noskova E. — разработка и реализация генетического алгоритма и метода автоматического перебора моделей демографической истории, проведение экспериментальных исследований (80%); Ulyantsev V. — рекомендации по постановке задачи, выбору и обоснованию теоретических основ научного исследования (10%); Koepfli K.P., O'Brien S.J. — консультирование при проведении экспериментальных исследований и написании статей (5%); Dobrynin P. — рекомендации по постановке задачи (5%).
2. В публикации [2] Noskova E. — разработка и реализация методов, программного обеспечения для вывода демографической истории популяций по генетическим данным, проведение экспериментальных исследований (85%); Abramov N., Iliutkin S., Sidorin A. — разработка программного обеспечения (10%); Dobrynin P., Ulyantsev V. — рекомендации по

постановке задач, выбору и обоснованию теоретических основ научного исследования (5%).

3. В публикации [3] Noskova E. — разработка и реализация метода байесовской оптимизации для вывода демографической истории популяций по генетическим данным, проведение экспериментальных исследований (90%); Borovitskiy V. — рекомендации по постановке задач, выбору и обоснованию теоретических основ научного исследования (10%).
4. В публикации [4] Noskova E. — вывод демографической истории трех популяций современного человека (10%); Ulyantsev V. — рекомендации по постановке задачи (5%); остальные соавторы — сбор и анализ генетических данных (85%).
5. В публикации [5] Noskova E. — вывод демографической истории двух и трех популяций голубых акул (10%); остальные соавторы — сбор и анализ генетических данных (90%).
6. В публикации [6] Noskova E. — разработка и тестирование программного обеспечения для симулирования генетических данных по демографической истории популяций (5%); остальные соавторы — разработка и тестирование программного обеспечения, проведение экспериментальных исследований (95%).
7. В публикации [7] Noskova E. — реализация демографических историй популяций в программном обеспечении для симулирования генетических данных по демографической истории популяций (5%); остальные соавторы — разработка программного обеспечения (95%).
8. В публикации [8] Noskova E. — разработка программного обеспечения для представления демографической истории популяций (5%); остальные соавторы — разработка программного обеспечения (95%).

Структура диссертационной работы

Диссертация состоит из введения, четырех глав, заключения и приложения. Полный объём диссертации составляет 396 страниц, включая 120 рисунков, 16 таблиц и восемь листингов. Список литературы содержит 171 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность научных исследований, проводимых в рамках данной диссертационной работы, описана степень разработки проблемы вывода демографической истории популяций по генетическим данным, обзор методов моделирования демографических историй, сформулированы цели и задачи, описаны научная новизна, теоретическая и практическая значимости работы, а также перечислены положения, выносимые на защиту.

В **первой главе** приводится обзор предметной области, который включает определение демографической истории популяций, описание существующих методов вывода демографических историй популяций по генетическим данным.

В **разделе 1.1** описаны основные определения популяционной генетики, используемые в данной работе. Она включает формальное определение демографической истории популяций. Популяционная генетика является важной областью генетики, изучающей изменение генетического состава популяций и их эволюцию. Она решает такие задачи, как определение структуры популяций, построение филогенетических деревьев и поиск демографической истории популяций.

Демографическая история популяций — история эволюции и развития популяций, которая включает в себя информацию о том, как популяции делились и образовывались, какова была численность популяций, интенсивность миграции, *коэффициенты инбридинга* — степень близкородственных связей, и много другого. Примеры визуального представления демографических историй представлены на рисунке Р.1. Информация о численности популяций и миграциях отображена шириной закрашенных областей и стрелками между ними. Время в демографических историях зачастую измеряется в поколениях или годах и отображено по оси ординат.

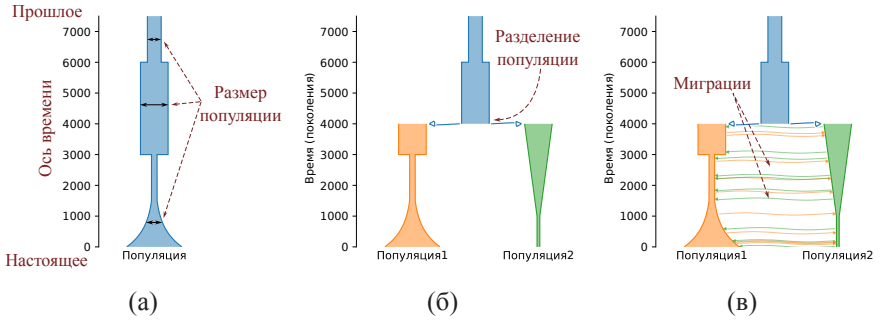


Рисунок Р.1 – Примеры визуального представления демографических историй одной и двух популяций

В **разделе 1.2** описана постановка задачи вывода демографической истории популяций по генетическим данным с использованием параметрических моделей, а также описаны основные компоненты существующих методов решения этой задачи. В разделе приведено краткое описание известных программных средств, реализующих эти методы, а именно *дадi*, *moments*, *momi2* и *momentsLD*.

Для вывода демографической истории популяций используются параметрические модели, которые представляют собой *метрические деревья с функциями на ребрах*. Использование моделей позволяет, во-первых, ограничить пространство поиска, а, во-вторых, использовать методы оптимизации для настройки значений их параметров по генетическим данным. На рисунке Р.2 изображен пример модели в виде метрического дерева с функциями на ребрах, которое описывает демографическую историю двух популяций.

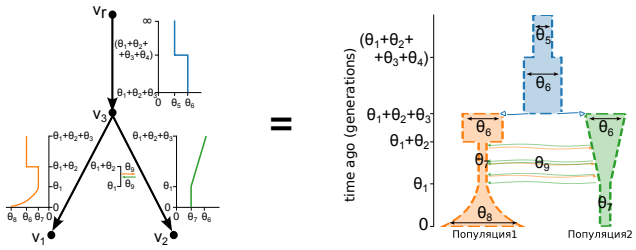


Рисунок Р.2 – Пример модели демографической истории двух популяции в виде метрического дерева с функциями на ребрах

Задача вывода демографической истории популяций по генетическим данным заключается в *настройке параметров* заданной модели — поиске параметров, обеспечивающих максимальное значение функции правдоподобия генетических данных (рисунок Р.3). Существующие программные решения отличаются интерфейсами спецификации моделей, методами вычисления правдоподобия и методами оптимизации для настройки параметров.

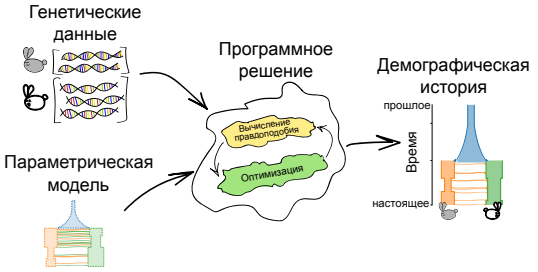


Рисунок Р.3 – Пример входа и выхода существующих программных решений для вывода демографической истории популяций по генетическим данным

В **разделе 1.3** описаны два класса моделей демографических историй, которые применяются в существующих решениях, а также методы сравнения моделей с разным числом параметров.

Модели первого класса используются в программных решениях *дадi*, *moments* и *momentsLD*. Они представляются в виде последовательности элементов временных интервалов, разделений, единичных миграций и элементов инбридинга (рисунок Р.4). Они имеют только непрерывные параметры, а динамики изменения численности (константная численность, линейное или экспоненциальное изменение) в этих моделях всегда фиксированы.

Модели второго класса применяются в программном решении *momi2*. Они представляются в виде набора событий изменения численности, разделения популяций и единичных миграций. Модели второго класса также включают только непрерывные параметры и имеют фиксированные динамики изменения численности. Однако они являются более ограниченными по сравнению с моделями первого класса, например, не поддерживают линейное изменение численности или непрерывные миграции.

Проблема выбора модели в общем случае состоит в том, что необходимо выбрать наиболее подходящую модель для данных. Если выбрана слишком простая модель — с малым числом параметров, она может не отображать всю информацию из данных. Если выбрана слишком сложная модель — с большим числом параметров, она может переобучиться на шуме в данных и в итоге неправильно моделировать реальный процесс. Для сравнения различных моделей и выбора наилучшей используют информационный критерий Акаике (AIC) [43], байесовский информационный критерий (BIC) [60] и тест отношения правдоподобия [61].

```python
import dadi

def model(params, ns, theta0, pts):
    Nanc, N1F, N2B, N2F, Tp, T = params

    # Задание сетки для численных вычислений
    xx = yy = dadi.Numerics.default_grid(pts)

    # Инициализация модели начальным размером популяции
    phi = dadi.PhiManip.phi_1D(xx, nu=Nanc, theta0=theta0)

    # Первый временной интервал
    # Функция изменения численности - константа N1F
    phi = dadi.Integration.one_pop(phi, xx, T=Tp, nu=N1F, theta0=theta0)

    # Второй элемент модели - разделение популяции
    phi = dadi.PhiManip.phi_1D_to_2D(xx, phi)

    # Функция изменения численности первой популяции - константа N1F
    # Задание функции изменения численности второй популяции
    n2_func = lambda t: N2B * (N2F / N2B) ** (t / T)
    # Третий элемент - второй временной интервал
    phi = dadi.Integration.two_pops(phi, xx, T=T, nu1=N1F, nu2=n2_func,
                                    theta0=theta0)

    # Вычисляем численными методами ожидаемую статистику данных
    sfs = dadi.Spectrum.from_phi(phi, ns, (xx,yy))
    return sfs
```

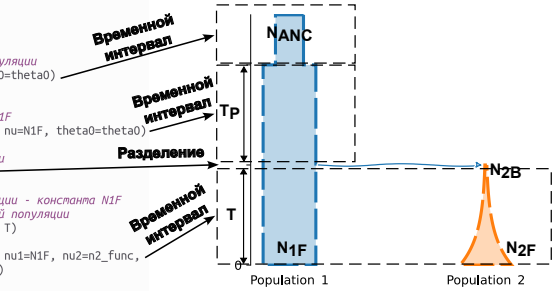
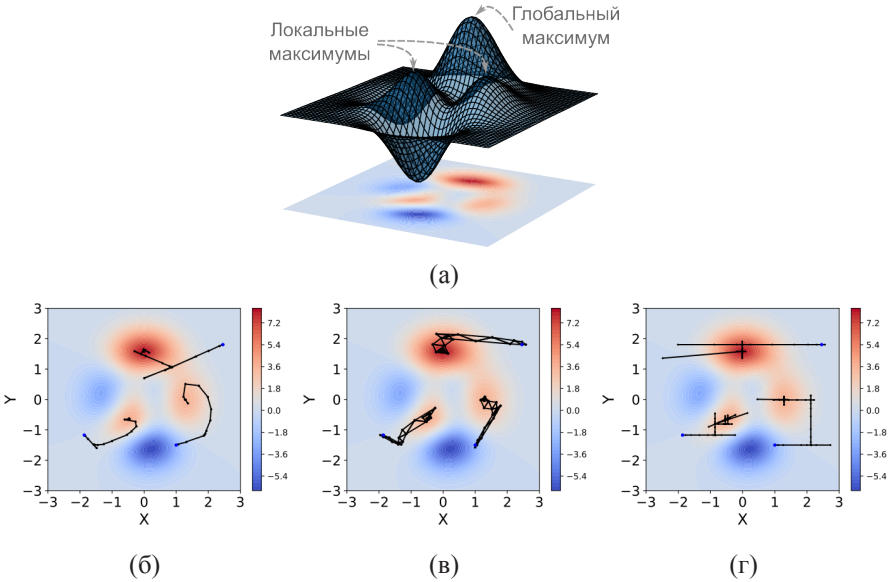


Рисунок Р.4 – Пример задания модели первого класса с использованием интерфейса библиотеки *дадi*

В **разделе 1.4** приведено описание существующих методов вычисления правдоподобия генетических данных при условии заданной демографической истории. В раздел включены определения основных понятий биологии и генетики, например, ДНК, генов, аллелей и генотипов. Также описаны основные используемые статистики данных: аллель-частотный спектр и статистики на основе неравномерного сцепления генов. Наконец, приведено описание методов вычисления правдоподобия, реализованных в программных решениях *дадi*, *moments*, *momi2* и *momentsLD*.

Раздел 1.5 содержит общее описание методов локальной и глобальной оптимизации, основные отличия этих двух групп, а также обзор существующих методов оптимизации для настройки параметров моделей демографических историй по генетическим данным. Преимущественно применяются методы локальной оптимизации такие, как метод Бройдена-Флетчера-Гольдфарба-Шанно (BFGS) [62–65], метод Нелдера-Мида [66] и метод Пауэлла [67] (рисунок Р.5). Существующие методы оптимизации для настройки параметров моделей ограничены выводом значений только непрерывных параметров и требуют вовлечения пользователя для задания начальных значений параметров модели и гиперпараметров метода, например, числа перезапусков. Использование методов локальной оптимизации не гарантирует нахождение глобального оптимума.



(а) (б) (в) (г)

Рисунок Р.5 – Примеры работы методов локальной оптимизации при поиске максимума функции, изображенной на рисунке (а): (б) метод BFGS, (в) метод Нелдера-Мида, (г) метод Пауэлла.

В **разделе 1.6** приведены существующие методы перебора моделей демографической истории популяций, а также модификации подходов сравнения моделей с разным числом параметров в условиях зависимости генетических данных. На момент начала исследований в 2017 году не существовало метода автоматического перебора моделей демографической истории популяций. Все сравнения моделей проводились пользователем вручную с использованием информационного критерия Акаике или теста отношения правдоподобия. После публикации первой статьи [1] диссертанта появилось единственное программное средство [54], реализующее аналог метода автоматического перебора моделей. Однако он ограничен анализом двух популяций и предполагает независимость данных. В общем случае, генетические данные имеют зависимости: определенные части генома наследуются вместе. Если данные зависимы, то информационный критерий Акаике и тест отношения правдоподобия будут ошибочно отдавать предпочтение моделям с большим числом параметров [68]. Существующие модификации этих конструкций [69] позволяют учитывать эти зависимости.

Во **второй главе** описаны разработанный расширенный класс моделей демографических историй, методы настройки их параметров на основе комбинации методов глобальной и локальной оптимизации, а также экспериментальные исследования разработанных моделей и методов.

Раздел 2.1 содержит описание разработанного класса расширенных моделей демографической истории, реализацию этих моделей и примеры использования. В качестве прототипа был выбран первый класс моделей. Разработанные модели включают новый тип параметров для настройки — дискретные параметры динамики изменения численности, которые могут представляться одной из трех зависимостей: постоянная численность, линейное или экспоненциальное изменение. Изображение предложенной модели, а также демографические истории при разных значениях параметра `Dyn` показаны на рисунке Р.6.

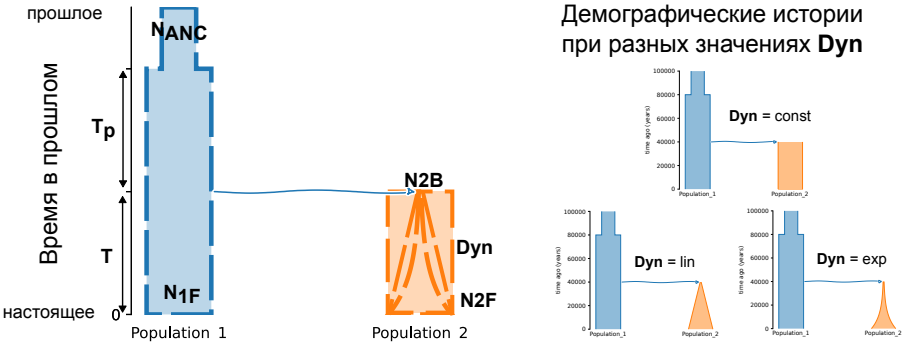


Рисунок Р.6 – Пример расширенной модели демографической истории двух популяций и демографические истории при разных значениях параметра `Dyn`

В **разделе 2.2** приведено формальное описание разработанного метода настройки параметров моделей расширенного класса на основе комбинации генетического алгоритма и метода локального поиска. Описаны общая схема, разработанные операторы мутации и скрещивания в генетическом алгоритме, реализация и примеры применения предложенного метода. Кроме того, в разделе приведены результаты автоматической настройки гиперпараметров генетического алгоритма для более эффективного решения поставленной задачи.

Раздел 2.3 описывает второй разработанный метод настройки параметров моделей расширенного класса — на основе комбинации байесовской оптимизации и метода локального поиска. Приведено описание байесовской оптимизации и ее компонент, существующего метода кросс-валидации для выбора некоторых гиперпараметров, а также реализации разработанного метода. Как и в случае с генетическим алгоритмом, были настроены гиперпараметры байесовской оптимизации. Проведены экспериментальные исследования для ручной настройки, на основе которых была предложена ансамблевая байесовская оптимизация.

Раздел 2.4 включает экспериментальные исследования разработанного метода настройки параметров на основе комбинации генетического алгоритма и локального поиска.

Сначала были проведены экспериментальные исследования разработанного метода в сочетании с методом вычисления правдоподобия, реализованным в *moments*. Проведено сравнение разработанного метода с методом Пауэлла с перезапусками и методом последовательных запусков метода Нелдера-Мида, реализованным в *moments-pipeline*. Для сравнения использовались модели первого класса и три набора симулированных данных. Сравнение показало, что разработанный метод (GA) позволяет более эффективно настраивать параметры моделей (таблица Р.1). Дополнительно были рассмотрены предложенные расширенные модели и была проведена настройка их параметров с использованием разработанного метода. Метод позволил корректно настроить параметры расширенных моделей, включая динамики изменения численности популяций.

Таблица Р.1 – Результаты экспериментальных исследований сравнения методов настройки параметров на симулированных данных трех популяций

	Метод Пауэлла с перезапусками	*moments-pipeline*	GA
Среднее число вычислений	22 475	19 452	21 651
Среднее $f^{moments}$	−11 178,62	−11 179,82	**−11 178,45**
Стандартное отклонение $f^{moments}$	0,40	0,72	**0,15**
Лучшее $f^{moments}$	−11 178,31	−11 178,59	**−11 178,29**

Затем были проведены три экспериментальных исследования разработанного метода в сочетании с методом вычисления правдоподобия, реализованным в *дадi*.

Метод был применен для реальных данных популяций кошачьей лягушки (*Scotobleps gabonicus*), которые ранее были проанализированы в [49] с применением *dadi-pipleine*. Для трех различных пар популяций были получены демографические истории с использованием тех же 12 моделей, которые были использованы в [49]. Для 92% моделей разработанный метод нашел параметры с большим значением правдоподобия, чем было получено с применением *dadi-pipleine*. В 5% случаев правдоподобие совпало и только для одной модели (3%) оно оказалось хуже.

На данных двух популяций американской пумы (*Puma concolor*) разработанный метод был сравнен с методом BFGS с перезапусками и методов BOBYQA с перезапусками. Использовались две модели, предложенные и проанализированные ранее в работе [51], без инбридинга и с инбридингом. Показано, что разработанный метод в среднем оказывается более эффективным, чем методы BFGS и BOBYQA (таблица Р.2). Дополнительная настройка параметров моделей с использованием расширенной области значений параметров позволила получить демографическую историю, имеющую значение правдоподобия выше, чем было получено ранее.

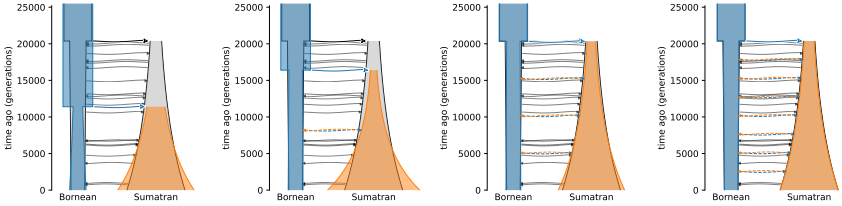
Таблица Р.2 – Результаты 100 запусков различных методов для поиска параметров модели с инбридингом для вывода демографической истории двух популяций пум

	BFGS		BOBYQA		GA
	1 запуск	16 запусков	1 запуск	4 запуска	
Число вычислений правдоподобия	394 ± 82	$6\,245 \pm 324$	$1\,605 \pm 1\,207$	$6\,095 \pm 2\,561$	$6\,193 \pm 2\,680$
Время CPU (мин.)	$1{,}3 \pm 1{,}4$	25 ± 19	12 ± 5	16 ± 7	93 ± 47
Лучшее правдоподобие	$-317\,370{,}88$	$-317\,370{,}88$	$\mathbf{-317\,239{,}48}$	$\mathbf{-317\,239{,}48}$	$-317\,239{,}49$
Среднее правдоподобие	$-1\,729\,870$	$-320\,947$	$-381\,979$	$-320\,503$	$\mathbf{-319\,451}$
Стандартное отклонение правдоподобия	$4\,339\,276$	$5\,029$	$115\,205$	$8\,753$	$7\,340$

Аналогичным образом разработанный метод был сравнен с методом BOBYQA с перезапусками на данных огородной капусты (*Brassica oleracea*), которые ранее были проанализированы в работе [51]. На этих данных метод BOBYQA с множественными запусками показал результаты лучше, чем разработанный метод, однако, отметим, что число перезапусков, необходимых для достижения эффективности метода BOBYQA, остается неизвестным в общем случае. Было рассмотрено расширенное пространство значений параметров, что позволило получить демографические истории, имеющие лучшее значение правдоподобия, чем получено ранее.

Затем, четыре существующих метода вычисления правдоподобия, реализованные в *дадi*, *moments*, *momi2* и *momentsLD*, были сравнены с использованием разработанного метода на данных орангутанга, симулированных с использо-

ванием библиотеки *stdpopsim* [6, 7]. Сравнение было проведено с применением различных моделей, включая модели расширенного класса. Было показано, что все методы позволяют восстановить исходную демографическую историю при применении корректных моделей. Использование моделей, которые не способны отразить исходную историю, например, не включающих непрерывные миграции, приводит к различиям в получаемых результатах (рисунок Р.7). Однако отметим, что основные характеристики популяций такие, например, как численность, настраиваются корректно и в таком случае.



(а) (б) (в) (г)

Рисунок Р.7 – Результаты настроенных моделей для метода вычисления правдоподобия, реализованного в *momi2* (а) модель без миграций, и модели с (б) одной, (в) тремя, (г) семью единичными миграциями

Наконец, была выведена демографическая история трех популяций современного человека на территории России: жителей Пскова, Новгорода и Якутии. Используемые данные ранее не были проанализированы. Параметры расширенной модели были настроены с помощью разработанного метода на основе комбинации генетического алгоритма и локального поиска. Полученная демографическая история (рисунок Р.8) согласуется с известной историей современного человека [56, 70].

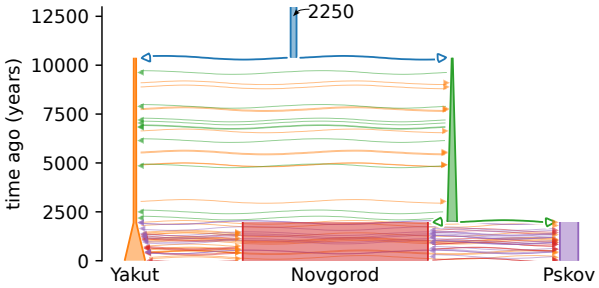
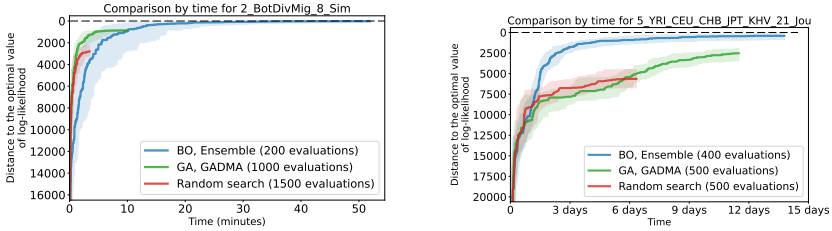


Рисунок Р.8 – Полученная демографическая история трех популяций современного человека

Раздел 2.5 описывает результаты экспериментальных исследований разработанного метода настройки параметров моделей на основе комбинации байесовской оптимизации и локального поиска.

Байесовская оптимизация была сравнена с генетическим алгоритмом на множестве наборов данных (рисунок Р.9). Было показано, что генетический алгоритм (GA) оказывается более эффективным в случае одной, двух и трех популяций. Байесовская оптимизация (BO) имеет более быструю сходимость, чем генетический алгоритм, если рассматривается более трех популяций. Применение байесовской оптимизации позволяет сократить время настройки параметров моделей на 50-80%, что приводит к значительному ускорению процесса на дни и даже недели.



(а) (б)

Рисунок Р.9 – Сходимость по времени методов настройки параметров моделей для (а) двух популяций, (б) пяти популяций

Также разработанный метод на основе комбинации байесовской оптимизации и метода локального поиска был использован для настройки параметров моделей демографических историй четырех и пяти популяций современного человека по реальным данным из статьи [55]. Разработанный метод позволил получить параметры, которые дают лучшее значение правдоподобия, чем найденные в [55] с помощью метода Пауэлла с перезапусками. Сравнение полученной демографической истории с более высоким значением правдоподобия и истории, полученной в оригинальной статье, представлено на рисунке Р.10.

В **третьей главе** описан разработанный метод автоматического перебора расширенных моделей демографической истории одной, двух и трех популяций, а также результаты его применения в сочетании с разработанным методом на основе комбинации генетического алгоритма и локального поиска.

В **разделе 3.1** приведено формальное описание разработанного метода. На вход подаются минимальные и максимальные ограничения на модель. На первом раунде метод строит модель, удовлетворяющую минимальным ограничением и выполняет настройку ее параметров с применением разработанного метода на основе генетического алгоритма. На каждом следующем раунде происходит изменение модели, увеличение числа ее параметров и последующая настройка

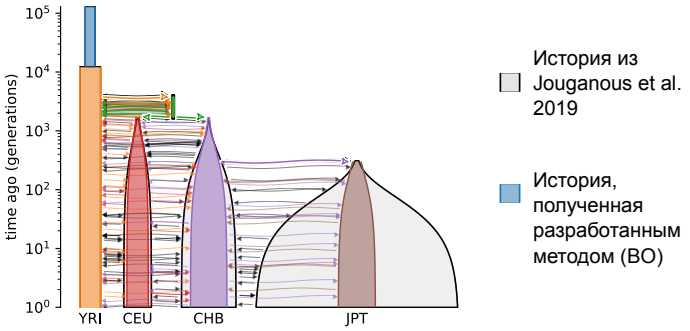


Рисунок Р.10 – Сравнение демографической истории, полученной разработанным методом (BO), и демографической истории из [55]

нового набора параметров. Работа метода останавливается, когда модель достигает максимальных ограничений. В конце происходит сравнение всех перебранных моделей с использованием информационного критерия Акаике и выбирается наилучшая. В качестве ограничений на модели было предложено число временных интервалов.

В **разделе 3.2** приведены экспериментальные исследования разработанного метода автоматического перебора моделей. Для генетических данных трех популяций современного человека была получена демографическая история «выхода из Африки», представленная на рисунке Р.11. Она согласуется с другими исследованиями [17, 45, 70] и имеет не только наилучшее значение правдоподобия, чем история, полученная ранее в [45] по тем же данным, но и лучшее значение информационного критерия Акаике.

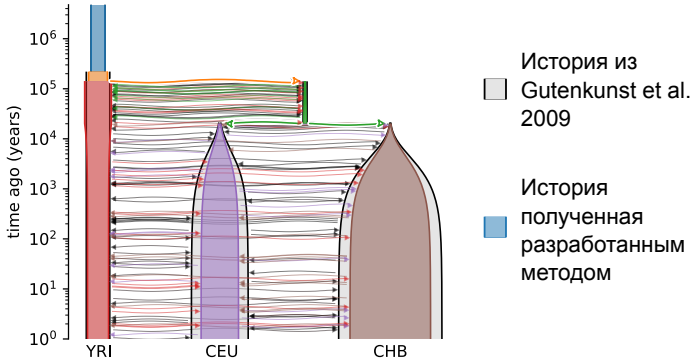


Рисунок Р.11 – Демографическая история, полученная разработанным методом

Для трех наборов генетических данных популяции кошачьей лягушки (*Scotobleps gabonicus*), на которых были построены модели в разделе 2.4 данной работы с использованием метода ручного перебора, также были получены модели демографических историй с использованием разработанного метода автоматического перебора. Из двух наборов данных полученные модели показали наилучшее значение информационного критерия Акаике среди всех рассмотренных конфигураций. В случае третьего набора данных полученная модель имеет значение информационного критерия Акаике, которое хуже, чем у лучшей модели, полученной в результате ручного перебора. Тем не менее, результаты позволяют выявить излишний параметр модели, и исключение этого параметра из конфигурации приводит к наилучшему значению информационного критерия Акаике.

Разработанный метод автоматического перебора расширенных моделей был использован при выводе демографической истории популяций голубой акулы. Генетические данные ранее не были проанализированы. Был разработан подход последовательного вывода демографической истории двух и трех популяций, в результате которого была получена демографическая история, представленная на рисунке Р.12. Полученные численности популяций согласуются с другими исследованиями [57, 58]. Апробация результатов, проведенная коллегами из области зоологии, позволила предположить, что разделение северной и южной популяций произошло в связи с палеоклиматическими событиями в эпоху голоцена [71–73].

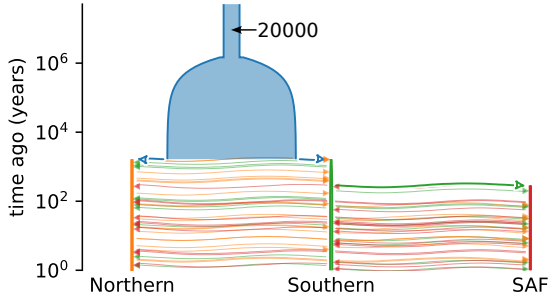


Рисунок Р.12 – Демографическая история трех популяций голубой акулы

В **четвертой главе** приведено описание программных комплексов, которые реализовывают разработанные методы или были использованы в данной работе.

Раздел 4.1 содержит описание программного комплекса GADMA (Global seach Algorithm for Demographic Model Analysis), который реализует разработанные модели расширенного класса, методы настройки параметров этих моделей и метод автоматического перебора расширенных моделей. Структура программного комплекса представлена на рисунке Р.13.

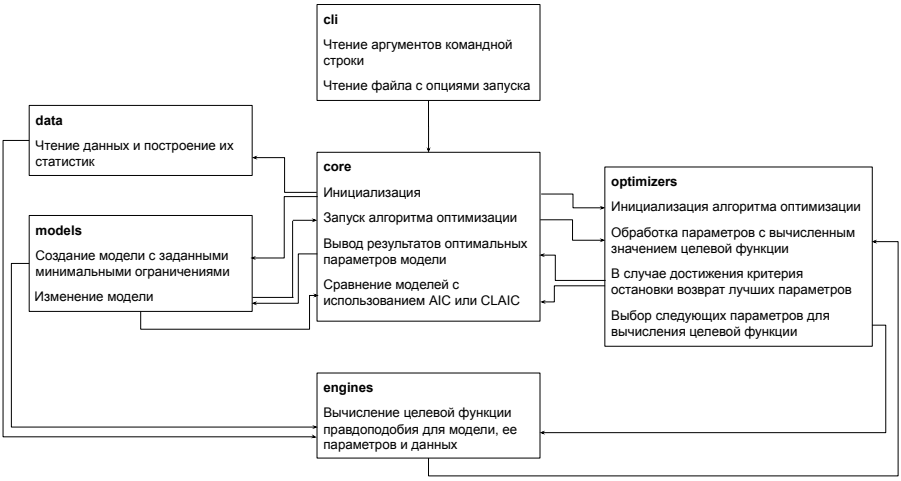


Рисунок Р.13 – Структура программного комплекса GADMA

В **разделе 4.2** приведено описание расширения библиотек *stdpopsim* и *demes*. Библиотека *stdpopsim* содержит каталог предопределенных видов и их демографических историй для более надежных симуляций генетических данных. Эта библиотека была расширена и использована при проведении экспериментальных исследований. Программное средство *demes* предназначено для текстового и визуального представления демографических историй. Эта библиотека была расширена реализацией линейного изменения численности популяции, а также была интегрирована в программный комплекс GADMA. Все изображения демографических историй в данной диссертации были получены с использованием библиотеки *demes*.

В **заключении** приведены основные результаты работы, которые состоят в следующем:

- проведено исследование текущего состояния предметной области, уточнение задачи и способов оценки результатов;
- формализована постановка задачи построения и настройки моделей метрических деревьев с функциями на ребрах на примере задачи вывода демографической истории популяций по генетическим данным;
- разработан метод автоматической настройки параметров моделей метрических деревьев с функциями на ребрах на основе комбинации методов глобальной и локальной оптимизации на примере задачи вывода демографической истории популяций по генетическим данным;
- разработан метод автоматического перебора моделей метрических деревьев с функциями на ребрах на примере задачи вывода демографической истории популяций по генетическим данным;

– спроектирован и реализован программный комплекс, включающий разработанные модели и методы для вывода демографической истории популяций по генетическим данным;
– проведены экспериментальные исследования, подтверждающие эффективность разработанных моделей и методов, а также их применимость для вывода демографической истории популяций по генетическим данным, проведен анализ результатов экспериментов.

Для оценки качества настройки моделей демографических историй в данной работе было использовано значение функции правдоподобия. Результаты экспериментов показывают, что метод настройки параметров моделей на основе комбинации генетического алгоритма и локального поиска позволил в 88% случаев (37 моделей из 42 протестированных) найти параметры модели, обеспечивающие лучшее значение правдоподобия, чем параметры, найденные существующими ранее методами. На симулированных данных разработанный метод позволил найти решения, которые на 97% ближе к оптимуму в случае одной популяции и на 66% ближе к оптимуму в случае трех популяций, чем решения, полученные существующими методами. Настройка гиперпараметров генетического алгоритма позволила ускорить реализацию в среднем на 10% с сохранением эффективности метода.

Была подтверждена эффективность метода настройки параметров моделей на основе байесовской оптимизации и локальной оптимизации в условиях сложновычислимной целевой функции. Разработанный метод позволил найти значения параметров, обеспечивающих лучшее значение правдоподобия, чем существующие методы, для двух ранее проанализированных данных четырех и пяти популяций. Было показано, что байесовская оптимизация достигает решения, близкого к оптимуму, на 50-80% быстрее, чем генетический алгоритм, в случае вывода демографической истории четырех и пяти популяций.

Метод автоматического перебора моделей позволяет автоматически строить и настраивать модели в заданных ограничениях на конфигурацию. Сравнение моделей демографических историй с разным числом параметров было осуществлено с использованием информационного критерия Акаике (AIC). Экспериментальные исследования показали, что в трех из четырех случаях метод позволил найти модель, обеспечивающую лучшее значение AIC, чем было получено ранее ручным перебором. В четвертом случае, полученная модель позволила установить излишние параметры в конфигурации и построить вложенную модель, которая в итоге обеспечила наилучшее значение AIC для данных.

В качестве перспективных направлений исследования можно выделить совершенствование метода автоматического перебора моделей с целью поиска оптимального набора параметров конфигурации, а также разработку методов настройки моделей метрического дерева с функциями на ребрах, которые позволяют осуществлять настройку не только функциональных параметров, но и поиск оптимальной структуры дерева.

Публикации автора по теме диссертации

Публикации в зарубежных изданиях, индексируемых в базах цитирования Web of Science или Scopus

1. ***Noskova E.***, *Ulyantsev V.*, *Koepfli K.-P.*, *O'Brien S. J.*, *Dobrynin P.* GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data // GigaScience. — 2020. — Vol. 9, no. 3. — giaa005. — DOI: 10.1093/gigascience/giaa005.
2. ***Noskova E.***, *Abramov N.*, *Iliutkin S.*, *Sidorin A.*, *Dobrynin P.*, *Ulyantsev V.* GADMA2: more efficient and flexible demographic inference from genetic data // GigaScience. — 2023. — Vol. 12. — giad059. — DOI: 10.1093/gigascience/giad059.
3. ***Noskova E.***, *Borovitskiy V.* Bayesian optimization for demographic inference // G3, Genes | Genomes | Genetics. — 2023. — Vol. 13, no. 7. — DOI: 10.1093/g3journal/jkad080. — jkad080.
4. *Zhernakova D. V.*, ..., *Ulyantsev V.*, ***Noskova E.***, ..., *O'Brien S. J.* Genome-wide sequence analyses of ethnic populations across Russia // Genomics. — 2020. — Vol. 112, no. 1. — Pp. 442–458. — DOI: 10.1016/j.ygeno.2019.03.007.
5. *Nikolic N.*, *Devloo-Delva F.*, *Bailleul D.*, ***Noskova E.***, ..., *Arnaud-Haond S.* Stepping up to genome scan allows stock differentiation in the worldwide distributed blue shark Prionace glauca // Molecular Ecology. — 2023. — Vol. 32, no. 5. — Pp. 1000–1019. — DOI: 10.1111/mec.16822.
6. *Adrion J. R.*, ..., ***Noskova E.***, ..., *Kern A. D.* A community-maintained standard library of population genetic models // eLife. — 2020. — Vol. 9. — e54967. — DOI: 10.7554/eLife.54967.
7. *Lauterbur M. E.*, ..., ***Noskova E.***, ..., *Gronau I.* Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations // eLife / ed. by Z. Gao, M. Przeworski. — 2023. — June. — Vol. 12. — DOI: 10.7554/eLife.84874.
8. *Gower G.*, *Ragsdale A. P.*, *Bisschop G.*, *Gutenkunst R. N.*, *Hartfield M.*, ***Noskova E.***, *Schiffels S.*, *Struck T. J.*, *Kelleher J.*, *Thornton K. R.* Demes: a standard format for demographic models // Genetics. — 2022. — Vol. 222, no. 3. — DOI: 10.1093/genetics/iyac131. — iyac131.

Synopsis

Thesis overview

The relevance. Metric tree models with functions on edges are used to analyze and predict various events of the real world, for example, processes represented as dynamic systems with variable structure [9, 10]. *Metric tree* is a graph that is a tree, where each edge is associated with an interval. In general, metric graphs with functions on edges have found wide application, for example, in the form of quantum graphs [11], which are used in physics in the study of quantum chaos [12], waveguides [13] and photonic crystals [14].

Model inferring is a set of actions aimed at selecting the configuration, defining the model parameters and adjusting their values in order to achieve a high correspondence of the modeling results to the data of a full-scale experiment. Expert data or assumptions about the object under study are usually required at various stages of model inferring. These data may be inaccurate, limited or unknown, which can negatively affect the accuracy and adequacy of the model. Automated construction methods allow to reduce the probability of human errors in model selection and parameters tuning.

When working with metric tree models with functions on edges, the participation of subject matter experts is resorted to. Expert data is used to determine the properties of functions on the edges of the tree. This information allows to establish the *configuration* of the model, where each function on the tree belongs to a given family and is characterized by functional parameters available for tuning. In the absence of expert data or in order to minimize the influence of the expert on the result, it is necessary to consider a set of all possible models differing in the types of functions and functional parameters. For example, when building models of demographic histories for each population, piecewise-defined functions are considered consisting of functions of the three most popular types: constant, linear, and exponential. Such an enumeration of configurations leads to an increase in the time required to infer the model. The greater the number of allowed function types is, the more time will be required. Additionally, it is required to keep track of the model complexity, the number of its parameters and overfitting.

Methods for tuning model parameters may also be limited in their degree of automation and require expert input. For example, local search methods require the involvement of an expert to determine initial parameter values, and the effectiveness of the tuning depends on this choice.

Thus, when modeling real-world events in the form of a metric tree with functions on edges it is *relevant* to develop specialized models and methods for automatic inference and tuning of models in order to minimize the influence of expert data on the result of modeling, which is considered in this dissertation on the example of the task of inferring demographic histories from genetic data.

A population is a group of individuals of the same species living in a specific area. The *demographic history of populations* is the history of populations' development and evolution, including events such as changes in population size, population splits, migration, and natural selection. The reconstruction of the demographic history from genetic data is called *demographic inference*. Demographic histories are essential for dating historical events which have left no written records [15, 16], and they hold significance in fields such as conservation genetics [5] and even medicine [17].

Various statistical and algorithmic methods allow inference of demographic history models in the form of metric trees with functions on the edges and tuning their continuous parameters from genetic data. In a case of demographic histories, the metric tree is the tree that defines the separation of populations, and the functions on the edges are the dynamics of population change. As dynamics, we consider piecewise defined functions consisting of functions of the three most popular types: constant, linear, and exponential. When building models, it is necessary to determine the number of time intervals, as well as the type of dynamics for each piecewise defined function.

Expert data is also involved at the stage of tuning the parameters of demographic history models, for which a combination of numerical simulation and optimization methods are used. Numerical modeling methods are used to calculate the likelihood function, which allows estimating the degree of model fit to genetic data. Local optimization methods are used to find the parameters that provide the maximum likelihood value. It is these methods that are limited in the degree of automation: they require expert data to determine the initial values of parameters, and their efficiency depends on this choice.

The task of demographic inference is further complicated by the need for the user to implement the program code of the model and the algorithm for parameters' tuning. Numerical modeling methods used by existing solutions have different capabilities and stability, and the user can apply several of them to compare the results. However, when applying different software solutions simultaneously, the user is faced with the need to specify the same models using different interfaces.

Thus, the development of methods for automatic construction and tuning for models of metric trees with functions on edges will lead to minimizing the influence of expert data, and, consequently, to improving the quality of modeling of real-world events using real data from experiments.

State of the art. Graph models are studied and applied to solve a wide range of problems. The works of A.M. Raygorodsky [18, 19] contain descriptions and examples of application of random graph models. Graph-based probabilistic models such as Bayesian networks are extensively presented in the works of I. Ben-Gal [20] for modeling industrial systems [21], classification [22], or identification of transcription factor binding sites [23]. L. Clark and D. Pregibon [24] described examples of applications of tree-based models, which include, for example, decision trees [25].

The theory of metric graphs was formed by the works of V.G. Boltiansky [26], P.S. Soltan [26, 27] and A. Dress [28]. The properties of metric trees and the met-

ric spaces generated by them have been studied by A. Dress [28], B. Buneman [29] and D. Aldous [10, 30, 31]. In the works of A.S. Matveev and S.I. Matveev [32–34] metric graphs were applied in the construction of coordination models for intelligent navigation.

The development of models approximating implicit functions is also actively pursued by many scientists. The most widespread application, described in the works of L. Fahrmeir [35] and R. Snee [36], these models have received for solving regression problems. When using piecewise-determined function models, the general form of the result functions is usually fixed, such as constructing piecewise-constant [37, 38], piecewise-linear [39], or piecewise-exponential [40] models. The number of function breakpoints as well as their positions are unknown characteristics of piecewise-defined function models. In [41, 42], methods for automatic model inference are presented to solve a piecewise-exponential regression problem, where the number of function breakpoints is determined using Bayesian information criterion (BIC) and Akaike information criterion (AIC) [43], respectively.

Models of metric trees with functions on graphs are a combination of metric tree models and functional models on edges. Quantum graphs that are metric graphs with differential operators on edges and their applications are discussed in detail in the works of G. Berkolaiko [11, 44]. Metric trees with functions on edges are used to model demographic histories of populations in the works of R. Gutenkunst [45], J. Kamm [46], A. Ragsdale and S. Gravel [47, 48]. However, the methods presented in these works assume that the user defines and fixes the general form of the piecewise-defined function on the edges of the tree, and sets the initial values of model parameters for tuning procedure that use local optimization methods. The works of D. Portik [49, 50] and R. Gutenkunst [51] presented global optimization methods for parameter tuning of population history models that minimize but still require user involvement. A general application of numerical optimization methods for different problems is presented in the classic paper by B.T. Polyak [52], and a description of modern global optimization methods can be found in [53].

At the time the author began his research (in 2017), there was no method for automatic model inference and tuning for metric tree models with functions on edges. By the end of the dissertation research, the first alternative solution for a method for automatic model selection emerged, applied for the demographic inference problem [54]. However, the method allows analyzing models defined in a specific catalog and only for inferring demographic histories of *two* populations. Furthermore, the selection of the best model is made under the assumption of data independence, which is not always correct.

The **aim** of this thesis is to improve the quality of computer modeling of real-world events by developing methods, and software packages for automatic inference and tuning of metric tree models with functions on edges.

In order to achieve this aim, the following **tasks** have been defined and completed:

- investigate the current state of the subject area, refine the problem, and determine methods for evaluating the results;
- formalize the problem of model inferring and tuning for metric tree models with functions on edges;
- develop a method for automatic tuning of metric tree models with functions on edges based on a combination of global and local optimization methods;
- develop a method for automatic selection of metric tree models with piecewise defined functions on edges;
- design and implement a software framework that incorporates the developed models and methods for inferring the demographic history of populations from genetic data;
- conduct experimental studies confirming the effectiveness of the developed models and methods, as well as their applicability for inferring the demographic history of populations from genetic data, analyze the results of experiments.

The **scientific novelty** of the thesis is as follows: (1) methods based on a combination of global and local optimization techniques for parameter tuning of a given metric tree model with functions on edges are developed; (2) method for automatic selection of metric tree model with piecewise-defined functions on edges that does not require expert involvement is developed.

The **theoretical significance** of the thesis lies in = extension of the classical formulation of the problem of tuning a metric tree model with functions on edges not only as a problem of tuning the parameters of a given model, but also as a problem of selecting the model itself automatically. The obtained modeling and tuning methods are applicable to arbitrary metric tree models with functions on edges. Moreover, the developed optimization methods can be used or adapted for optimization problems in other scientific fields.

The **practical significance** of the thesis is determined by:

- the extension of the scientific and practical toolkit of bioinformaticians with methods and algorithms for demographic inference;
- the open-access source code of the developed software framework GADMA, which is available for reuse at the following address: `https://github.com/ctlab/GADMA`;
- the applicability of the developed methods for the analysis of genetic data;
- incorporation of the developed method based on a genetic algorithm into a third-party software [54].

Principal statements of the thesis:

1. The method of modeling and parameter tuning of metric tree models with functions on edges based on field experiment data, that contains models with continuous functional parameters, characterized in that for the purpose of automatic tuning without involving expert data it uses models with discrete parameters that define families of functions, as well as global optimization methods — genetic algorithm and Bayesian optimization, and a set of programs implementing it.
2. The method of automatic selection of metric tree model with functions on edges with different number of parameters and tuning of these parameters basen on field experiment data, that contains the Akaike information criterion for model comparison, characterized in that in order to increase the level of automatization and to provide an opportunity to tune not only the model parameters, but also the model itself, it includes a method of increasing the number of time intervals for piecewise-defined functions on edges of a tree, as well as a set of programs implementing it.

Research methods. The study utilized optimization methods, numerical methods, probability theory and mathematical statistics, machine learning techniques, and methods for conducting experimental research.

Soundness and correctness of scientific results obtained in this thesis are ensured by the correct utilization of methods, the formulation of well-justified problem statements, and the extensive experimental investigations that cover the developed technologies and algorithms. The population demographic histories obtained with the developed methods on verified simulated data are consistent with the original histories used for modeling. The results obtained from real data align with previously published studies [45, 55–59].

Compliance with specialty requirements. In accordance with the specialty passport 1.2.2 — «Mathematical modeling, numerical methods, and software frameworks (computer science)» the dissertation belongs to the following fields of research:

Point 2 of the specialty passport «Development, justification, and testing of efficient computational methods using modern computer technologies». Methods for tuning the parameters of metric tree models with functions on edges based on numerical optimization methods were developed, justified and tested.

Point 4 of the specialty passport «Development of new mathematical methods and algorithms for interpretation of natural experiment on the basis of its mathematical model». This dissertation study presents methods for building metric tree models with functions on edges from natural experiment data in order to analyze real world events.

Dissemination. The main results of the thesis were presented at the following conferences:

- International Congress «VII Congress of the Vavilov Society of Geneticists and Breeders dedicated to the 100th anniversary of the Department of Genetics, SPbSU, and associated symposia», 2019, St. Petersburg, Russia;
- Moscow Conference on Computational Molecular Biology, 2019, Moscow, Russia;
- Probabilistic Modeling in Genomics, 2019, Aussois, France;
- Probabilistic Modeling in Genomics, 2021, virtual;
- Moscow Conference on Computational Molecular Biology, 2021, Moscow, Russia;
- Probabilistic Techniques in Analysis: Spaces of Holomorphic Functions, 2021, Sochi, Russia;
- LI Scientific and educational conference of ITMO University, 2022, ITMO University, St. Petersburg, Russia;
- Probabilistic Modeling in Genomics, 2022, Oxford, UK;
- The XI Congress of Young Scientists, 2022, ITMO University, St. Petersburg, Russia;
- Conservation Genomics at the Population Level, 2022, Cambridge, UK;
- Probabilistic Modeling in Genomics, 2023, Cold Spring Harbor, NY, USA;
- The XII Congress of Young Scientists, 2023, ITMO University, St. Petersburg, Russia;
- Society for Molecular Biology and Evolution Meeting (SMBE23), 2023, Ferrara, Italy.

Awards

- Bronze Award in the 17th Human-Competitive Awards category at the Genetic and Evolutionary Computation Conference (GECCO) virtual conference in 2020.
- Winner of the System Biology Fellowship from Skolkovo Institute of Science and Technology for the project «Computational methods for unsupervised demographic inference of multiple populations from genomic data» in 2021. The number of winners is five per country per year.

Publications

Based on the results presented in the thesis, eight articles were published in peer-reviewed scientific journals included in the international abstract databases and citation systems Scopus and Web of Science.

Personal contribution

1. In the publication [1] Noskova E. — development and implementation of genetic algorithm and method for automatic selection of demographic history model from genetic data, conduction of the experimental studies (80%);

Ulyantsev V. — supervision on problem formulation, selection, and justification of theoretical foundations of scientific research (10%); Koepfli K.P., O'Brien S.J. — advice in conducting experimental studies and writing a paper (5%); Dobrynin P. — supervision on problem formulation (5%).

2. In the publication [2] Noskova E. — development and implementation of methods, software for demographic inference from genetic data, conduction of the experimental studies (85%); Abramov N., Iliutkin S., Sidorin A. — software development (10%); Dobrynin P., Ulyantsev V. — supervision on problem formulation, selection, and justification of theoretical foundations of scientific research (5%).
3. In the publication [3] Noskova E. — development and implementation of the Bayesian optimization method for demographic inference from genetic data, conduction of the experimental studies (90%); Borovitskiy V. — recommendations on problem formulation, selection, and justification of theoretical foundations of scientific research (10%).
4. In the publication [4] Noskova E. — demographic inference of the history of three modern human populations (10%); Ulyantsev V. — supervision (5%); other co-authors — collection and analysis of the genetic data (85%).
5. In the publication [5] Noskova E. — demographic inference of the history of two and three populations of blue sharks (10%); the other co-authors — collection and analysis of the genetic data (90%).
6. In the publication [6] Noskova E. — software development and testing of the genetic data simulations (5%); other co-authors — software development, testing, and conduction of the experimental studies (95%).
7. In the publication [7] Noskova E. — implementation of published demographic histories in the software for genetic data simulations (5%); other co-authors — software development (95%).
8. In the publication [8] Noskova E. — software development for the representation of the demographic histories (5%); other co-authors — software development (95%).

.

Scope and structure of the work

The dissertation consists of an introduction, four chapters, a conclusion, and an appendix. The full length of the dissertation is 396 pages, including 120 figures, 16 tables and eight listings. The list of references contains 171 references.

Thesis contents

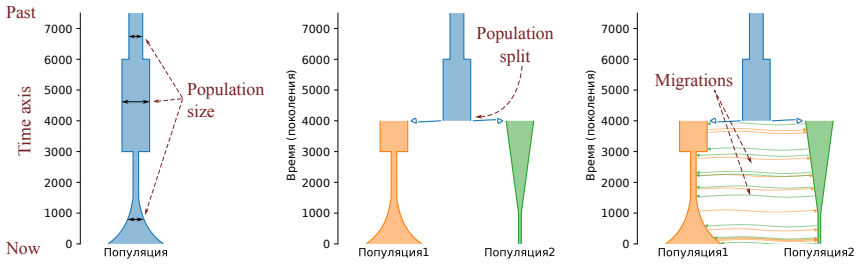
The relevance of the scientific research conducted within the framework of this dissertation is justified in the **introduction**. The state of the art of the demographic inference from genetic data is described, and an overview of the methods for modeling demographic histories is provided. The goals and objectives of the research are formulated, and the scientific novelty, theoretical significance, and practical implications of the work are described. Additionally, the principal statements of the thesis are listed.

The **first chapter** presents a comprehensive overview of the subject area. It includes the definition of the demographic history of populations and the description of the existing methods for inferring demographic histories from genetic data.

The fundamental definitions in population genetics used in this dissertation are described in **Section 1.1**. Population genetics is an important field within genetics that investigates the changes in the genetic composition of populations and their evolution. It addresses various tasks such as determining population structure, constructing phylogenetic trees, and inferring the demographic history of populations. The section begins with the formal definition of the demographic history of populations.

Demographic history of populations refers to the populations' history of evolution and development. It includes information about population splits, population size changes, migration rates, *inbreeding coefficients* (the degree of consanguineous relationships), and more. Examples of visual representations of demographic histories are shown in Figure S.1. The colored areas demonstrate how populations split — the population tree. Information about population size is depicted by the width of those areas, and arrows between them represent migrations. Time in demographic histories is often measured in generations or years and is represented along the y-axis.

Demographic inference is the reconstruction of the demographic history from genetic data. The problem of demographic inference and its solutions are described in **Section 1.2**. The existing solutions aim to find demographic history using parametric models. The section includes the description of the main methods and components



(а) (б) (в)

Figure S.1 – Examples of visual representations of the demographic histories of one and two populations

of existing methods for demographic inference. A brief description of well-known software tools implementing these methods, namely ∂a∂i, *moments*, *momi2*, and *momentsLD*, is also provided.

Parametric models are used for searching demographic histories of populations. These models are metric trees with functions defined on edges. Models usage allows constraining the search space and tuning the model parameters using optimization methods. Figure S.2 shows a model example in the form of metric tree with functions on edges that describes the demographic history of two populations.

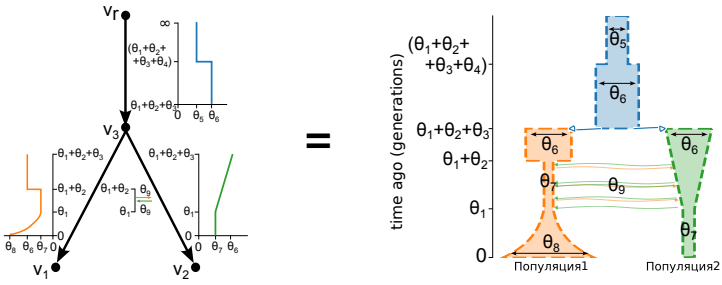


Figure S.2 – Example of the model of demographic history of two populations in the form of a metric tree with functions on edges

The problem of inferring the demographic history of populations from genetic data involves *parameter tuning* of a given parametric model, i.e. finding model parameters that maximize the likelihood function of the genetic data (Figure S.3). Existing software solutions differ in their model specification interfaces, likelihood computation methods, and parameter optimization methods.

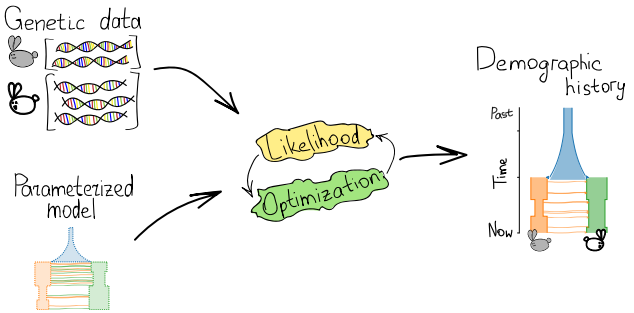


Figure S.3 – Example input and output of existing software solutions for the demographic history inference from genetic data

Two classes of demographic history models used in existing solutions are described in **Section 1.3**, along with methods for comparing models with different numbers of parameters.

Models of the first class are used in the software solutions ∂a∂i, *moments*, and *momentsLD*. They are represented as a sequence of elements corresponding to time intervals, population splits, pulse migrations, and inbreeding events (Figure S.4). These models have only continuous parameters, and the dynamics of population size (constant, linear, or exponential) are fixed within these models.

Models of the second class are used in the software solution *momi2*. They are represented as a set of events related to changes in population size, population splits, and single migrations. Models of the second class also have only continuous parameters and fixed dynamics of population size. However, they are more limited compared to models of the first class, for example, they do not support linear population size changes or continuous migrations.

The model selection problem is generally stated as follows: choose the model that is most suitable for the data. A simple model with a small number of parameters may not reflect some information in the data. On contrary, the complex model with a large number of parameters may overfit to noise in the data and ultimately model the underlying process incorrectly. To compare different models and choose the best one, the Akaike information criterion (AIC)[43], Bayesian information criterion (BIC)[60], and likelihood ratio test [61] are commonly used.

An overview of existing methods for computing the likelihood of genetic data given a specified demographic history is provided in **Section 1.4**. The section includes definitions of key biological and genetic concepts, such as DNA, genes, alleles, and genotypes. The main data statistics are described, including the allele frequency spectrum and statistics based on linkage disequilibrium. Finally, the methods for like-

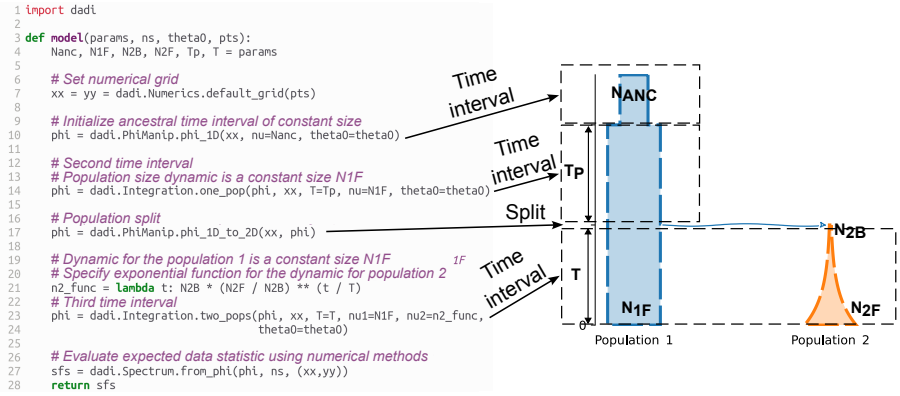


Figure S.4 – Example specification of the first class model using the ∂a∂i library interface

lihood evaluation implemented in the software solutions ∂a∂i, *moments*, *momi2*, and *momentsLD* are described.

In **Section 1.5**, a general description of local and global optimization methods is provided, highlighting the key differences between these two groups, along with an overview of existing optimization methods for parameter tuning in demographic history models using genetic data. Principally, local optimization methods such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [62–65], the Nelder-Mead method [66], and the Powell method [67] are employed for demographic inference (Figure S.5). Existing optimization methods for parameter tuning in models are limited to inference of continuous parameters only and require user involvement to specify initial model parameter and method hyperparameters, such as the number of restarts. The use of local optimization methods does not guarantee finding the global optimum.

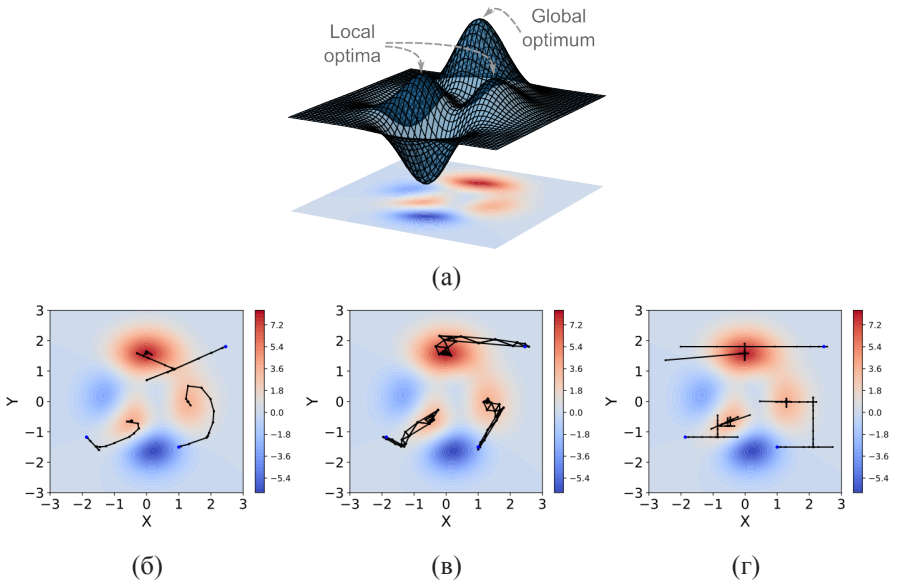


Figure S.5 – Examples of local optimization methods applied to maximize the function shown in panel (a): (b) BFGS method, (c) Nelder-Mead method, (d) Powell method.

Existing methods for selection of the best demographic model are described in **Section 1.6** alongside with modified approaches for model comparison. At the beginning of the research in 2017, there was no method available for automatic model selection for demographic inference. All model comparisons were performed manually by users using the Akaike information criterion or likelihood ratio test. A single software tool [54] implementing an alternative automatic model selection method became available after the publication of the first article of this dissertation [1]. How-

ever, it is limited to the analysis of two populations and assumes data independence. In general, genetic data contain dependencies, as certain genomic regions are inherited together. The Akaike information criterion and likelihood ratio test incorrectly favor models with more parameters when the data is dependent [68]. Existing modifications of these criteria [69] account for these dependencies.

Chapter 2 describes the developed class of extended demographic models, two methods for parameter tuning using a combination of global and local optimization methods, and experimental studies of the developed models and methods.

Section 2.1 provides a description of the developed class of extended demographic models, its implementation, and usage examples. The first class of models was used as a prototype of the class of extended models. The proposed extended models include a new type of parameters for tuning, which are discrete parameters representing one of three dependencies: constant population size, linear change, or exponential change. The representation of the proposed model and the demographic histories for different parameter values (`Dyn`) are shown in Figure S.6.

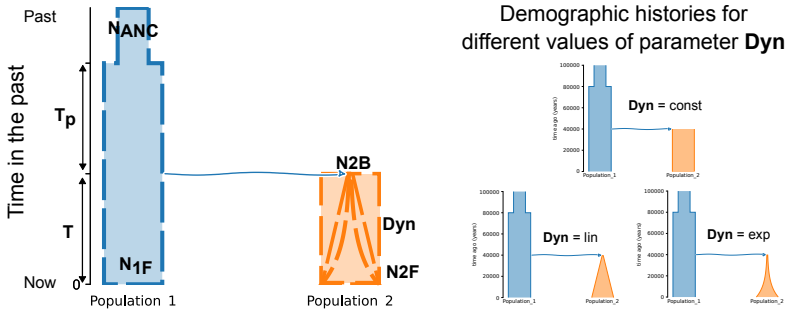


Figure S.6 – Example of an extended demographic model for two populations and the corresponding demographic histories for different values of the parameter `Dyn`

The formal description of the first proposed method for parameters tuning of the extended models in **Section 2.2**. It is based on a combination of a genetic algorithm and local search is presented. The section includes the description of the general scheme of method, the operators of mutation and crossover in the genetic algorithm. The implementation details, and examples of applying the proposed method are also included. Additionally, the section presents the results of hyperparameter tuning of the genetic algorithm that was performed to increase efficiency of the method.

Section 2.3 describes the second developed method for parameter tuning of extended models, that is based on a combination of Bayesian optimization and local search. The description of Bayesian optimization and its components, an existing cross-validation method for hyperparameter selection, and the implementation of the developed method are provided. Similar to the genetic algorithm, hyperparameters of Bayesian optimization are tuned. Based on conducted experimental studies for manual tuning, an ensemble Bayesian optimization approach was proposed.

Section 2.4 includes conducted experimental studies of the developed method for parameter tuning that is based on a combination of a genetic algorithm and local search.

First, experimental studies of the developed method in combination with the likelihood computation method implemented in *moments* are conducted. A comparison is made between the developed method and the Powell method with restarts and the sequential runs of the Nelder-Mead method implemented in *moments-pipeline*. Models from the first class and three sets of simulated data are used for comparison. The comparison shows that the developed method (GA) allows for more effective parameter tuning of models (Table S.1). Furthermore, the proposed extended models are considered, and their parameters are tuned using the developed method. The method correctly finds the parameters of the extended models, including the dynamics of population size changes.

Table S.1 – Results of experimental studies for comparing parameter tuning methods on simulated data of three populations

	Powell's method with restarts	*moments-pipeline*	GA
Mean number of likelihood evaluations	22,475	19,452	21,651
Mean $f^{moments}$	−11,178.62	−11,179.82	**−11,178.45**
Standard deviation of $f^{moments}$	0.40	0.72	**0.15**
Best $f^{moments}$	−11,178.31	−11,178.59	**−11,178.29**

Next, three experimental studies of the developed method in combination with the likelihood computation method implemented in ∂a∂i are conducted.

The method is applied to real data of Gaboon forest frog (*Scotobleps gabonicus*), previously analyzed in [49] using *dadi-pipeline*. Demographic histories for three different pairs of populations are obtained using the same 12 models used in [49]. For 92% of the models, the developed method finds parameters with higher likelihood values compared to those obtained using *dadi-pipeline*. In 5% of cases, the likelihood values are the same, and only for one model (3%) it is worse.

On data of two puma populations (*Puma concolor*), the developed method is compared to the BFGS method with restarts and the BOBYQA method with restarts. Two models, proposed and analyzed previously in [51], without and with inbreeding, are used. The results show that the developed method, on average, outperforms the BFGS and BOBYQA methods (Table S.2). Additional parameter tuning using an extended parameter bounds provides a demographic history with a higher likelihood value than previously achieved.

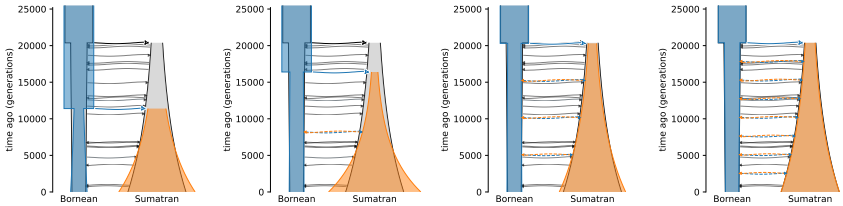
Similarly, the developed method is compared to the BOBYQA method with restarts on domestic cabbage data (*Brassica oleracea*), previously analyzed in [51]. On these data, the BOBYQA method with multiple restarts shows better results than the developed method. However, it should be noted that the number of restarts required

Table S.2 – Results of 100 repeats of different methods for parameter tuning in case of model 2 with inbreeding for two puma populations

	BFGS		BOBYQA		GA
	1 restart	16 restarts	1 restart	4 restarts	
Number of likelihood evaluations	394 ± 82	$6{,}245 \pm 324$	$1{,}605 \pm 1{,}207$	$6{,}095 \pm 2{,}561$	$6{,}193 \pm 2{,}680$
Time CPU (min)	1.3 ± 1.4	25 ± 19	12 ± 5	16 ± 7	93 ± 47
Best likelihood	$-317{,}370.88$	$-317{,}370.88$	$\mathbf{-317{,}239.48}$	$\mathbf{-317{,}239.48}$	$-317{,}239.49$
Mean likelihood	$-1{,}729{,}870$	$-320{,}947$	$-381{,}979$	$-320{,}503$	$\mathbf{-319{,}451}$
Standard deviation	$4{,}339{,}276$	$5{,}029$	$115{,}205$	$8{,}753$	$7{,}340$

for the BOBYQA method to achieve this efficiency is unknown in general. The demographic inference with extended parameter bounds provides histories with higher likelihood values than before.

Then, four existing likelihood computation methods implemented in *∂a∂i*, *moments*, *momi2*, and *momentsLD*, are compared using the developed method on orangutan data simulated using the *stdpopsim* library [6, 7]. The comparison is made using different models, including extended models from the proposed class. It is shown that all methods are able to recover the true demographic history when correct models are used. Using misspecified models, that could not capture the true history, such as models without continuous migrations, leads to differences in the obtained results (Figure S.7). However, it should be noted that key population characteristics, such as population size, are properly found even in such cases.



(а) (б) (в) (г)

Figure S.7 – Results of tuned models for the likelihood computation method implemented in *momi2* (a) model without migrations, and models with (b) one, (c) three, (d) seven pulse migrations

Finally, the demographic history of three populations of modern humans in Russia is inferred: the inhabitants of Pskov, Novgorod, and Yakutia. The data used in this analysis has not been previously analyzed. The parameters of the extended model are tuned using the developed method based on the combination of genetic algorithm and local search. The obtained demographic history (Figure S.8) is consistent with the known history of modern humans [56, 70].

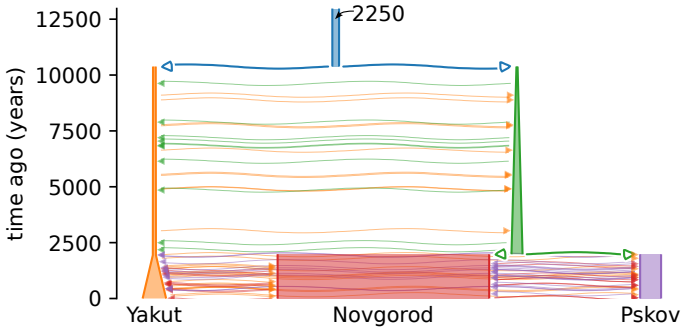
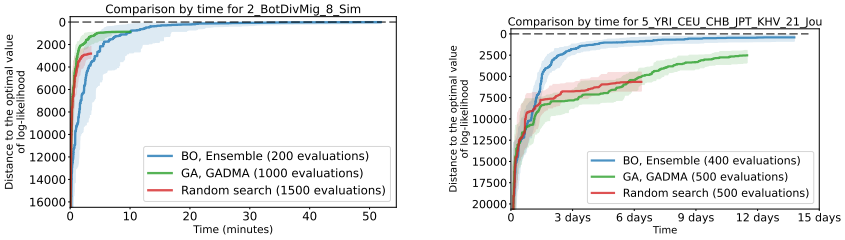


Figure S.8 – Obtained demographic history of three populations of modern humans

Section 2.5 describes the results of conducted experimental studies of the developed method for parameter tuning, that is based on the combination of Bayesian optimization and local search.

Bayesian optimization is compared to the genetic algorithm on a set of datasets (Figure S.9). It was shown that the genetic algorithm (GA) is more efficient in the case of one, two, and three populations. Bayesian optimization (BO) has faster convergence than the genetic algorithm when considering more than three populations. Applying Bayesian optimization allows reducing the time required for parameter tuning by 50-80% which leads to significant acceleration of the process by days and even by weeks.



(а) (б)

Figure S.9 – Convergence over time of parameter tuning methods for (a) two populations, (b) five populations.

Furthermore, the developed method based on the combination of Bayesian optimization and local search is used to tune the parameters of demographic models for four and five populations using real data of modern humans from [55]. The developed method allows obtaining parameters that yielded a higher likelihood than those found in [55] using the Powell's method with restarts. Figure S.10 demonstrates a compari-

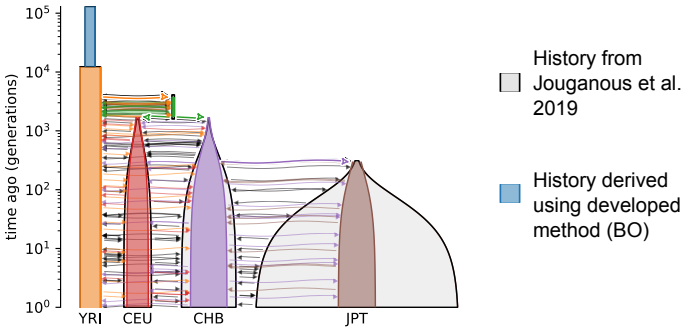


Figure S.10 – Comparison of the demographic history obtained using the developed method (BO) and the demographic history from [55].

son between the resulting demographic history with a higher likelihood and the history obtained in the original study.

The developed method of automatic model selection for demographic inference of one, two, and three populations is described in the **third chapter**, along with the results of its application in combination with the developed method based on a genetic algorithm.

A formal description of the developed method is provided in **Section 3.1**. The method takes the minimum and maximum constraints on the model as input. In the first round, the method constructs a model that satisfies the minimum constraint and performs parameter tuning using the developed genetic algorithm-based method. In each next round, the model is modified, the number of parameters is increased, and a new set of parameters is tuned. The method stops when the model reaches the maximum constraints. Finally, all explored models are compared using the Akaike information criterion, and the best model is selected. The number of time intervals in the model is proposed as the constraint for the models.

Section 3.2 includes experimental studies of the developed automatic model selection method. The demographic history of the «Out of Africa» scenario for genetic data of three modern human populations is obtained, as shown in Figure S.11. The result is consistent with other studies [17, 45, 70]. It has not only a higher likelihood value than the history obtained in [45] using the same data, but also is better according to the Akaike information criterion.

The automatic model selection method was applied to the real data of the Gaboon forest frog (*Scotobleps gabonicus*). Models for three different pairs of populations were constructed in Section 2.4 using manual brute force. The models obtained through the automatic selection method exhibited the best Akaike information criterion values among all the configurations considered for two out of three population pairs. In the case of the third population pair, the obtained model had a worse Akaike information criterion value compared to the best model obtained through manual brute

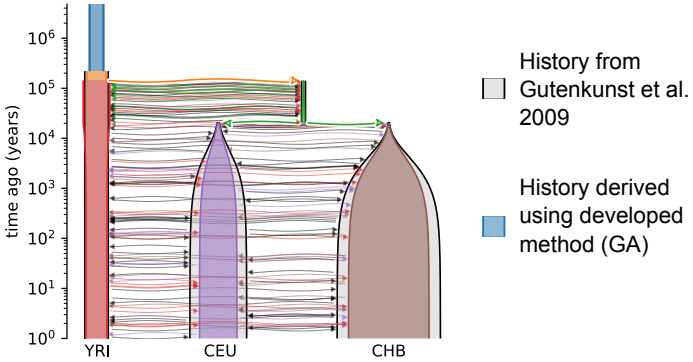


Figure S.11 – Demographic history obtained by the developed method

force. However, the results allow identifying a redundant model parameter, and excluding this parameter from the configuration resulted in the best Akaike information criterion value.

The developed method of automatic selection of extended models is used to infer the demographic history of blue shark populations. The genetic data was not previously analyzed. A sequential approach to inferring the demographic history of two and three populations is developed, resulting in the demographic history shown in Figure S.12. The inferred population sizes are consistent with other studies [57, 58]. Validation of the results by colleagues in the field of zoology suggests that the split between the northern and southern populations of blue shark occurred in connection with paleoclimatic events during the Holocene epoch [71–73].

The description of the software tools that implement the developed methods or are used in this work is provided in the **fourth chapter**.

Section 4.1 contains a description of the GADMA (Global Search Algorithm for Demographic Model Analysis) software framework, which implements the developed

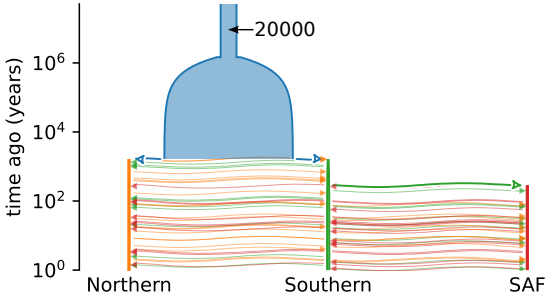


Figure S.12 – Demographic history of three blue shark populations

class of extended models, parameter tuning methods, and automatic mdoel selection method. The structure of the software package is shown in Figure S.13.

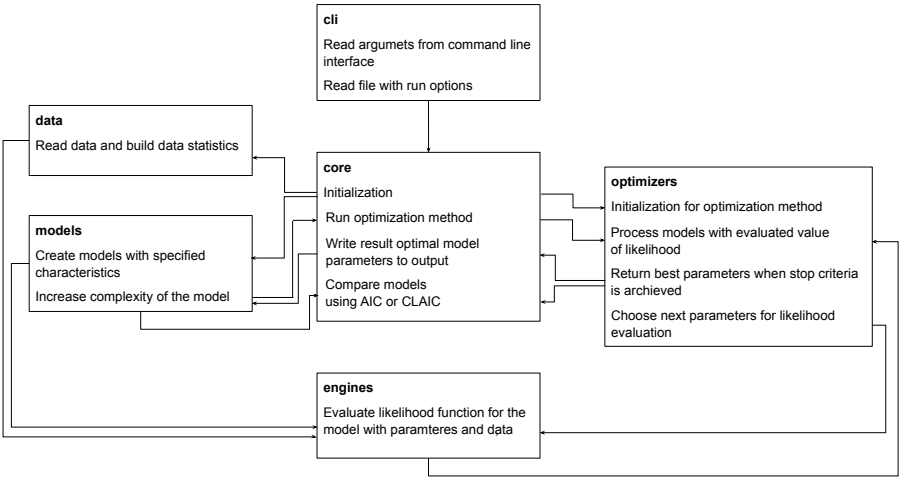


Figure S.13 – Structure of the GADMA software framework.

The extension of the *stdpopsim* and *demes* libraries is described in **Section 4.2**. The *stdpopsim* library provides a catalog of predefined species and their demographic histories for more reliable genetic data simulations. This library was extended and used in the experimental studies. The *demes* software is designed for textual and visual representation of demographic histories. This library is extended with the implementation of linear population size changes and is integrated into the GADMA software framework. All the visualizations of demographic histories in this thesis are obtained using the *demes* library.

To **conclude**, the results of this study are:

- The current state of the research field has been investigated, clarifying the problem and methods for evaluating results;
- The formalization of the problem of building and tuning models of metric trees with functions on edges considered on the example of the demographic inference from genetic data.
- The method for automatic tuning of the models of metric trees with functions on edges based on a combination of global and local optimization methods was developed considered on the example of the demographic inference from genetic data;
- The method for automatic selection of model of metric tree with functions on edges was developed considered on the example of the demographic inference from genetic data;

– The software package that incorporates the developed models and methods for inferring the demographic history of populations from genetic data was designed and implemented;
– experimental studies confirming the effectiveness of the developed models and methods, as well as their applicability for inferring the demographic history of populations from genetic data were carried out, and the results of the experiments were analyzed.

The value of the likelihood function was used to evaluate the quality of demographic history models in this work. Experimental results show that the method of model parameter tuning based on a combination of genetic algorithm and local search allowed to find model parameters that provide a better likelihood value in 88% of cases (37 models out of 42 tested) than the parameters found by existing methods. Using simulated data, the developed method allowed us to find solutions that are 97% closer to the optimum in the case of one population and 66% closer to the optimum in the case of three populations than the solutions obtained by existing methods. The tuning of the hyperparameters of the genetic algorithm allowed to speed up the implementation by 10% on average while maintaining the efficiency of the method.

The effectiveness of the method for tuning model parameters based on Bayesian optimization and local optimization under conditions of a computationally complex target function was confirmed. The developed method made it possible to find parameter values providing a better likelihood value than existing methods for two previously analyzed data of four and five populations. It was shown that Bayesian optimization achieves a solution close to the optimum 50-80% faster than the genetic algorithm in the case of inferring the demographic history of four and five populations.

The method of automatic model selection allows to automatically build and tune models within given configuration constraints. The comparison of models of demographic histories with different numbers of parameters was performed using the Akaike Information Criterion (AIC). Experimental studies showed that in three out of four cases, the method was able to find a model that provided a better AIC value than was previously obtained by manual brute force. In the fourth case, the resulting model allowed to identify a redundant parameter in the configuration and build a nested model that finally provided the best AIC value for the data.

As promising areas of research we can highlight the improvement of the method of automatic model selection in order to find the optimal set of configuration parameters, as well as the development of methods for tuning metric tree models with functions on the edges, which allow not only the tuning of functional parameters, but also the search for the optimal tree structure.

Author's publications on the topic of the thesis

Publications indexed in Web of Science or Scopus

1. ***Noskova E.***, *Ulyantsev V.*, *Koepfli K.-P.*, *O'Brien S. J.*, *Dobrynin P.* GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data // GigaScience. — 2020. — Vol. 9, no. 3. — giaa005. — DOI: 10.1093/gigascience/giaa005.
2. ***Noskova E.***, *Abramov N.*, *Iliutkin S.*, *Sidorin A.*, *Dobrynin P.*, *Ulyantsev V.* GADMA2: more efficient and flexible demographic inference from genetic data // GigaScience. — 2023. — Vol. 12. — giad059. — DOI: 10 . 1093 / gigascience/giad059.
3. ***Noskova E.***, *Borovitskiy V.* Bayesian optimization for demographic inference // G3, Genes | Genomes | Genetics. — 2023. — Vol. 13, no. 7. — DOI: 10.1093/g3journal/jkad080. — jkad080.
4. *Zhernakova D. V.*, ..., *Ulyantsev V.*, ***Noskova E.***, ..., *O'Brien S. J.* Genome-wide sequence analyses of ethnic populations across Russia // Genomics. — 2020. — Vol. 112, no. 1. — Pp. 442–458. — DOI: 10.1016/j.ygeno.2019.03.007.
5. *Nikolic N.*, *Devloo-Delva F.*, *Bailleul D.*, ***Noskova E.***, ..., *Arnaud-Haond S.* Stepping up to genome scan allows stock differentiation in the worldwide distributed blue shark Prionace glauca // Molecular Ecology. — 2023. — Vol. 32, no. 5. — Pp. 1000–1019. — DOI: 10.1111/mec.16822.
6. *Adrion J. R.*, ..., ***Noskova E.***, ..., *Kern A. D.* A community-maintained standard library of population genetic models // eLife. — 2020. — Vol. 9. — e54967. — DOI: 10.7554/eLife.54967.
7. *Lauterbur M. E.*, ..., ***Noskova E.***, ..., *Gronau I.* Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations // eLife / ed. by Z. Gao, M. Przeworski. — 2023. — June. — Vol. 12. — DOI: 10.7554/eLife.84874.
8. *Gower G.*, *Ragsdale A. P.*, *Bisschop G.*, *Gutenkunst R. N.*, *Hartfield M.*, ***Noskova E.***, *Schiffels S.*, *Struck T. J.*, *Kelleher J.*, *Thornton K. R.* Demes: a standard format for demographic models // Genetics. — 2022. — Vol. 222, no. 3. — DOI: 10.1093/genetics/iyac131. — iyac131.

Введение

Актуальность темы исследования. Модели метрических деревьев с функциями на ребрах применяются для анализа и прогнозирования различных явлений реального мира, например, процессов, представимых в виде динамических систем с переменной структурой [9, 10]. Под *метрическими деревьями* (metric trees) понимают граф, являющийся деревом, где каждому ребру поставлен в соответствие интервал. В общем виде, метрические графы с функциями на ребрах нашли широкое применение, например, в виде квантовых графов [11], которые используются в физике при изучении квантового хаоса [12], волноводов [13] и фотонных кристаллов [14].

Построение модели представляет собой набор действий, направленных на выбор конфигурации, определение параметров модели и настройку их значений с целью достижения высокого соответствия результатов моделирования данным натурного эксперимента. На различных этапах построения модели часто требуются экспертные данные или предположения об исследуемом объекте. Эти данные могут быть неточными, ограниченными или неизвестными, что может негативно сказаться на точности и адекватности модели. Методы автоматизированного построения позволяют уменьшить вероятность человеческих ошибок при выборе модели и настройке ее параметров.

При работе с моделями метрических деревьев с функциями на ребрах прибегают к участию предметных специалистов. В первую очередь экспертные данные используются для определения свойств функций на ребрах дерева. Эта информация позволяет установить *конфигурацию* модели, где каждая функция на дереве принадлежит заданному семейству и характеризуется функциональными параметрами, доступными для настройки. В условиях отсутствия экспертных данных или для минимизации влияния специалиста на получение результата приходится рассматривать множество всех возможных моделей, отличающихся типами функций и функциональными параметрами. Например, при построении моделей демографических историй для каждой популяции в качестве динамики изменения численности обычно рассматриваются кусочно-заданные функции, состоящие из функций трех наиболее популярных типов: константная, линейная и экспоненциальная. Такой перебор конфигураций приводит к увеличению временных затрат при построении модели, тем больших, чем больше допустимых типов функций. Дополнительно, требуется следить за сложностью модели, числом ее параметров и переобучением.

Методы для настройки параметров моделей также могут быть ограничены в степени автоматизации и требовать экспертных данных. Например, при использовании методов локального поиска требуются вовлечение специалиста для определения начальных значений параметров, и эффективность настройки зависит от этого выбора.

Таким образом, при моделировании явлений реального мира в виде метрического дерева с функциями на ребрах *актуальна* разработка специализированных моделей и методов для автоматического построения и настройки моделей с целью минимизации влияния экспертных данных на результат моделирования, что рассматривается в данной диссертации на примере задачи вывода демографических историй по генетическим данным.

Популяция — это группа особей одного вида, живущих на одной территории. *Демографическая история популяций* — это исторический процесс их развития и эволюции, который включает в себя такие явления, как изменения численности популяций, разделения популяций, миграция и отбор. Демографические истории используются для датирования исторических событий, не оставивших письменных свидетельств [15, 16], а также играют важную роль в области консервативной генетики [5] и даже в медицине [17].

Различные статистические и алгоритмические методы позволяют строить модели демографических историй в виде метрических деревьев с функциями на ребрах и настраивать их непрерывные параметры по генетическим данным. В случае демографических историй, метрическое дерево является деревом, которое определяет разделение популяций, а функции на ребрах — динамиками изменения численности популяций. В качестве динамик рассматривают кусочно-заданные функции, состоящие из функций трех наиболее популярных типов: константная, линейная и экспоненциальная. При построении моделей требуется определить число временных интервалов, а также тип динамики для каждой кусочно-заданной функции.

Вовлечение специалиста также происходит и на этапе настройки параметров моделей демографической истории популяций, для чего используются комбинация методов численного моделирования и методов оптимизации. Методы численного моделирования используются для вычисления функции правдоподобия, которая позволяет оценить степень соответствия модели генетическим данным. Для поиска параметров, обеспечивающих максимальное значение правдоподобия, используются методы локальной оптимизации. Именно эти методы ограничены в степени автоматизации: они требуют экспертных данных для определения начальных значений параметров, а их эффективность зависит от этого выбора.

Задача вывода демографической истории популяций дополнительно усложняется необходимостью реализации пользователем программного кода модели и алгоритма вывода ее параметров. Методы численного моделирования, используемые существующими решениями, имеют разные возможности и стабильность работы, и пользователь может применить несколько из них для сравнения результатов. Однако при применении различных программных решений одновременно, пользователь сталкивается с необходимостью задавать одни и те же модели с использованием разных интерфейсов.

Таким образом, развитие методов автоматического построения и настройки метрических деревьев с функциями на ребрах приведет к минимизации

влияния экспертных данных, и, следовательно, к повышению качества моделирования явлений реального мира по данным натурного эксперимента.

Степень разработки проблемы. Модели графов исследуются и применяются для решения широкого круга задач. В работах А.М. Райгородского [18, 19] приведены описания и примеры применения моделей случайных графов. Графовые вероятностные модели, такие как байесовские сети, обширно представлены в работах И. Бена-Гала [20] для моделирования индустриальных систем [21], классификации [22] или идентификации сайтов связывания транскрипционных факторов [23]. Л. Кларк и Д. Прегибон [24] описали примеры применения моделей, основанных на деревьях, к которым относятся, например, решающие деревья [25].

Теория метрических графов была сформирована работами В.Г. Болтянского [26], П.С. Солтана [26, 27] и А. Дресса [28]. Свойства метрических деревьев и метрических пространств, порожденных ими, были изучены А. Дрессом [28], Б. Бунеманом [29] и Д. Олдосом [10, 30, 31]. В работах А.С. Матвеева и С.И. Матвеева [32–34] метрические графы были применены при построении координационных моделей для интеллектуальной навигации.

Разработкой моделей, приближающих неявные функции, также активно занимаются многие ученые. Наиболее широкое применение, описанное в работах Л. Фармейра [35] и Р. Сни [36], эти модели получили для решения задач регрессии. При использовании моделей кусочно-заданных функций обычно фиксируют общий вид формирующих функций, например, строят кусочно-постоянные [37, 38], кусочно-линейные [39] или кусочно-экспоненциальные [40] модели. Число точек смены функции, а также их положение являются неизвестными характеристиками моделей кусочно-заданных функций. В работах [41, 42] рассмотрены методы автоматического построения таких моделей для решения задачи кусочно-заданной регрессии, где число точек смены функции определяется с использованием байесовского информационного критерия (BIC) и информационного критерия Акаике (AIC) [43] соответственно.

Модели метрических деревьев с функциями на графах являются комбинацией моделей метрических деревьев и функциональных моделей на ребрах. Квантовые графы, которые являются метрическими графами с дифференциальными операторами на ребрах, и их приложения подробно рассмотрены в работах Г. Берколайко [11, 44]. Метрические деревья с функциями на ребрах используются для моделирования демографических историй популяций в работах Р. Гутенкунста [45], Д. Камма [46], А. Рэгсдейла и С. Гравеля [47, 48]. Однако методы, представленные в этих работах, предполагают, что пользователь определяет и фиксирует общий вид кусочно-заданной функции на ребрах дерева, а также задает начальные значения параметров настройки параметров методами локальной оптимизации. В работах Д. Портика [49, 50] и Р. Гутенкунста [51] были представлены методы глобальной оптимизации для настройки параметров моделей демографических историй, которые минимизируют, однако все еще требуют во-

влечение пользователя. Общее применение методов численной оптимизации для решения задач представлено в классической работе Б.Т. Поляка [52], а описание современных методов глобальной оптимизации в работе [53].

На момент начала исследований автором (в 2017 году) не существовало метода автоматического построения и настройки моделей метрических деревьев с функциями на ребрах. К концу диссертационного исследования появилось первое альтернативное решение для метода автоматического перебора моделей на примере задачи вывода демографических историй [54]. Однако метод позволяет анализировать модели, определенные специфичным каталогом и только для вывода демографической истории *двух* популяций, а выбор наилучшей модели происходит в предположении независимости данных, что не всегда является корректным.

Целью настоящей диссертации является повышение качества[1] компьютерного моделирования явлений реального мира за счет автоматизации построения и настройки моделей метрических деревьев с функциями на ребрах.

Для решения цели в диссертации решаются следующие **задачи**:
- исследование текущего состояния предметной области, уточнение задачи и способов оценки результатов;
- формализация постановки задачи построения и настройки моделей метрического дерева с функциями на ребрах;
- разработка метода автоматической настройки моделей метрического дерева с функциями на ребрах на основе комбинации методов глобальной и локальной оптимизации;
- разработка метода автоматического перебора моделей метрического дерева с кусочно-заданными функциями на ребрах;
- проектирование и реализация программного комплекса, включающего разработанные модели и методы для вывода демографической истории популяций по генетическим данным;
- проведение экспериментальных исследований, подтверждающих эффективность разработанных моделей и методов, а также их применимость для вывода демографической истории популяций по генетическим данным, анализ результатов экспериментов.

Научная новизна диссертации состоит в том, что: (1) разработаны методы на основе комбинации методов глобальной и локальной оптимизации для настройки параметров заданной модели метрического дерева с функциями на ребрах; (2) разработан метод автоматического перебора моделей метрического

[1] Качество моделей в данной работе определяется степенью соответствия настроенной модели данным натурного эксперимента. В случае задачи вывода демографических историй популяций качество определяется значением функции правдоподобия, полученным численными методами за фиксированное время настройки модели.

дерева с кусочно-заданными функциями на ребрах, не требующий вовлечения эксперта на этапе выбора параметров рассматриваемых моделей.

Теоретическая значимость работы определяется расширением классической постановки задачи настройки модели метрического дерева с функциями на ребрах не только как задачи настройки параметров заданной модели, но и как задачи выбора самой модели путем автоматического перебора. Полученные методы моделирования и настройки применимы для произвольных моделей метрического дерева с функциями на ребрах. Более того, разработанные методы оптимизации могут быть использованы или адаптированы для задач поиска оптимальных параметров в других научных областях.

Практическую значимость работы определяют:

- расширение научно-практического инструментария специалистов-биоинформатиков методами и алгоритмами для вывода демографических историй популяций;
- открытый программный код разработанного программного комплекса GADMA, который доступен к переиспользованию по адресу `https://github.com/ctlab/GADMA`;
- применимость разработанных методов для анализа генетических данных;
- внедрение разработанного метода на основе генетического алгоритма в стороннее программное решение [54].

На защиту выносятся положения, обладающие научной новизной:

1. Метод моделирования и настройки параметров моделей метрических деревьев с функциями на ребрах по данным натурного эксперимента, содержащий модели с непрерывными функциональными параметрами, отличающийся тем, что с целью автоматической настройки без привлечения экспертных данных в нем используются модели с дискретными парамерами, определяющими семейства функций, а также методы глобальной оптимизации — генетический алгоритм и байесовская оптимизация, и реализующий его комплекс программ.
2. Метод автоматического перебора моделей метрических деревьев с функциями на ребрах с разным числом параметров и настройки этих параметров по данным натурного эксперимента, содержащий сравнение моделей с использованием информационного критерия Акаике, отличающийся тем, что с целью повышения уровня автоматизации и обеспечения возможности настраивать не только параметры модели, но и саму модель, он включает метод увеличения числа временных интервалов для кусочно-заданных функций на ребрах дерева, а также реализующий его комплекс программ.

Методы исследования. В работе использованы методы оптимизации, численные методы, методы теории вероятности и математической статистики, методы машинного обучения и методы проведения экспериментальных исследований.

Достоверность научных результатов обусловлена корректным использованием методов, обоснованием постановки задач, экспериментальными исследованиями, покрывающими разработанные технологии и алгоритмы. Демографические истории, полученные разработанными методами на проверяемых симулированных данных, согласуются с исходными историями, используемыми для моделирования. Результаты, полученные на реальных данных, согласуются с опубликованными ранее исследованиями [45, 55–59].

Соответствие паспорту специальности. Полученные научные результаты соответствуют следующим пунктам паспорта специальности 1.2.2 — «Математическое моделирование, численные методы и комплексы программ (технические науки)».

Пункт 2 паспорта специальности «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий». Были разработаны, обоснованы и протестированы методы настройки параметров моделей метрического дерева с функциями на ребрах, основанные на методах численной оптимизации.

Пункт 4 паспорта специальности «Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели». В диссертационном исследовании представлены методы для построения моделей метрического дерева с функциями на ребрах по данным натурного эксперимента с целью анализа явлений реального мира.

Апробация результатов работы

Основные результаты работы были представлены на следующих конференциях:

- Международный конгресс «VII съезд Вавиловского общества генетиков и селекционеров, посвященный 100-летию кафедры генетики СПбГУ, и ассоциированные симпозиумы», 2019, Санкт-Петербург, Россия;
- Moscow Conference on Computational Molecular Biology, 2019, Москва, Россия;
- Probabilistic Modeling in Genomics, 2019, Осуа, Франция;
- Probabilistic Modeling in Genomics, 2021, онлайн;
- Moscow Conference on Computational Molecular Biology, 2021, Москва, Россия;
- Вероятностные методы в анализе: пространства голоморфных функций, 2021, Сочи, Россия;

– LI Научная и учебно-методическая конференция Университета ИТМО, 2022, Университет ИТМО, Санкт-Петербург, Россия;
– Probabilistic Modeling in Genomics, 2022, Окфорд, Великобритация;
– XI Конгресс молодых ученых, 2022, Университет ИТМО, Санкт-Петербург, Россия;
– Conservation Genomics at the Population Level, 2022, Кембридж, Великобритания;
– Probabilistic Modeling in Genomics, 2023, Колд Спринг Харбор, США;
– XII Конгресс молодых ученых, 2023, Университет ИТМО, Санкт-Петербург, Россия;
– Society for Molecular Biology and Evolution Meeting (SMBE23), 2023, Феррара, Италия.

Награды

– Бронзовая награда в номинации 17th Human-Competitive Awards на онлайн конференции The Genetic and Evolutionary Computation Conference (GECCO) в 2020 году.
– Победитель конкурсной программы поддержки исследовательских проектов System Biology Fellowship от Сколковского института науки и технологий по проекту «Computational methods for unsupervised demographic inference of multiple populations from genomic data» в 2021 году. Число победителей — пять на всю страну в год.

Публикации

По результатам, представленным в диссертации, было опубликовано восемь статей в рецензируемых научных журналах, входящих в международные реферативные базы данных и системы цитирования Scopus и Web of Science.

Личный вклад автора

1. В публикации [1] Noskova E. — разработка и реализация генетического алгоритма и метода автоматического перебора моделей демографической истории, проведение экспериментальных исследований (80%); Ulyantsev V. — рекомендации по постановке задачи, выбору и обоснованию теоретических основ научного исследования (10%); Koepfli K.P., O'Brien S.J. — консультирование при проведении экспериментальных исследований и написании статей (5%); Dobrynin P. — рекомендации по постановке задачи (5%).
2. В публикации [2] Noskova E. — разработка и реализация методов, программного обеспечения для вывода демографической истории популяций по генетическим данным, проведение экспериментальных исследований (85%); Abramov N., Iliutkin S., Sidorin A. — разработка программного обеспечения (10%); Dobrynin P., Ulyantsev V. — рекомендации по

постановке задач, выбору и обоснованию теоретических основ научного исследования (5%).

3. В публикации [3] Noskova E. — разработка и реализация метода байесовской оптимизации для вывода демографической истории популяций по генетическим данным, проведение экспериментальных исследований (90%); Borovitskiy V. — рекомендации по постановке задач, выбору и обоснованию теоретических основ научного исследования (10%).
4. В публикации [4] Noskova E. — вывод демографической истории трех популяций современного человека (10%); Ulyantsev V. — рекомендации по постановке задачи (5%); остальные соавторы — сбор и анализ генетических данных (85%).
5. В публикации [5] Noskova E. — вывод демографической истории двух и трех популяций голубых акул (10%); остальные соавторы — сбор и анализ генетических данных (90%).
6. В публикации [6] Noskova E. — разработка и тестирование программного обеспечения для симулирования генетических данных по демографической истории популяций (5%); остальные соавторы — разработка и тестирование программного обеспечения, проведение экспериментальных исследований (95%).
7. В публикации [7] Noskova E. — реализация демографических историй популяций в программном обеспечении для симулирования генетических данных по демографической истории популяций (5%); остальные соавторы — разработка программного обеспечения (95%).
8. В публикации [8] Noskova E. — разработка программного обеспечения для представления демографической истории популяций (5%); остальные соавторы — разработка программного обеспечения (95%).

Структура диссертационной работы

Диссертация состоит из введения, четырех глав, заключения и приложения. Полный объём диссертации составляет 396 страниц, включая 120 рисунков, 16 таблиц и восемь листингов. Список литературы содержит 171 наименований.

Финансирование

Автор признателен компании JetBrains Research за финансовую поддержку работы в 2017–2021 годах. Работа выполнена также благодаря финансированию от проекта 5-100, программы НИР МиА университета ИТМО и благодаря гранту System Biology Fellowship от Сколковского института науки и технологий и компании Philip Morris International.

Глава 1. Обзор предметной области

Развитие методов секвенирования привело в наши дни к накоплению большого объема генетических данных. Вместе с тем происходило развитие методов для анализа этих данных в области биоинформатики. *Популяционная генетика* является важной областью генетики, изучающей изменение генетического состава популяций и их эволюцию. Она рассматривает такие важные понятия, как генетические вариации, частоты аллелей и генотипов, дрейф генов и отбор. Благодаря пониманию механизмов, определяющих генетическое разнообразие и изменения в популяциях, популяционная генетика может пролить свет на широкий спектр биологических явлений от происхождения видов до распространения инфекционных заболеваний.

Одной из важных задач популяционной генетики является задача вывода демографической истории популяций — истории их эволюции, которая включает в себя информацию о численности популяций в прошлом, времени разделений и темпы миграций. История эволюции сохраняется в геномах особей, и ее можно реконструировать, используя различные статистические и алгоритмические методы.

В разделе 1.1 приведено описание демографической истории популяций. В разделе приведены различные примеры демографических историй, описано их визуальное изображение, используемое в данной работе, а также дано формальное определение объекта демографической истории с математической точки зрения.

Раздел 1.2 описывает общую схему существующих методов вывода демографической истории популяций по генетическим данным. Эти методы используют параметрические модели демографической истории для сужения области поиска, а также методы настройки параметров этих моделей по генетическим данным. В разделе приведено определение и примеры параметрических моделей. Затем описан процесс настройки параметров моделей, который заключается в поиске значений параметров, дающих максимальное значение правдоподобия для генетических данных. Приведен список существующих программных решений и краткое описание их методов. В конце раздела сформулирована задача поиска демографической истории.

В разделе 1.3 описаны основные классы параметрических моделей, которые используются в существующих программных решениях. Раздел включает формальные определения этих моделей, а также примеры их спецификации с использованием существующих библиотек. Более того, приведены основные методы сравнения моделей с разным числом параметров.

Раздел 1.4 включает подробное описание существующих методов вычисления правдоподобия генетических данных при заданной демографической истории популяций. Эти методы являются методами численного имитационного моделирования.

В разделе 1.5 приведен обзор существующих методов оптимизации для настройки параметров моделей демографических историй, которые, в основном, являются методами локальной оптимизации.

Раздел 1.6 включает обзор методов перебора моделей демографической истории популяций. Перебор моделей позволяет получить более надежный результат при решении задачи вывода демографической истории популяций.

1.1. Демографическая история популяций

Вид — это группа организмов, которые могут размножаться между собой и давать потомство, способное также размножаться. Видовое понятие является одним из основных в биологии, так как оно позволяет классифицировать живые организмы на различные таксоны и изучать их в рамках конкретной научной области.

Популяция — это группа организмов одного вида, находящихся в определенной географической области и взаимодействующих друг с другом. Популяции могут быть различных размеров и иметь разную структуру, исходя из того, как организмы этой популяции взаимодействуют между собой и с окружающей средой.

На рисунке 1 представлены две популяции газелей вида *Dama gazelle*. Популяция mhorr обитала на западе Африки, а популяция addra на востоке континента. Из-за различных мест обитания эти две группы особей называются отдельными популяциями. В настоящее время отдельные особи обеих популяций встречаются только в зоопарках и частных коллекциях по всему миру.

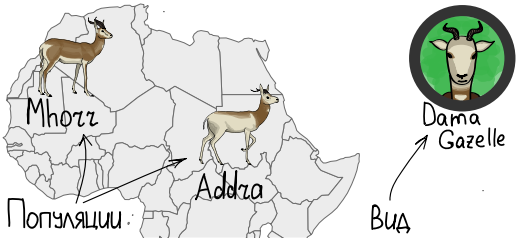


Рисунок 1 – Пример двух популяций газели вида *Dama gazelle*; 1) популяция mhorr, 2) популяция addra

Эффективный размер популяции N_e соответствует числу размножающихся особей в популяции. В более общем смысле N_e — это концепт, используемый для описания того, какой размер имела бы идеальная популяция с таким же уровнем генетического разнообразия, как и реальная популяция. Идеализированные популяции основаны на нереалистичных, но удобных упрощениях, таких как случайное спаривание, одновременное рождение каждого нового поколения,

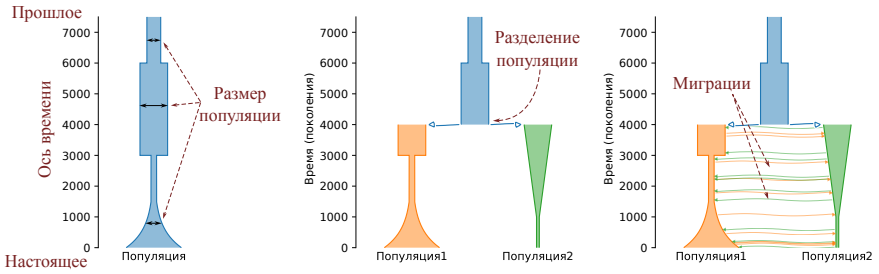
постоянный размер популяции и равное число детей для одного родителя. Эффективный размер популяции N_e обычно меньше, чем фактический размер популяции из-за случайного разброса генетических вариантов между поколениями и влияния генетических дрейфов, мутаций и естественного отбора. Чем меньше эффективный размер популяции, тем выше риск утраты генетического разнообразия и возникновения генетической деградации популяции.

Инбридинг — это получение потомства от спаривания или размножения особей, которые являются близкородственными на генетическом уровне. Инбридинг приводит к увеличению гомозиготности генома, что может увеличить вероятность поражения потомства рецессивными признаками [74]. Коэффициент инбридинга является мерой инбридинга, которая отражает процент позиций генома с генетической информацией, которая была унаследована от одного предка [75]. Чем этот коэффициент выше, тем выше уровень близкородственного скрещивания в популяции.

Неформально говоря, демографическая история популяций — это история развития и эволюции популяций, которая включает в себя информацию о дереве разделения популяций, численности популяций в прошлом, миграциях, отборе и многом другом. Формальное определение демографической истории популяций будет дано далее.

Примеры визуального представления демографических историй показаны на рисунке 2. Демографическую историю можно изображать различными способами. В диссертации используется представление, которое было предложено в работе [8]. На рисунке 2а представлена демографическая история одной популяции. Ось абсцисс соответствует числу поколений в прошлом, ноль — настоящее время. Время в демографических историях измеряется в поколениях, так как в процессе эволюции генетический материал передается от одного поколения к другому. Демографическая история — древовидная структура, она, в частности, задает филогенетическое дерево популяций, которое отображает как популяции разделялись. Однако демографическая история также содержит информацию о численности популяций и миграциях: это отображено шириной веток дерева и стрелками между ними. Под численностью или размером популяции здесь и далее будет пониматься эффективный размер популяции. Ширина раскрашенных областей соответствует размеру популяции в конкретный момент времени, а число стрелок зависит от степени миграций между популяциями.

Более подробно, на рисунке 2а изображена история о том, что размер популяции в далеком прошлом был равен 5000 особей, 6000 поколений назад численность популяции возросла в два раза и оставалась постоянной на протяжении 3000 поколений, за последние 3000 поколений популяция пережила «бутылочное горлышко», когда ее размер составлял всего 2000 особей и экспоненциальный рост в течение последних 1500 поколений до текущего размера в 20000 особей. Рисунок 2б представляет демографическую историю изоляции двух популяций. Для удобства представления популяция-предок до разделения и ее размер изображены синим цветом, а популяции 1 и 2, образованные разделением



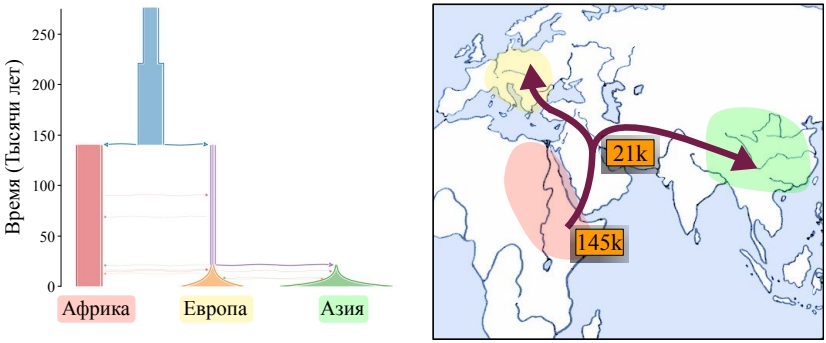
(а) (б) (в)

Рисунок 2 – Примеры визуального представления демографических историй одной и двух популяций

популяции-предка, изображены оранжевым и зеленым цветом соответственно. История называется изоляцией, так как популяции 1 и 2 не имели контактов в виде миграций после образования разделением. Третья демографическая история, изображенная на рисунке 2в, является историей двух популяций с миграциями. Миграция изображена стрелками между областями, соответствующими популяциям, между которыми происходила миграция.

Изучение демографической истории популяций имеет большое значение для понимания биологических процессов [17], в том числе для определения стратегий охраны и восстановления угрожаемых видов. Они дополняют имеющиеся археологические данные об исторических событиях, которые не оставили письменных свидетельств, таких, например, как континентальные миграции популяций человека [15, 16].

Рассмотрим пример того, как демографическая история популяций может дополнять археологические данные (рисунок 3). Происхождение человека в Африке — это теория, согласно которой вид современного человека *Homo sapiens* возник в Африке около 200 тысяч лет назад, а затем распространился по всему миру [56]. По данным археологических исследований первые люди покинули Африку вероятнее всего через территорию современной Саудовской Аравии, и распространились по всему миру. Этот процесс называется «выходом из Африки» и является одним из ключевых событий в истории человечества. Демографическая история популяций позволяет датировать такие события как «выход из Африки», миграция в Европу или Азию. На рисунке 3а представлен пример демографической истории, полученной по генетическим данным для трех популяций современного человека: из Африки, Европы и Азии. Рисунок 3б демонстрирует карту, показывающую примерное перемещение групп людей, согласно археологическим исследованиям. Имея демографическую историю, можно обозначить времена этих перемещений на карте, например, можно утверждать, что «выход из Африки» произошел примерно 145 тысяч лет назад.



(а) (б)

Рисунок 3 – Пример демографической истории трех популяций современного человека и карта перемещения этих популяций, построенная по археологическим данным.

В современной литературе строгого определения демографической истории нет. Демографическая история может быть сколь угодно подробной, например, можно представить, что она описывает все геномы всех особей, которые когда-либо присутствовали в популяции или, более того, включает координаты перемещения этих особей по земному шару. Однако, в современных исследованиях все же обычно не рассматривают настолько подробные объекты. Вместо этого под демографической историей понимают историю разделения популяций, численности в каждый момент времени и темпы миграции. В данной работе исследованы аналогичные объекты. Опишем строгое понятие демографической истории популяций, которое использовано в данной работе.

Пусть имеется вершина-корень, к которой присоединено полное бинарное дерево с P листьями. Такое дерево определяет структуру разделения популяций. Бинарное дерево называется полным, если у каждого узла есть либо два дочерних элемента, либо ноль дочерних элементов. Каждый лист дерева и входящее в него ребро ассоциированы с одной из текущих популяций, а каждый узел дерева со своим входящим ребром ассоциирована с популяциями в прошлом. Например, вершина-родитель двух листьев, соответствующих популяциям 1 и 3, будет соответствовать их общей популяции, которая в какой-то момент в прошлом разделилась и образовала популяции 1 и 3.

Определение 1. Дерево разделений P популяций — дерево $T = \langle V_T, E_T \rangle$ с корнем в вершине v_r такое, что поддерево $T^\star = T \setminus \{v_r\}$ без корня является полным бинарным деревом, в котором множество листьев занумеровано числами от 1 до P.

Пример дерева разделений для четырех популяций изображен на рисунке 4. Листья занумерованы числами от 1 до 4, которые соответствуют номерам популяций.

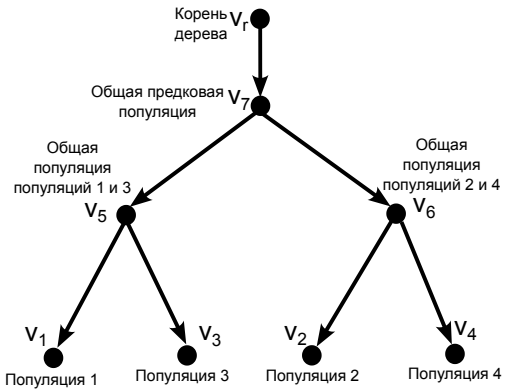


Рисунок 4 – Пример дерева разделений популяций

Рассмотрим дерево разделений P популяций. Каждое ребро графа ассоциировано с какой-то популяцией в настоящем или в прошлом. Пусть на каждом ребре e задано множество из времени образования $t_e \in \mathbb{R}_+$ соответсвующей популяции, времени ее разделения $t_e^d \in \mathbb{R}_+$ и функции изменения численности популяции $g(t) : [t_e, t_e^d] \to \mathbb{R}_+$. Время разделения популяции — это время, когда она перестала существовать. Заметим, что для каждой внутренней вершины дерева время образования, ассоциированное с исходящим ребром, должно совпадать со временем разделения, ассоциированным с входящим ребром. Время отображается в поколениях в прошлом, поэтому $t_e^d < t_e$. На функцию изменения численности никаких ограничений не накладывается, она не обязана быть непрерывной. Пример дерева разделений с заданными временами образования и разделения популяций, а также функциями изменения численности изображен на рисунке 5.

Обратим внимание, что такое дерево является метрическим графом. *Метрическим графом* называется граф, каждому ребру которого соответствует интервал. Каждому ребру e рассмотренного дерева ассоциирован отрезок времени $[t_e^d, t_e]$, в течении которого существовала популяция, следовательно, это метрический граф. Ребро, исходящее из корня, является открытым — ему соответствует бесконечный интервал $[t_e^d, \infty]$.

Рассмотрим множество миграций между популяциями, которые бывают двух типов: единичные и непрерывные. Единичная миграция определяется как событие перемещения группы особей из одной популяции в другую в определенный момент времени в прошлом. Если особи перемещаются непрерывно на

протяжении какого-то интервала времени, то такая миграция называется непрерывной. Та популяция, из которой происходит миграция, называется *популяцией-истоком*, а та, в которую происходит миграция особей — *популяцией-стоком*.

Определение 2. Единичная миграция $\mathfrak{M}_q \in \widetilde{\mathscr{M}}$ — это четверка $\langle e_1, e_2, t, m\rangle$, где e_1 — популяция-исток, e_2 — популяция-сток, t — время единичной миграции, m — интенсивность, равная числу особей, которое переместилось.

Определение 3. Непрерывная миграция $\mathfrak{M}_q \in \mathscr{M}$ — это пятерка $\langle e_1, e_2, t^s, t^e, m\rangle$, где e_1 — популяция-исток, e_2 — популяция-сток, t^s — время начала непрерывной миграции, $t^e < t^s$ — время окончания непрерывной миграции, m — интенсивность, равная среднему числу особей, которое перемещается каждое поколение.

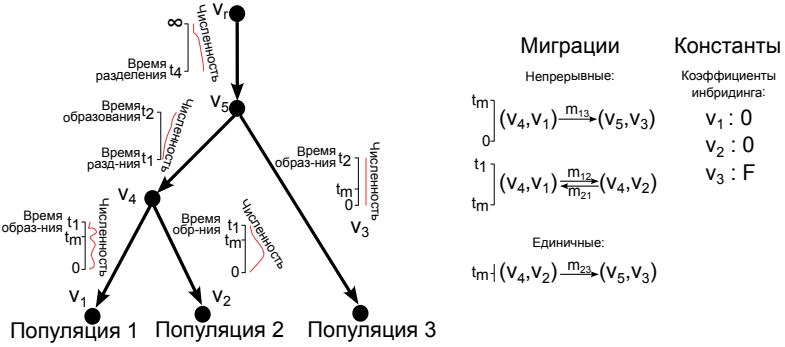


Рисунок 5 – Пример демографической истории популяций как дерева разделений с заданными временами и функциями численности, набора миграций и дополнительных констант

Определение 4. Демографическая история $\mathcal{D}$ для P популяций — это четверка $\langle T, \mathfrak{G}, \mathfrak{M}, \mathcal{C}\rangle$, где $T = \langle V, E\rangle$ — дерево разделения популяций, $\mathfrak{G} : E \to \mathbb{R}_+ \times \mathbb{R}_+ \times \mathcal{F}_{\mathbb{R}_+\to\mathbb{R}_+}$ — отображение, которое для каждого ребра e ставит в соответствие множество $\langle t_e, t_e^d, g(t)\rangle$, где t_e и t_e^d — время образования и разделения популяции соответственно, а функция $g(t) : [t_e, t_e^d] \to \mathbb{R}_+$ определяет численность популяции в каждый момент ее существования, а $\mathfrak{M} = \{\mathfrak{M}_q\}$, $\mathfrak{M}_q \in \widetilde{\mathscr{M}} \cup \mathscr{M}$ — набор единичных и непрерывных миграций, $\mathcal{C} : V \to \mathbb{R}^C$ — набор дополнительных констант для популяций. На отображение $\mathfrak{G}$ накладывается следующее ограничение: для любой внутренней вершины v время образования исходящего ребра $t_{v,\text{child}(v))}$ равно времени разделения входящего ребра $t^d_{(\text{parent}(v))}$. Для любой единичной миграции $\mathfrak{M}_q = \langle e_1, e_2, t, m\rangle$ выполняется: $e_1, e_2 \in E$ и $t \in [t_{e_1}, t^d_{e_1}] \cap [t_{e_2}, t^d_{e_2}]$. Для любой непрерывной миграции $\mathfrak{M}_q = \langle e_1, e_2, t^s, t^e, m\rangle$ выполняется: $e_1, e_2 \in E$ и $t^s, t^e \in [t_{e_1}, t^d_{e_1}] \cap [t_{e_2}, t^d_{e_2}]$.

В качестве дополнительных констант для популяций в данной работе будут рассмотрены коэффициенты инбридинга $\mathcal{C}(v_i) = F_i,\ i = 1, \ldots, P$, заданные для листьев $\{v_i\}_{i=1}^{P}$ дерева T.

Рисунок 5 демонстрирует пример демографической истории популяций как набора из дерева разделений, отображения на его вершинах и множества миграций.

1.2. Методы вывода демографической истории популяций по генетическим данным

Основы для методов вывода демографической истории популяций заложил японский биолог М. Кимура в своих работах 1962 [76, 77] и 1969 [78] годах, а также ученые В. Хилл и А. Робертсон в работах 1966 [79] и 1968 [80] годов. Эти методы стали активно развивать в конце XX века для вывода отдельных характеристик демографических историй, например, степени роста численности популяции [81]. В начале XXI века стали появляться методы для вывода более сложных демографических историй с большим числом характеристик таких, как численность, время разделения и темпы миграции [45, 55, 82].

В общем случае, задача вывода демографической истории популяций по генетическим данным состоит в поиске наилучшей демографической истории из всего множества возможных историй. Для определения наилучшей демографической истории используется значение правдоподобия, которое определяет насколько хорошо история описывает генетические данные. Таким образом, задача состоит в поиске демографической истории с наибольшим значением правдоподобия для данных. Напомним, что демографическая история включает в себя функции изменения численности, и поэтому, поиск по всему пространству возможных историй — это, в том числе, и поиск по пространству функций, что вызывает трудности.

В результате для упрощения задачи поиска существующие методы ограничивают пространство рассматриваемых демографических историй, используя параметрические модели, или просто модели. Приведем два эквивалентных определения параметрической модели демографической истории популяций.

Определение 5. Параметрическая модель демографической истории — это множество $\{\mathcal{D}_\theta\}_{\theta \in \Theta}$ демографических историй популяций, параметризованное набором параметров θ.

Определение 6. Параметрическая модель демографической истории — отображение $\mathcal{M} : \Theta \to \{\mathcal{D}\}$, которое любому набору значений параметров θ модели ставит в соответствие демографическую историю популяций $\mathcal{D}_\theta$.

Приведем пример модели демографической истории. Пусть задана одна популяция и известно, что ее численность всегда была постоянна. Все демографические истории одной популяции имеют одинаковое дерево разделений, состоящее из одной вершины. Таким образом, все демографические истории одной популяции отличаются только функцией изменения численности.

Зная, что численность популяции постоянна, можно задать модель M_1 с одним параметром $\theta = (\theta_1)$, который будет соответствовать константному размеру популяции. Таким образом, модель будет отображать пространство параметров в демографические популяции, у которой функция изменения численности будет равна $g(t) = \theta_1$. Пример описанной модели M_1 представлен на рисунке 6. На рисунке 6а показано отображение из пространства параметров в множество демографических историй. Рисунок 6б демонстрирует визуальное изображение модели демографической истории, которое будет использовано в данной работе. На рисунке представлено изображение демографической истории и схематично указано какую характеристику демографической истории регулирует параметр θ_1. Для того чтобы продемонстрировать что это не одна демографическая история, а множество, изображение модели дополнительно содержит пунктирные линии.

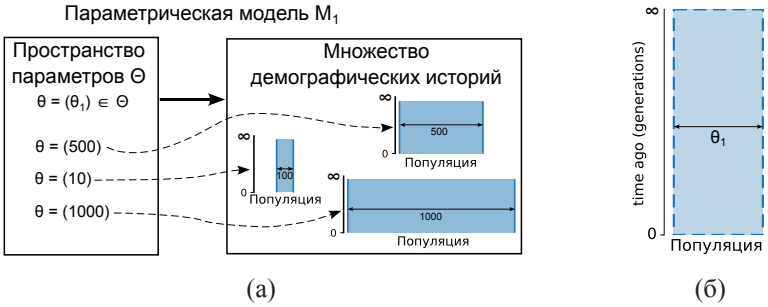


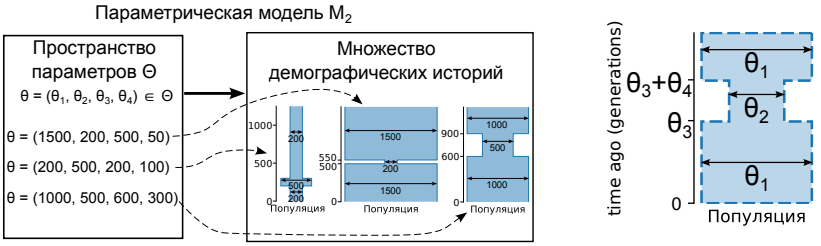
Рисунок 6 – Пример модели M_1 демографической истории одной популяции с одним параметром

Одни и те же параметры модели могут задавать разные характеристики демографических историй. Например, на рисунке 7 представлена модель M_2, имеющая четыре параметра. Эта модель представляет множество демографических историй одной популяции, у которых функция изменения численности определяется следующим образом:

$$g(t) = \begin{cases} \theta_1, & \text{если } t \leqslant \theta_3, \\ \theta_2, & \text{если } \theta_3 \leqslant t \leqslant \theta_3 + \theta_4, \\ \theta_2, & \text{если } t \geqslant \theta_3 + \theta_4. \end{cases}$$

Рисунок 7а изображает модель M_2, как отображение из пространства параметров в множество демографических историй. Рисунок 7б демонстрирует визуальное изображение модели.

Модель M_2 описывает демографическую историю с тремя периодами константной численности, при этом параметр θ_1 задает и численность до момента времени $\theta_3 + \theta_4$, и после времени θ_3.



(а) (б)

Рисунок 7 – Пример модели M_2 демографической истории одной популяции с четырьмя параметрами

Заметим, что множество демографических историй, определенных моделью M_1, вложено в множество историй, определенное моделью M_2 (рисунок 8). Тогда будем говорить, что модель M_1 является вложенной в модель M_2.

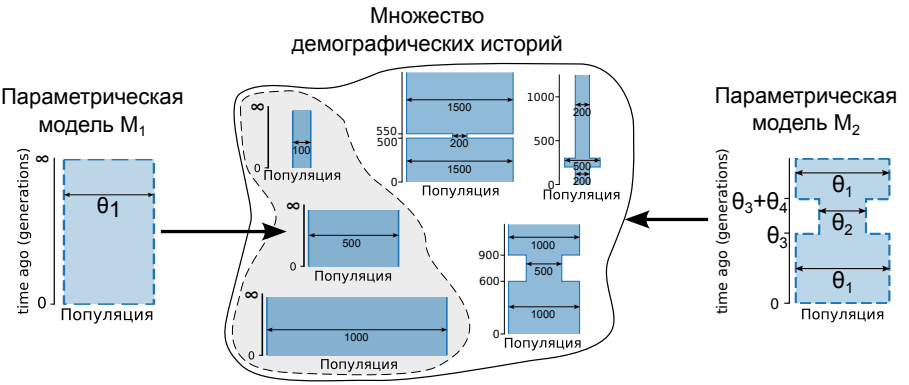
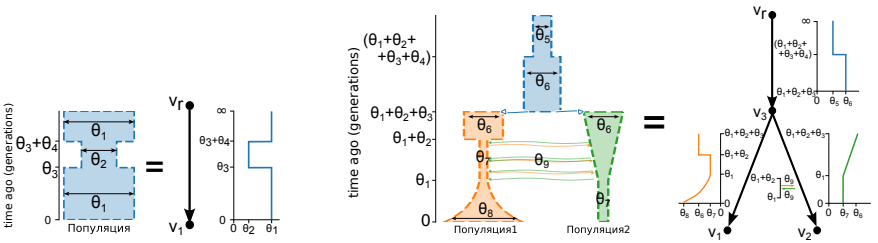


Рисунок 8 – Пример вложенных моделей: модель M_1 вложена в модель M_2

Модели не всегда вложены друг в друга, иногда их множества демографических историй просто пересекаются. Как следствие, одна и та же демографическая история может соответствовать разным моделям.

Рассмотренное определение параметрических моделей является общим. Однако обычно параметры моделей отображают определенные характеристики объектов моделирования. Поэтому в данной работе будут рассмотрены модели демографических историй определенного класса математических объектов — **метрических деревьев с функциями на ребрах**. Определение демографических историй, приведенное в разделе 1.1 позволяет представить их модели в виде метрических деревьев с функциями на ребрах, которые определяют изменение численности популяций во времени, темпы непрерывных и единичных миграций. На рисунке 9 представлены метрические графы с функциями на ребрах как модели демографических историй. На рисунке 9а представлена модель демографической истории одной популяции, изображенной ранее на рисунке 7. На рисунке 9б представлена модель демографической истории двух популяций.



(а) (б)

Рисунок 9 – Метрические деревья с функциями на ребрах как модели демографических историй а) одной популяции, б) двух популяций

Вернемся к задаче вывода демографической истории популяций по генетическим данным, которая состоит в поиске демографической истории с максимальным значением правдоподобия для данных. Используя описанные параметрические модели, пользователь может ограничить пространство поиска, а также использовать методы оптимизации для перебора значений параметров, а следовательно, и демографических историй для выбора наилучшей.

Процесс поиска демографической истории популяций по генетическим данным с использованием параметрических моделей выглядит следующим образом (рисунок 10):

- ограничение пространства рассматриваемых демографических историй путем задания параметрической модели;
- настройка значений параметров модели по генетическим данным.

Пусть имеются генетические данные P популяций, представленные в виде последовательностей геномов или частей геномов длины G. Обозначим данные как $\mathfrak{D} = \{\mathfrak{D}_j = \{\mathfrak{d}_j^i\}_{i=1}^{n_j}\}_{j=1}^{P}$, где n_j равно числу представленных особей из популяции j, а $\mathfrak{d}_j^i \in \{A, T, G, C\}^G$ — генетическая последовательность особи i из популяции j. Пусть также задано семейство моделей демографической истории

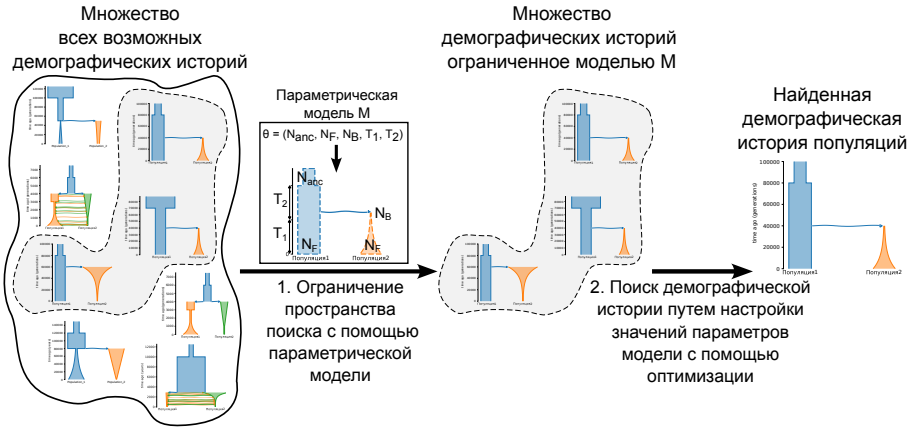


Рисунок 10 – Общая схема поиска демографической истории популяций с использованием параметрической модели

параметризованное вектором параметров $\theta \in \Theta$, где Θ — множество всех возможных значений параметров, как непрерывных, так и дискретных. Обозначим такое семейство моделей как $\{\mathcal{M}(\theta)\},\ \theta \in \Theta$.

Рассмотрим функцию $f_{\mathcal{M}}(\theta, \mathfrak{D})$, которая принимает на вход параметры θ модели $\mathcal{M}$, а также генетические данные $\mathfrak{D}$ и возвращает вероятность наблюдать данные $\mathfrak{D}$ при условии модели $\mathcal{M}$ с заданными параметрами θ. Функция $f_{\mathcal{M}}(\theta, \mathfrak{D})$ называется функцией правдоподобия и может быть определена различными способами. Например, если рассмотреть метод аппроксимации диффузией, используемый в ∂a∂i, то функция f симулирует ожидаемый аллель-частотный спектр по заданной модели $\mathcal{M}$ демографической истории с параметрами θ и вычисляет значение правдоподобия симулированного спектра и наблюдаемого аллель-частотного спектра, построенного из генетических данных $\mathfrak{D}$. Более подробное описание методов вычисления функции $f_{\mathcal{M}}$ представлены в разделе 1.4.3.

Сформулируем задачу вывода демографической истории по генетическим данным следующим образом.

На **вход** подается:

– Генетические данные $\mathfrak{D} = \{\mathfrak{D}_j = \{\mathfrak{d}_j^i\}_{i=1}^{n_j}\}_{j=1}^{P},\ \mathfrak{d}_j^i \in \{A, T, G, C\}^G$ для P популяций,
– Параметризованная модель демографической истории $\{\mathcal{M}(\theta)\},\ \theta \in \Theta$,
– Множество Θ всех возможных значений параметров.

На **выход** получаем:

– Набор $\theta \in \Theta$ значений параметров, который максимизирует значение функции f:

$$\theta : f_{\mathcal{M}}(\theta, \mathfrak{D}) \to \max$$

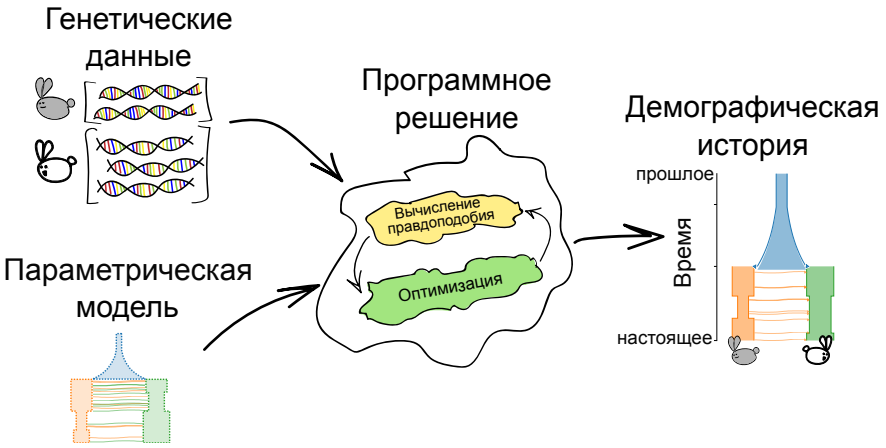


Рисунок 11 – Пример входа и выхода существующих программных решений для вывода демографической истории популяций по генетическим данным

Существует несколько **методов вывода демографической истории популяций по генетическим данным** [45, 46, 55, 83–85]. Все они требуют задания параметрической модели $\mathcal{M}$ демографической истории, а также предоставляют метод настройки ее параметров θ по генетическим данным. Они представляют собой совокупность двух компонент (рисунок 11). Первая компонента позволяет вычислить значение правдоподобия демографической истории популяций $\mathcal{D}$ и генетических данных $\mathfrak{D}$. Вторая компонента существующих методов — это оптимизация для поиска параметров заданной модели. На вход методы принимают генетические данные $\mathfrak{D}$ и параметризованную модель $\mathcal{M}$ демографической истории популяций. В качестве ответа существующие решения предоставляют настроенные параметры θ заданной модели $\mathcal{M}$, которые имеют максимальное значение правдоподобия. Таким образом, задача вывода демографической истории популяций по генетическим данным сводится к задаче поиска оптимальных параметров для заданной модели истории.

В таблице 1 приведены наиболее популярные программные средства для вывода демографической истории популяций по генетическим данным. Они отличаются интерфейсами спецификации параметрических моделей, методами вычисления правдоподобия и методами настройки параметров моделей. Их можно разделить на две группы. Первая из них включает программные средства, которые реализуют полный набор методов вывода демографической истории популяций, включая интерфейс спецификации моделей, метод вычисления правдоподобия и метод оптимизации. Вторая группа программных средств использует методы вычисления правдоподобия, реализованные в средствах первой группы, и предоставляет отличные методы оптимизации.

Таблица 1 – Существующие программные средства для вывода демографической истории популяций по генетическим данным

Программное средство	Год	Интерфейс для спецификации моделей	Метод вычисления правдоподобия	Методы оптимизации	Требуются начальные параметры	Число популяций
дaдi	2009	Да (модели I класса)	Аппроксимация диффузией	Четыре метода локальной оптимизации плюс один метод глобальной оптимизации (2020 г.)	Да	До трех
moments	2017	Да (модели I класса)	Метод моментов для статистики частоты аллелей	Четыре метода локальной оптимизации	Да	До пяти
momentsLD	2019	Да (модели I класса)	Метод моментов для статистик неравновесного сцепления генов	Четыре метода локальной оптимизации	Да	Произвольное
momi2	2020	Да (модели II класса)	Непрерывная модель Морана	Один метод усеченный метод Ньютона	Да	Произвольное
dadi-pipeline	2017	Нет (интерфейс *дaдi*)	Метод из *дaдi*	Один метод множественного запуска метода Нелдера-Мида	Нет	До трех
moments-pipeline	2019	Нет (интерфейс *moments*)	Метод из *moments*	один метод множественного запуска метода Нелдера-Мида	Нет	до пяти

К первой группе известных программных средств относятся четыре библиотеки на языке Python: *дaдi*, *moments*, *momentsLD* и *momi2*. Они требуют написания вручную пользователем программного кода для вывода демографической истории популяций.

Вход:

- генетические данные; характеристики популяций (скорости мутации);
- написанный вручную пользователем программный код для вывода демографической истории, который включает:
  - спецификацию модели определенного класса с непрерывными параметрами;
  - начальные значения параметров модели или способ их генерации случайным образом;
  - выбор метода оптимизации (BFGS, метод Пауэлла, метод Нелдера-Мида или метод BOBYQA).
  - число перезапусков выбранного метода оптимизации для разных значений начальных параметров.

Выход:

– демографическая история популяций, как модель с настроенными значениями параметров, которая имеет максимальное значение правдоподобия с генетическими данными.

Методы вычисления правдоподобия, реализованные в этих библиотеках, являются методами имитационного моделирования (рисунок 12). Они моделируют процесс эволюции согласно заданной демографической истории для вычисления или симулирования ожидаемой статистики данных. Значение правдоподобия вычисляется, как вероятность наблюдать генетические данные при условии полученной ожидаемой статистики. Библиотеки *дадi*, *moments* и *momi2* используют статистику, основанную на частотах мутаций, которая называется аллель-частотный спектр, в то время, как библиотека *momentsLD* использует набор статистик, основанных на неравновесном сцеплении генов.

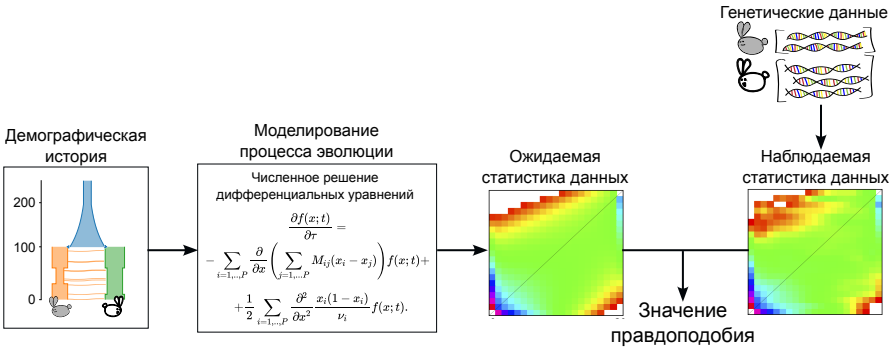


Рисунок 12 – Общая схема существующих методов вычисления значения правдоподобия

Библиотека *дадi* — одно из первых программных средств для вывода сложных демографических историй, предложенная в 2009 году [45]. Она реализует метод аппроксимации диффузии для симулирования ожидаемого аллель-частотного спектра и вычисления значения правдоподобия. Этот метод заключается в построении и численном решении дифференциального уравнения диффузии методом Чанга-Купера [86]. Интерфейс библиотеки *дадi* позволяет задать только модели первого класса, которые будут определены далее. Эти модели содержат исключительно непрерывные параметры, а динамики изменения численности популяций в них фиксированы. Для настройки параметров указанных моделей в *дадi* предложен выбор из четырех методов локальной оптимизации — BFGS [62–65], L-BFGS-B [87], метод Нелдера-Мида [66] и метод Пауэлла [67]. В 2020 году в новую версию библиотеки был включен первый метод глобальной оптимизации — метод BOBYQA [88]. Однако все включенные методы оптимизации *дадi* требуют начальных параметров модели, заданных от пользователя [45].

Библиотека *moments*, предложенная в 2017 году, основана на библиотеке *дaдi* и заменяет уравнение диффузии для симулирования ожидаемой статистики аллель-частотного спектра, используемое в *дaдi*, на систему линейных уравнений [55]. При построении этой системы используется метод моментов, который дает название библиотеке и заключается в применении аппроксимации методом складного ножа (jackknife) [89]. Для поиска решения системы используется численный метод Кранка-Николсона [90]. Интерфейс библиотеки *moments* очень схож с библиотекой *дaдi*, она также работает с моделями первого класса и предоставляет выбор из тех же четырех методов локальной оптимизации, что и первая версия *дaдi*.

Программное средство *momentsLD* является модулем библиотеки *moments* для работы со статистиками неравновесного сцепления генов [47]. Этот модуль является независимым от других модулей библиотеки *moments*, поэтому он выделен как отдельное программное средство. Метод вычисления правдоподобия симулирует набор ожидаемых статистик неравновесного сцепления генов, используя рекурсивные уравнения Хила-Робертсона [80]. Для поиска решений этих уравнений применяется метод моментов и численный метод Кранка-Николсона, разработанные для библиотеки *moments* [47, 55]. Как и библиотека *moments*, модуль *momentsLD* работает с моделями первого класса и реализует четыре метода локальной оптимизации для настройки их параметров.

Стоит отметить, что сложность методов вычисления правдоподобия растет с увеличением рассматриваемого числа популяций. Как следствие, некоторые программные средства поддерживают только ограниченное число популяций для анализа. Так, например, *дaдi* и *moments* имеют экспоненциальную сложность методов вычисления правдоподобия и позволяют анализировать только до трех и пяти популяций соответственно.

Библиотека *momi2* была разработана в 2020 году [46]. Она симулирует ожидаемую статистику аллель-частотного спектра, используя метод динамического программирования, основанный на модели Морана, которая моделирует изменение частот мутаций между поколениями в популяции. Этот метод является сопряженным методам, реализованным в *дaдi* и *moments*, однако имеет отличную от них стабильность работы. Метод вычисления правдоподобия в *momi2* имеет линейную сложность от числа популяций и поэтому позволяет анализировать произвольное число популяций, но не поддерживает непрерывные миграции и линейный закон изменения численности. Библиотека *momi2* позволяет работать с другим классом моделей, который в данной работе называется вторым классом моделей и описан далее. Эти модели, как и модели первого класса, имеют исключительно непрерывные параметры, а динамики изменения численности в них зафиксированы. Кроме этого, модели второго типа не поддерживают линейную динамику и могут иметь произвольный порядок событий в истории популяций. Для настройки параметров моделей библиотека *momi2* реализует усеченный метод Ньютона.

Часто авторы при создании программных средств таких, как *∂a∂i*, *moments*, *momentsLD* и *momi2*, концентрируются на разработке метода вычисления правдоподобия, оставляя оптимизацию классическим методам локальной оптимизации, реализованным в общедоступных популярных библиотеках таких, например, как SciPy [91]. Эти методы требуют начальной оценки значений параметров и осуществляют поиск оптимума в окрестности этой точки. Для более надежного поиска рекомендуется использовать метод множественного запуска из разных начальных точек, однако он не реализован в упомянутых программных средствах.

В 2017 и 2019 годах в работах Д. Портика были представлены программные средства *dadi pipeline* и *moments pipeline* как оболочки для библиотек *∂a∂i* и *moments* соответственно [49, 50]. Они реализуют метод множественного запуска локальной оптимизации Нелдера-Мида.

1.3. Методы моделирования демографической истории популяций

Для поиска демографических историй используются параметрические модели, или просто модели. В данном разделе приведено описание моделей демографических историй, которые применяются в существующих программных средствах. В подразделе 1.3.1 приведено описание, определение и примеры моделей первого класса, которые используются в программных решениях *∂a∂i*, *moments* и *momentsLD*. Подраздел 1.3.2 содержит описание, определение и примеры моделей второго класса, спецификация которых реализована в библиотеке *momi2*.

1.3.1. Модели первого класса

Для определения первого класса моделей, который используется в библиотеках *∂a∂i*, *moments* и *momentsLD*, приведем определение элементов этих моделей: временного интервала и разделения.

Временной интервал описывает некий промежуток во времени определенной длины, в течение которого для каждой популяции из набора заданы начальная и конечная численность, а также функция изменения численности: константная, линейная, экспоненциальная. Элемент разделения определяет популяцию, которая разделилась.

Модель первого класса — это параметрическая модель демографической истории, которая описывается набором временных интервалов и разделений. Приведем формальные определения.

Определение 7. Элемент временного интервала $\mathcal{I}$ — это шестерка $\langle p, T, \mathfrak{N}^{\text{start}}, \mathfrak{N}^{\text{end}}, \mathfrak{M}, \mathfrak{d} \rangle$, где $p \in \mathbb{N}$ — число популяций, T — время продолжительности временного интервала, $\mathfrak{N}^{\text{start}} = \{N_1^s, \ldots, N_p^s\}$ — численности каждой из популяций в начале временного интервала, $\mathfrak{N}^{\text{end}} = \{N_1^e, \ldots, N_p^e\}$ — численности каждой из популяций в конце, $\mathfrak{M} = \{m_{i,j}\}_{i \neq j},\ i, j = 1 \ldots, p$ — темпы

непрерывных миграций между популяциями, $\mathfrak{d} = \{d_1, \ldots, d_p\},\ d_i \in \{0, 1, 2\}$ — закон изменения численности.

Определение 8. Характеристиками $\chi(\mathcal{I})$ временного интервала $\mathcal{I}$ называется множество $\{T, N_1^s, \ldots, N_p^s, N_1^e, \ldots, N_p^e, m_{1,2}, \ldots, m_{p,p-1}\}$.

Определение 9. Элемент единичной миграции $\mathcal{A}$ — это тройка $\langle i^{\text{from}}, i^{\text{to}}, m\rangle$, где i^{from} — популяция-исток, i^{to} — популяция-сток, m — интенсивность единичной миграции.

Определение 10. Характеристиками $\chi(\mathcal{A})$ элемента единичной миграции $\mathcal{A}$ называется множество $\{m\}$.

Определение 11. Элемент разделения $\mathcal{S}$ — это двойка чисел $\langle p, i\rangle$, где p — число популяций до разделения, $i \in \{1, \ldots, p\}$ — индекс разделившейся популяции. Популяция с индексом i разделяется на две популяции с индексами i и $p + 1$. Элемент разделения не имеет характеристик — $\chi(\mathcal{S}) = \varnothing$.

Определение 12. Элемент инбридинга $\mathcal{W}$ — это набор коэффициентов инбридинга $\{F_i\}_{i=1}^p$, где p — число популяций.

Определение 13. Характеристиками $\chi(\mathcal{W})$ элемента инбридинга $\mathcal{W}$ называется множество $\{F_1, \ldots, F_p\}$.

Определение 14. Модель первого класса для демографической истории P популяций — параметрическая модель для демографической истории P популяций, которая представляется в виде тройки $\langle\Theta, \mathcal{E}, \mathfrak{F}\rangle$, где $\Theta \subset \mathbb{R}_+^d$ — множество значений непрерывных параметров модели, $\mathcal{E} = \{E_i\}_{i=1}^K,\ E_i \in \mathcal{I} \cup \mathcal{A} \cup \mathcal{S} \cup \mathcal{W}$ — последовательность элементов временных интервалов, единичных миграций и разделений, $\mathfrak{F} : \Theta \to \bigcup \chi(E_i)$ — отображение параметров модели в набор характеристик элементов.

На рисунке 13 приведен пример параметрической модели, которая относится к первому классу моделей. Она описывает демографические истории двух популяций, у которых размер предковой популяции до разделения равен сумме размеров новообразованных популяций после разделения. Заметим, что такую модель можно представить в виде последовательности $\{I_1, S_1, I_2\}$, где I_1, I_2 — элементы временных интервалов, а S_1 — элемент разделения. Отображение $\mathfrak{F}$ задает зависимости между параметрами и характеристиками модели. Оно может быть взаимно-однозначным — каждой характеристике элементов ставить параметр в соответствие. Однако обычно это не так, и число параметров модели строго меньше числа характеристик всех ее элементов. Элемент инбридинга не включен в модель, так как инбридинг не рассматривается в этой модели, однако его можно включить как последний элемент и задать отображение в коэффициенты, равные нулю.

Программные средства *дадi*, *moments*, *momentsLD* — библиотеки на языке Python для работы с моделями первого класса. Каждая из них позволяет специфицировать модель первого класса и настроить ее параметры по генетическим данным. Модель демографической истории реализуется с использованием этих библиотек, как процедура на языке программирования Python.

Приведем пример как пользователь может специфицировать модель первого класса с помощью интерфейса *дадi*. Для этого рассмотрим модель демографической истории двух популяций, изображенную на рисунке 14. Схематичное изображение модели представлено на рисунке 14а. Из определения модели, как параметрического семейства демографических историй, следует, что модель при каких-то значениях параметров является демографической историей. На рисунке 14б приведены демографические истории, которые соответствуют модели со следующими значениями параметров:

1. `Nanc`: 7200, `Tp`: 40000, `N1F`: 13000, `T`: 40000, `N2B`: 500, `N2F`: 12500;
2. `Nanc`: 7200, `Tp`: 80000, `N1F`: 35000, `T`: 20000, `N2B`: 500, `N2F`: 12500;
3. `Nanc`: 7200, `Tp`: 20000, `N1F`: 13000, `T`: 60000, `N2B`: 30000, `N2F`: 500;
4. `Nanc`: 30000, `Tp`: 30000, `N1F`: 13000, `T`: 40000, `N2B`: 500, `N2F`: 12500;

Данная модель соответствует тому, что давно в прошлом была одна популяция размера `Nanc` особей, затем она в какой-то момент начала меняться. Сначала был временной интервал продолжительностью `Tp` поколений, в течение которого размер популяции был константным и равным `N1F` особей. После окончания этого интервала от этой популяции отделилась вторая популяция. После разделения на протяжении `T` поколений (второй временной интервал) первая популяция имела ту же константную численность `N1F` особей, а вторая популяция имела экспоненциальное изменение численности от `N2B` до `N2F` особей. После этого наступил настоящий момент времени, когда существуют обе рассматриваемые популяции. Все только что описанные параметры `Nanc`, `Tp`, `N1F`, `T`, `N2B` и `N2F` — это параметры рассматриваемой модели.

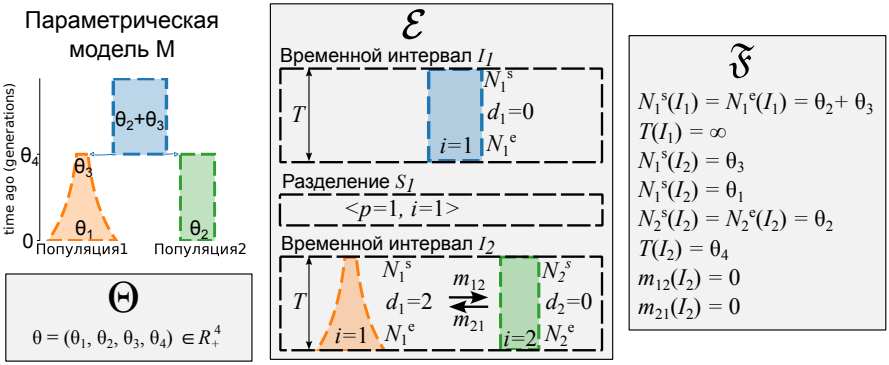


Рисунок 13 – Пример модели $M = \langle \Theta, \mathcal{E}, \mathfrak{F} \rangle$ первого класса

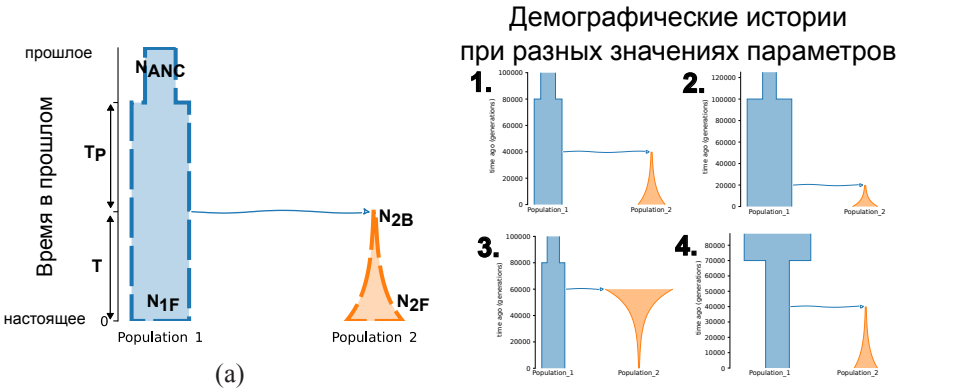


Рисунок 14 – Модель демографической истории с параметрами и демографические истории при разных значениях параметров

Это модель первого класса, так как она представляется в виде последовательности $\{I_1, I_2, S_1, I_3\}$, где I_1, I_2, I_3 — элементы временных интервалов, а S_1 — элемент разделения. Эти элементы показаны на рисунке 15.

Теперь рассмотрим как будет выглядеть интерфейс *да∂*i для задания этой модели. Для программной реализации модели требуется реализовать процедуру `model` на языка программирования Python, которая на вход принимает переменные — параметры модели. Рисунок 15 демонстрирует вид этой процедуры с использованием *да∂*i для задания модели. В теле функции последовательно определяются элементы временных интервалов и зависимость их характеристик от параметров модели: сначала создается первый элемент временного интервала I_1 с константной численностью `Nanc` одной популяции, затем создается второй временной интервал I_2 длины $T(I_2) =$ `Tp` с константной численностью $N_1^s(I_2) = N_1^e(I_2) =$ `N1F` одной популяции, потом следует элемент разделения первой популяции на две и, наконец, третий элемент временного интервала длины `T` для двух популяций, одна из которых имеет константную численность $N_1^s(I_3) = N_1^e(I_3) =$ `N1F`, а вторая — экспоненциальное изменение $d_2(I_3) = 2$ от $N_1^s(I_3) =$ `NB` до $N_1^e(I_3) =$ `N2F`.

Такой способ задания модели имеет ряд неудобств для пользователя, например, создание сетки `xx` для численных вычислений с использованием аргумента `pts`, а также передача `xx` и объекта `phi` во все используемые процедуры библиотеки. Это следствие того, что реализуемая процедура напрямую вычисляет значение ожидаемой статистики данных `sfs` для специфицированной демографической истории. Объект `phi` является решением уравнения диффузии, которое находится с применением численных методов с сеткой `xx`. Модель, за-

```python
import dadi

def model(params, ns, theta0, pts):
    Nanc, N1F, N2B, N2F, Tp, T = params

    # Задание сетки для численных вычислений
    xx = yy = dadi.Numerics.default_grid(pts)

    # Инициализация модели начальным размером популяции
    phi = dadi.PhiManip.phi_1D(xx, nu=Nanc, theta0=theta0)

    # Первый временной интервал
    # Функция изменения численности - константа N1F
    phi = dadi.Integration.one_pop(phi, xx, T=Tp, nu=N1F, theta0=theta0)

    # Второй элемент модели - разделение популяции
    phi = dadi.PhiManip.phi_1D_to_2D(xx, phi)

    # Функция изменения численности первой популяции - константа N1F
    # Задание функции изменения численности второй популяции
    n2_func = lambda t: N2B * (N2F / N2B) ** (t / T)
    # Третий элемент - второй временной интервал
    phi = dadi.Integration.two_pops(phi, xx, T=T, nu1=N1F, nu2=n2_func,
                                    theta0=theta0)

    # Вычисляем численными методами ожидаемую статистику данных
    sfs = dadi.Spectrum.from_phi(phi, ns, (xx,yy))
    return sfs
```

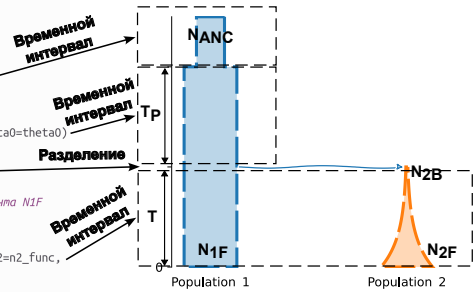


Рисунок 15 – Пример задания модели демографической истории с использованием интерфейса библиотеки *∂a∂i*

данная описанным образом с помощью библиотеки *∂a∂i*, может быть использована исключительно только для *∂a∂i*.

Величина `theta0` равна $4 \cdot \mu \cdot L$, где μ — это скорость мутации особей рассматриваемого вида, а L — длина генетической последовательности генетических данных. Значения μ и L, а следовательно и `theta0`, определяются пользователем. Множитель «4» присутствует в формуле в силу сложившихся области популяционной генетики обозначений для модели бесконечного числа сайтов [45].

Библиотека *moments* имеет схожий интерфейс спецификации моделей первого класса. Рисунок 16 демонстрирует пример спецификации модели, изображенной на рисунке 14, с использованием библиотеки *moments*. Величина `theta0` — та же самая, что используется в *∂a∂i*.

```python
import moments

def model(params, ns, theta0):
    Nanc, N1F, N2B, N2F, Tp, T = params

    # Инициализация модели начальным размером популяции
    sts = moments.LinearSystem_1D.steady_state_1D(ns[0] + ns[1])
    fs = moments.Spectrum(sts, N=Nanc, theta0=theta0)

    # Первый временной интервал
    # Функция изменения численности - константа N1F
    fs.integrate([N1F], Tp, theta0=theta0)

    # Второй элемент модели - разделение популяции
    fs = moments.Manips.split_1D_to_2D(fs, ns[0], ns[1])

    # Задание функции изменения численности обеих популяций
    n_func = lambda t: [N1F, N2B * (N2F / N2B) ** (t / T)]
    # Третий элемент - второй временной интервал
    fs.integrate(n_func, T, theta0=theta0)

    return fs
```

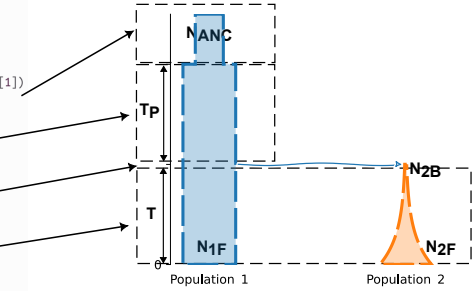


Рисунок 16 – Пример задания модели демографической истории с использованием интерфейса библиотеки *moments*

Пример спецификации модели первого класса с использованием библиотеки *momentsLD* изображен на рисунке 17. Следует отметить применение величин `rho` и `theta` в данной спецификации. Величина `rho` равна $4 \cdot$`Nanc`$\cdot$`r_bins`, где `r_bins` — набор генетических расстояний для вычисления статистик неравновесного сцепления генов. Величина `theta` — вероятность позиции мутировать хотя бы в одной хромосоме в популяции. Она равна $4 \cdot$ `Nanc` $\cdot \mu$, где μ — вероятность мутации одной позиции в одной особи популяции. Множитель «4» присутствует, так как численность `Nanc` является эффективной численностью диплоидной популяции и равна числу женских диплоидных хромосом. Следовательно, общее число хромосом будет равно $2 \cdot 2 \cdot$ `Nanc`.

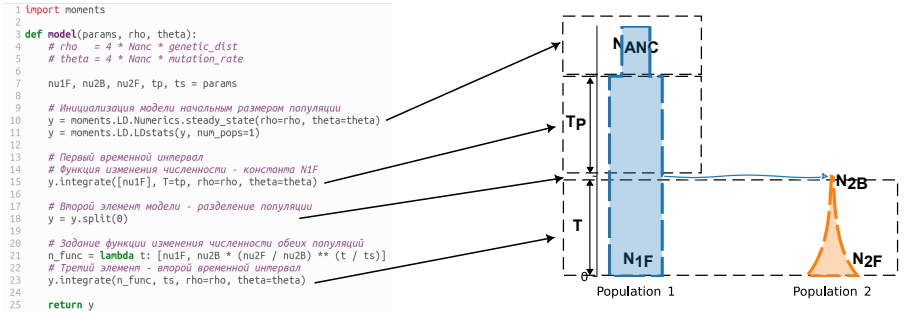


Рисунок 17 – Пример задания модели демографической истории с использованием интерфейса библиотеки *momentsLD*

Заметим, что спецификация с использованием библиотеки *momentsLD* не включает в себя параметры модели `Nanc`, `Tp`, `N1F`, `T`, `N2B` и `N2F`. Вместо этого величины `rho` и `theta` по их определению уже содержат параметр `Nanc`, а спецификация включает набор других параметров, которые определены следующим образом:

$$\texttt{nu1F} = \texttt{N1F}/\texttt{Nanc},$$

$$\texttt{nu2B} = \texttt{N2B}/\texttt{Nanc},$$

$$\texttt{nu2F} = \texttt{N2F}/\texttt{Nanc},$$

$$\texttt{tp} = \texttt{Tp}/(2 \cdot \texttt{Nanc}),$$

$$\texttt{ts} = \texttt{T}/(2 \cdot \texttt{Nanc}).$$

Спецификация библиотеки *momentsLD* требует использования параметров, определенных специальными преобразованиями относительно параметра численности популяции-предка, что вызывает дополнительные трудности при применении библиотеки.

Можно выделить следующие недостатки моделей первого класса:

– имеют только непрерывные параметры;
– динамики изменения численности (константная численность, линейное или экспоненциальное изменение) фиксированы в моделях.

При использовании библиотек *дaдi*, *moments* и *momentsLD* для спецификации моделей можно выделить следующие недостатки:
– библиотеки позволяют работать только с моделями первого класса;
– каждая модель специфицируется вручную с использованием специфичного интерфейса библиотек *дaдi*, *moments* и *momentsLD*;
– модели нельзя переиспользовать. Модель, заданную с помощью одной из библиотек *дaдi*, *moments* или *momentsLD*, можно применять только для вывода демографической истории с использованием этой же библиотеки.

1.3.2. Модели второго класса

Библиотека *momi2* работает с другим классом моделей демографической истории, чем *дaдi*. Модель второго класса — это параметрическая модель демографической истории, которая представляется в виде набора событий изменения численности и разделений. Событие изменения численности задает численность популяции в какой-то момент времени в прошлом, а также описывает ее изменение до этого момента, используя экспоненциальный закон. Константная численность воспринимается как экспоненциальное изменение со степенью, равной нулю. События разделений описывают отделение одной популяции от другой, они маркируют вершины дерева разделений модели индексами текущих популяций так, что индекс вершины-родителя всегда равен индексу строго одного из его потомков. Модели второго класса не включают события инбридинга, что связано как с определением модели, так и с тем, что библиотека *momi2* не поддерживает коэффициенты инбридинга, отличные от нуля.

Определение 15. Событие изменения численности C — это четверка $\langle p, T, N, r\rangle$, где p — индекс популяции, численность которой изменилась, T — время окончания изменения численности, N — численность популяции в конце, r — степень экспоненциального изменения популяции.

Определение 16. Характеристиками $\chi(C)$ события изменения численности C называется множество $\{T, N, r\}$.

Определение 17. Событие единичной миграции A — это тройка $\langle i^{from}, i^{to}, m\rangle$, где i^{from} — популяция-исток, i^{to} — популяция-сток, m — интенсивность единичной миграции.

Определение 18. Характеристиками $\chi(A)$ события единичной миграции $\mathcal{A}$ называется множество $\{m\}$.

Определение 19. Событие разделения U — это двойка $\langle p_{from}, p_{to} \rangle$, где p_{from} — индекс популяции, от которой произошло отделение, p_{to} — индекс популяции, которая образовалась. Событие разделения не имеет характеристик — $\chi(U) = \varnothing$.

Определение 20. Модель второго класса для демографической истории P популяций — параметрическая модель для демографической истории P популяций, которая представляется в виде тройки $\langle \Theta, \mathcal{E}, \mathfrak{F} \rangle$, где $\Theta \subset \mathbb{R}_+^d$ — множество значений непрерывных параметров модели, $\mathcal{E} = \{E_i\}_{i=1}^K$, $E_i \in C \cup A \cup U$ — набор событий изменения численности, единичных миграций и разделений, $\mathfrak{F} : \Theta \to \bigcup_i \chi(E_i)$ — отображение параметров модели в характеристики событий.

Приведем пример модели второго класса. Модель первого класса, изображенная на рисунке 13, является также и моделью второго класса. Рисунок 18 изображает ее представление, как модели второго класса.

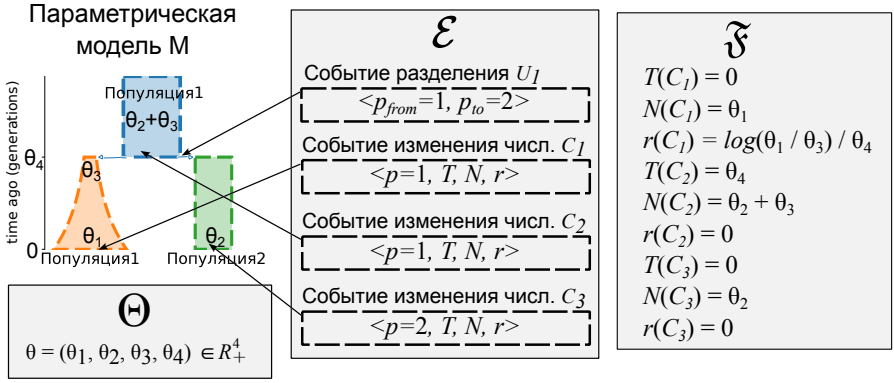


Рисунок 18 – Пример модели $M = \langle \Theta, \mathcal{E}, \mathfrak{F} >$ второго класса

Однако первый и второй классы моделей различны. Приведем пример модели, которая является моделью второго класса и не относится к моделям первого класса. Опишем ее следующим образом.

Рассмотрим две популяции. Первая из них существовала с момента существования вида (∞ поколений назад) и ее начальная численность была `Nanc` особей. Эта первая популяция в какой-то момент времени `Tc` поколений назад изменила свою численность, и размер популяции стал равен `N1F` особей. Вторая популяция отделилась от первой `T` поколений назад и имела экспоненциальный рост численности со степенью `r2` и ее численность в настоящий момент составляет `N2F` особей. Все только что описанные параметры `Nanc`, `Tc`, `N1F`, `T`, `r2` и `N2F` — это параметры рассматриваемой модели.

Изображение этой модели представлено на рисунке 19а. На рисунке 19б приведены демографические истории, которые соответствуют модели со следующими значениями параметров:

1. `Nanc`: 7200, `Tc`: 80000, `N1F`: 13000, `T`: 40000, `r2`: $8 \cdot 10^{-5}$, `N2F`: 12500;
2. `Nanc`: 7200, `Tc`: 100000, `N1F`: 35000, `T`: 20000, `r2`: $16 \cdot 10^{-5}$, `N2F`: 12500;
3. `Nanc`: 7200, `Tc`: 30000, `N1F`: 13000, `T`: 60000, `r2`: $5 \cdot 10^{-5}$, `N2F`: 12500;
4. `Nanc`: 30000, `Tc`: 30000, `N1F`: 13000, `T`: 60000, `r2`: $-1 \cdot 10^{-5}$, `N2F`: 10000;

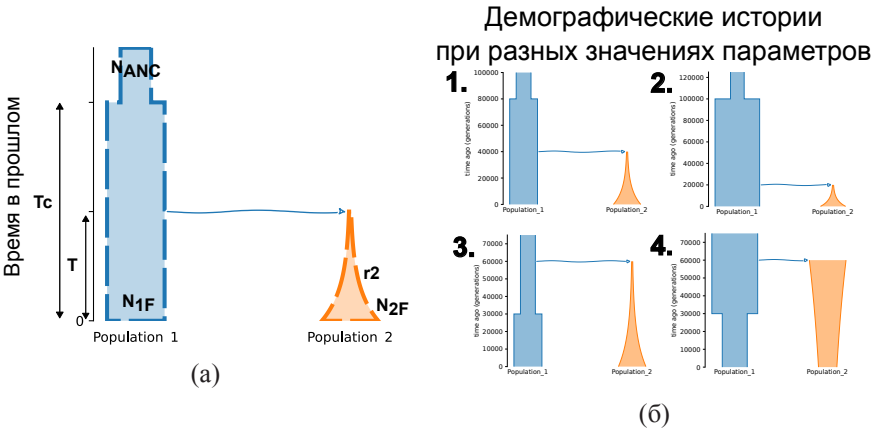


Рисунок 19 – Модель демографической истории с параметрами и демографические истории при разных значениях параметров

Параметры общие для моделей, изображенных на рисунках 14а и 19а, имеют одинаковые названия. Однако модель на рисунке 19а имеет два новых параметра: `r2` и `Tc`. Заметим, что если значение параметра `Tc` больше, чем значение параметра `T`, то эту модель можно легко перевести в модель, показанную на рисунке 14а, используя следующие уравнения для значений параметров `Tp` и `N2B`:

$$\texttt{Tp} = \texttt{Tc} - \texttt{T}$$

$$\texttt{N2B} = \texttt{N2F} \cdot e^{-\texttt{r2} \cdot \texttt{T}}.$$

При реализации модели с использованием библиотеки *momi2* пользователю требуется задать объект специального класса `momi.DemographicModel`. Это способствует переиспользованию этого класса для других программных средств. Для задания модели требуется спецификация объектов параметров-переменных, для каждого из которых можно указать границы, однако библиотека предоставляет значения границ по умолчанию. Рисунок 20 показывает реализацию модели с применением библиотеки *momi2*.

Перечислим недостатки моделей второго класса:

– имеют только непрерывные параметры;
– динамики изменения численности (константная численность или экспоненциальное изменение) зафиксированы в модели;

```python
import momi

model = momi.DemographicModel(mutation_rate=1.25e-8)

# Спецификация параметров модели
model.add_size_param("Nanc")
model.add_size_param("N1F")
model.add_size_param("N2F")
model.add_growth_param("r2")
model.add_time_param("Tp")
model.add_time_param("T")

# Задаем все события в обратном порядке: от настоящего времени
# Два листа-популяции
model.add_leaf("1", N="N1F", g=0)
model.add_leaf("2", N="N2F", g="r2")

# Популяция 2 произошла от популяции 1, задаем разделение
# Так как время в обратном порядке, сливаем 2 популяцию в первую
model.move_lineages("2", "1", t="T")

# Изменяем размер первой популяции
model.set_size("1", N="Nanc", t="Tp")
```

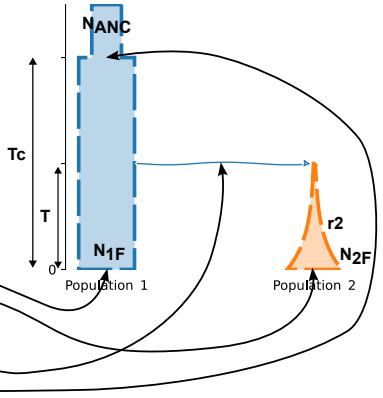


Рисунок 20 – Пример задания модели демографической истории с использованием интерфейса библиотеки *momi2*

– поддерживает только два закона изменения численности: константный и экспоненциальный. Невозможно, например, задать линейное изменение численности;
– не поддерживают непрерывные миграции и коэффициенты инбридинга;
– порядок событий, таких как изменение численности или разделение популяций, не зафиксирован и может меняться в зависимости от значений параметров.

При использовании библиотеки *momi2* для спецификации моделей можно выделить следующие недостатки:

– библиотека позволяет работать только с моделями второго класса;
– позволяет использовать определенный набор параметров моделей (параметр численности, параметр времени, параметр степени экспоненциального изменения);
– каждая модель задается вручную с применением специфичного интерфейса библиотеки *momi2*.

1.3.3. Методы сравнения моделей с разным числом параметров

Проблема выбора модели заключается в том, что необходимо выбрать наиболее подходящую модель для данных. Однако, если выбрать модель, которая слишком проста — содержит мало параметров, то она может быть не способна отображать всю информацию из данных и привести к недообучению. Если выбрать слишком сложную модель с большим числом параметров, она может быть слишком гибкой и переобучиться на шуме генетических данных, который всегда в них присутствует.

Выбор модели с оптимальным числом параметров является важной задачей в машинном обучении и статистике. Для этого широко применяются кросс-валидация и регуляризация. Кросс-валидация состоит в разделении данных на несколько непересекающихся подмножеств и обучении модели на одном из них, а оценке на другом. Таким образом, процедура кросс-валидации предоставляет более точную оценку производительности модели. Регуляризация позволяет уменьшить сложность модели и предотвратить переобучение. Это достигается путем добавления штрафа за большие значения параметров в функцию потерь. Благодаря этому модель предпочитает более простые решения, что обеспечивает лучшую обобщающую способность.

Также в выборе оптимальной модели помогают критерии Акаике и Байеса. Они учитывают как точность модели, так и ее сложность и выбирают модель с наименьшей сложностью и наибольшей вероятностью правильного описания данных. Информационный критерий Акаике (AIC) определяет модель, которая максимизирует отношение правдоподобия к числу параметров [43]. Он определяется как:

$$\mathrm{AIC}(\mathcal{M}, \mathfrak{D}) = 2 \cdot k - 2 \cdot \log \mathcal{L}(\theta^* | \mathfrak{D}),$$

где k — число параметров θ модели $\mathcal{M}$, $\mathcal{L}(\theta^* | \mathfrak{D})$ — максимальное значение функции правдоподобия модели $\mathcal{M}$ для данных $\mathfrak{D}$. Модель с наименьшим значением AIC считается наилучшей.

Байесовский информационный критерий (BIC) определяет модель, которая максимизирует правдоподобие, учитывая штраф за сложность модели, отличный от AIC-критерия [60]. BIC учитывает более строгое условие на сложность модели, что приводит к выбору более простых моделей, чем AIC. Это достигается путем умножения значения логарифма правдоподобия на размер выборки и вычитания из этого произведения числа параметров, умноженного на логарифм размера выборки. При этом формула для вычисления BIC выглядит следующим образом:

$$\mathrm{BIC}(\mathcal{M}, \mathfrak{D}) = k \cdot \log(n) - 2 \cdot \log \mathcal{L}(\theta^* | \mathfrak{D}),$$

где k — число параметров θ модели $\mathcal{M}$, $\mathcal{L}(\theta^* | \mathfrak{D})$ — максимальное значение функции правдоподобия модели $\mathcal{M}$ для данных $\mathfrak{D}$, а n — размер выборки данных $\mathfrak{D}$. Модель с наименьшим значением BIC считается наилучшей.

Если необходимо выбрать лучшую модель из двух, где одна вложена в другую, можно применить тест отношения правдоподобия (likelihood ratio test, LRT). Предположим, что задана полная или расширенная модель $\mathcal{M}_{full}$ с параметрами θ_{full} и вложенная модель $\mathcal{M}_{nested}$ с параметрами $\theta_{nested} \subset \theta_{full}$. Можно рассматривать вложенную модель $\mathcal{M}_{nested}$ как полную, у которой подмножество параметров $\psi = \theta full \setminus \theta_{nested}$ имеет фиксированные значения $\psi_i = C_i$, где $i = 1, 2, \ldots, d$, а d — это число таких параметров ψ. Необходимо проверить гипотезу $H_0 : \psi = \bar{C}$ на выборочных данных.

Для этого можно использовать тестовую статистику λ_{LRT} отношения правдоподобий:

$$\lambda_{LRT} = 2 \log \frac{\mathcal{L}(\theta^*_{full})}{\mathcal{L}(\theta^*_{nested})} = 2(\log \mathcal{L}(\theta^*_{full}) - \log \mathcal{L}(\theta^*_{nested})),$$

где $\log \mathcal{L}(\theta^*_{full}), \log \mathcal{L}(\theta^*_{nested})$ — это максимальные значения логарифма правдоподобия полной и вложенной моделей соответственно. Если гипотеза H_0 верна, то тестовая статистика λ_{LRT} имеет распределение $\chi^2(d)$ хи-квадрат с d степенями свободы. Если значение статистики λ_{LRT} превышает критическое значение распределения при заданном уровне значимости, то ограничения отвергаются и предпочтение отдается более сложной полной модели $\mathcal{M}_{full}$. В противном случае предпочтение отдается более простой вложенной модели $\mathcal{M}_{nested}$.

1.4. Методы и программные комплексы для вычисления правдоподобия генетических данных при условии заданной демографической истории

В данном разделе приведено описание существующих методов и программных комплексов для вычисления правдоподобия генетических данных при условии заданной демографической истории, которое используется при выводе демографической истории популяций. Подраздел 1.4.1 содержит описание основных определений биологии и генетики, которые применяются в данной работе при работе с генетическими данными. В подразделе 1.4.2 включены описание и примеры статистик генетических данных, которые используются для вычисления значения правдоподобия. Описание существующих методов и программных комплексов для вычисления правдоподобия приведено в подразделе 1.4.3.

1.4.1. Основные понятия биологии и генетики

Конечной целью биологических исследований является понимание живых организмов и их функционирования. Одним из основных объектов изучения является генетический материал, который управляет наследственностью и определяет все процессы в организме. Главными компонентами генетического материала являются ДНК (дезоксирибонуклеиновая кислота), РНК (рибонуклеиновая кислота) и белки. На рисунке 22 представлена иллюстрация основных элементов генетического материала клетки, подробное описание которых приводится ниже.

Дезоксирибонуклеиновая кислота (ДНК) является основной молекулой наследственности и содержит информацию о строении и функционировании организма. Она представляет собой двухцепочечную структуру, образованную последовательностью нуклеотидов. ДНК состоит из четырех типов нуклеотидов (аденин, гуанин, цитозин и тимин), которые образуют две комплементарные цепи, связанные между собой водородными связями. ДНК измеряется в единицах длины нуклеотидной цепи — в *парах оснований* (bp). Она определяется только

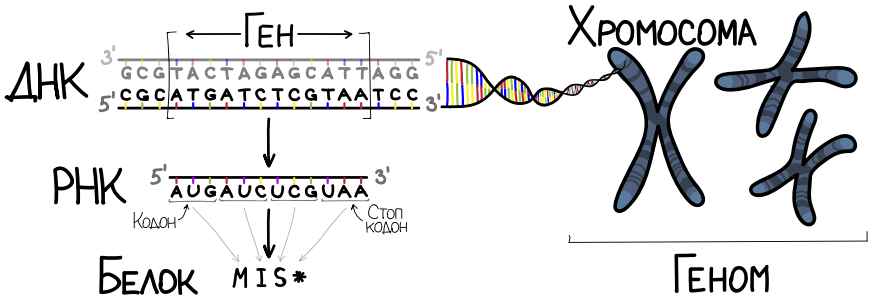


Рисунок 21 – Главные компоненты генетического материала клетки.

одной цепочкой — вторая достраивается согласно комплементарности нуклеотидов, поэтому ДНК можно рассматривать как последовательность $\mathfrak{D}$ длины g символов из четырех возможных букв — A (аденин), C (цитозин), G (гуанин) и T (тимин):

$$\mathfrak{D} = [\mathfrak{d}_1, \mathfrak{d}_2, \ldots, \mathfrak{d}_g], \quad \forall \mathfrak{d}_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}.$$

ДНК содержит *гены* — участки, которые кодируют информацию для синтеза белка или РНК молекулы. Кроме генов, в геноме могут быть и другие участки ДНК, такие, как регуляторные элементы, которые могут влиять на экспрессию генов, или некодирующие участки, которые не содержат информации для синтеза белка или РНК молекулы. Если рассматривать ДНК как последовательность символов $\mathfrak{D}$, то гены — это подстроки этой строки $\mathfrak{D}$.

На рисунке 21 представлена ДНК длины 18 пар оснований, состоящая из двух цепочек — кодирующей и комплементарной. На показанном фрагменте ДНК выделен участок — ген, с которого происходит считывание РНК.

Рибонуклеиновая кислота (РНК) выполняет различные функции в клетке, включая передачу информации от ДНК к белкам и участие в процессе синтеза белков. РНК содержит те же нуклеотиды, что и ДНК, за исключением тимина, который заменяется урацилом. В отличие от ДНК, молекула РНК обычно одноцепочечная. РНК можно рассматривать как последовательность R длины p символов из четырех возможных букв — A (аденин), C (цитозин), G (гуанин) и U (урацил):

$$R = [r_1, sr_2, \ldots, r_p], \quad \forall r_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}.$$

На рисунке 22 показана цепочка РНК, полученная из гена на ДНК с помощью процесса, называемого транскрипцией. РНК является посредником между ДНК и белками. Каждая тройка нуклеотидов в РНК называется кодоном и кодирует определенную аминокислоту в белке. Процесс трансляции позволяет синтезировать белок из молекулы РНК. На рисунке 22 продемонстрированы кодоны на цепочке РНК и белок, транслированный из них.

Белки выполняют различные функции в организме, включая катализ химических реакций, транспорт веществ, поддержание структуры клеток и участие в

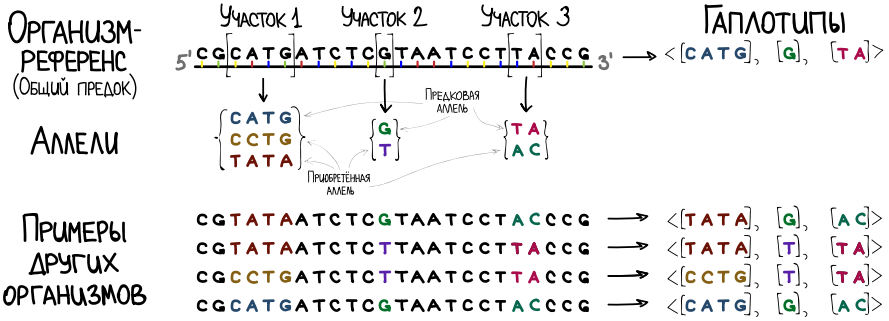


Рисунок 22 – Понятие аллели как варианта участка ДНК и генотипа как совокупность аллелей организма.

иммунной системе. Белки состоят из аминокислот, которые связываются между собой пептидными связями. Последовательность аминокислот в белке определяет его структуру и функцию.

ДНК упаковывается в *хромосомы*, которые могут иметь различные формы, в зависимости от типа организма и стадии клеточного цикла. У бактерий хромосома обычно представляет собой кольцевую молекулу ДНК. У более сложных организмов хромосомы имеют форму линейных нитей и находятся в ядре клетки. Во время клеточного деления хромосомы обычно конденсируются в более компактные формы, что позволяет им удобнее перемещаться внутри клетки и обеспечивает более точное распределение генетической информации между дочерними клетками. На рисунке 22 показано сворачивание цепочки ДНК в хромосому, которая имеет типичный для деления клетки вид в форме буквы X.

Гаплоидные организмы имеют один комплект хромосом в своих клетках, тогда как диплоидные организмы имеют два комплекта гомологичных хромосом. У *диплоидных организмов* каждый из двух комплектов хромосом наследуется от каждого из родителей в процессе сексуального размножения, который называется мейозом. Многие бактерии, грибы и растения являются гаплоидными, а большинство животных, включая человека, диплоидные. У человека и других млекопитающих обычно есть 23 пары гомологичных хромосом (46 хромосом в общей сложности).

Гамета — репродуктивная клетка, участвующая в половом размножении, которая имеет гаплоидный набор хромосом.

Геном — это полный набор генетической информации, содержащейся во всех хромосомах организма. Конечный размер генома может варьироваться у различных видов. Например, геном бактерии *Escherichia coli* состоит из одной хромосомы размером примерно $4,6$ миллиона пар оснований, в то время как геном человека состоит из 23 пар гомологичных хромосом, содержащих более трех миллиардов пар оснований.

Аллель — это одна из нескольких форм или вариантов, конкретного участка ДНК. Этот участок может быть геном, регуляторным элементом, некодирующей ДНК или другим типом ДНК-последовательности. Аллели могут быть различными версиями одного и того же гена, например, различающимися в одном нуклеотиде или в нескольких нуклеотидах. Различные аллели могут влиять на фенотипические (наблюдаемые) черты, такие как цвет глаз или тип крови. В популяционной генетике также часто используют понятие *генотипа* — набор аллелей, которые определенный организм несет в определенных участках ДНК. Генотипы могут совпадать у разных особей и зависят от выбранного набора участков. Если две особи имеют одинаковый генотип на всех вариабельных локусах генома, то это означает полное совпадение ДНК и то, что особи — идентичные близнецы. На рисунке 22 показаны примеры аллелей и генотипов. На фрагменте ДНК выделены три вариативных участка, для каждого из которых указаны возможные аллели в группе организмов. Первый участок длиной в четыре пары оснований имеет три возможных варианта последовательности, и, следовательно, три аллели, а второй и третий участки имеют по две аллели каждый. Второй участок соответствует одному нуклеотиду на позиции 12, для которого указаны две аллели: G и T. Также на рисунке 22 приведены примеры генетической информации нескольких организмов из группы и их генотипы. Например, генотипом организма-референса является набор аллелей этого организма в указанных участках — набор $[\mathrm{CATG}, \mathrm{G}, \mathrm{TA}]$.

1.4.2. Используемые статистики генетических данных

Для вывода демографической истории популяций обычно не используют полные геномы, потому что их обработка может быть очень трудоемкой и затратной. Полные геномы представляют собой большие объемы данных, и их анализ может потребовать значительных вычислительных ресурсов. Вместо этого исследователи часто используют более компактные статистики генетических данных [70]. В данном разделе описаны два наиболее популярных типа данных: аллель-частотный спектр и группа статистик, основанных на неравновесном сцеплении генов.

Аллель-частотный спектр является одним из наиболее популярных представлений генетических данных [70, 92]. Аллель-частотный спектр — это совместное распределение частот приобретенных аллелей у P популяций. Приобретенные аллели — это аллели, которые образовались в результате эволюции путем мутаций от общего аллеля-предшественника. Аллель-частотный спектр P популяций — это P-мерный тензор $A \in \mathbb{N}^{(n_1+1)\times(n_2+1)\times\ldots\times(n_P+1)}$, где n_i равно числу хромосом в i-й популяции [93]. Каждый элемент спектра равен числу локусов, в которых приобретенная аллель встретилась у определенного числа особей в каждой из популяций. Таким образом, каждый элемент спектра $A[d_1, \ldots, d_P] \in \mathbb{N},\ d_i \in [0, n_i]$ равен числу локусов, где приобретенная аллель встретилась у d_1 особей первой популяции, d_2 особей второй популяции и т. д.

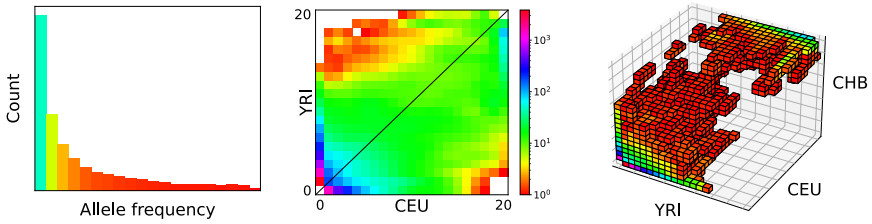
На рисунке 23 представлен пример построения аллель-частотного спектра для небольшого генома длины семь пар оснований. Референсная последовательность позволяет определить приобретенные мутации для последовательностей пяти особей. Три особи относятся к первой популяции и две ко второй, их приобретенные аллели выделены красным цветом. Для каждой позиции вычислим частоту встречаемости приобретенной аллели в каждой из популяций. Например, для второй позиции T — предковая аллель, а C — приобретенная. Аллель C имеет частоту два в первой и второй популяциях, так как она присутствует у двух особей (второй и третьей) первой популяции и у двух особей (первой и второй) второй популяции. Среди данных вторая позиция единственная с частотой два в обеих группах, поэтому аллель-частотный спектр A имеет значение $A[2, 2] = 1$, что изображено красным цветом на тепловой карте. Существует три позиции (1, 3, 6), где приобретенная аллель встречалась у одной особи в первой популяции и ни у нуля особей во второй, поэтому $A[1, 0] = 3$.

Примеры аллель-частотного спектра для одной, двух и трех популяций представлены на рисунке 24. Рисунок 25 демонстрирует зависимость аллель-частотного спектра двух популяций от демографической истории.

Статистики, основанные на неравновесном сцеплении генов также являются популярными представлениями генетических данных. Неравномерное сцепление генов (linkage disequilibrium) — это явление, когда определенные аллели двух генов находятся в тесной связи друг с другом — наследуются вместе чаще, чем ожидалось бы, если бы гены находились на разных хромосомах и независимо переносились друг от друга в процессе размножения.

Неравномерное сцепление генов возникает из-за наличия взаимодействия между аллелями разных генов, находящихся в близости друг от друга на хро-

Рисунок 23 – Пример построения аллель-частотного спектра для двух популяций



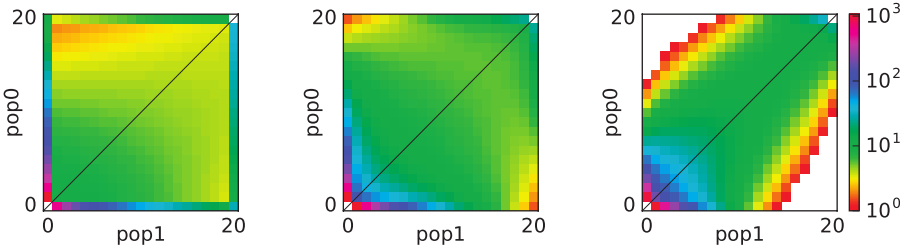
(а) (б) (в)

Рисунок 24 – Примеры аллель-частотного спектра для: а) одной популяции; б) двух популяций; в) трех популяций

мосоме. Это может происходить, например, из-за наличия мутации в одном из генов, которая приводит к изменению частоты аллелей в связанном гене. Также неравномерное сцепление генов может быть следствием естественного отбора, когда некоторые комбинации аллелей предпочтительнее для выживания и размножения, чем другие.

Генетическое расстояние — это мера расстояния между генетическими локусами, которая определяется на основе частоты рекомбинации между ними. Один из способов измерения генетического расстояния — использование единицы измерения «морган» (М) и «сантиморган» (сМ). Расстояние ρ в один сантиморган указывает на вероятность в 1% того, что два локуса будут разделены рекомбинацией в процессе мейоза. Сто сантимогранов составляют один морган.

$$\rho(x, y) = 1\text{сМ} = 100\text{М} \Leftrightarrow P(\text{рекомбинация между } x \text{ и } y) = 0.01,$$



(а) (б) (в)

Рисунок 25 – Примеры аллель-частотного спектра двух популяций, которые соответствуют разным демографическим историям двух популяций: а) изоляция, отсутствие миграции; б) между популяциями существовала небольшая миграция; в) сильная миграция между популяциями

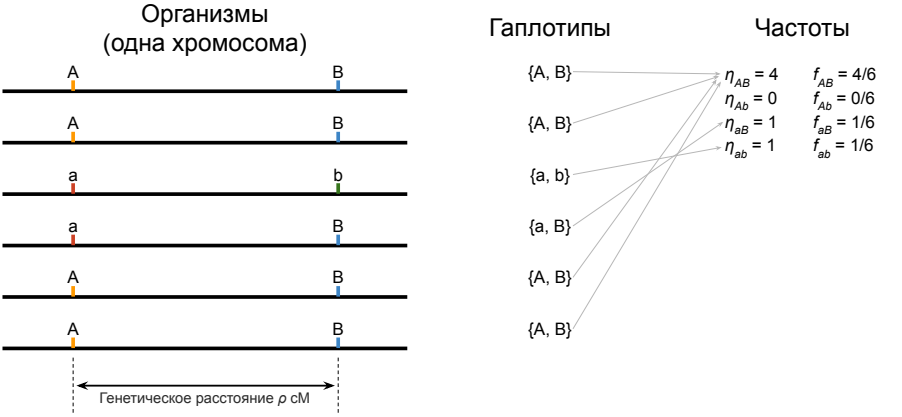


Рисунок 26 – Пример абсолютных и относительных частот гаплотипов

где x, y — физические позиции двух локусов на хромосоме. Если два генетических локуса находятся на разных хромосомах, то их генетическое расстояние считается равным 50 cM, поскольку вероятность рекомбинации между ними составляет 50%. Если они находятся на одной хромосоме, то генетическое расстояние между ними может быть менее 50 cM, поскольку вероятность рекомбинации между ними будет меньше 50%.

Рассмотрим два локуса, находящиеся на некотором фиксированном расстоянии ρ cM на одной хромосоме, что означает, что вероятность рекомбинации между этими локусами равна $r = \rho/100$. Пусть локусы биаллельны — каждый содержит одну из двух аллелей: первый локус аллели A или a, и второй — B или b. Тогда возможно образование четырех гаплотипов AB, Ab, aB и ab. Обозначим их относительные частоты, как $f_{AB}, f_{Ab}, f_{aB}, f_{ab}$. Частоты могут рассматриваться как относительные, так и абсолютные. Обозначим абсолютные частоты, как $\eta_{AB}, \eta_{Ab}, \eta_{aB}$ и η_{ab}. Если частоты абсолютные, то в сумме они будут давать размер рассмотренной выборки хромосом: $\eta_{AB} + \eta_{Ab} + \eta_{aB} + \eta_{ab} = n$, если относительные, то их сумма будет равна единице: $f_{AB} + f_{Ab} + f_{aB} + f_{ab} = 1$. На рисунке 26 изображен пример абсолютных и относительных частот гаплотипов AB, Ab, aB и ab.

Частоты в следующем поколении, при условии случайного набора с повторениями гаплотипов из предыдущего, зависят от текущих частот и от вероятности рекомбинации r между локусами [94]. Примеры передачи гаплотипов от родителей детям и их вероятности показаны на рисунке 27. Например, если родитель имеет гаплотипы AB и Ab, то вероятность передачи гаплотипа AB ребенку равна вероятности передачи Ab и составляет 50%, а родитель с комбинациями AB и ab передаст AB или ab с равной вероятностью $(1-r)/2$ и Ab или aB с вероятностью $r/2$.

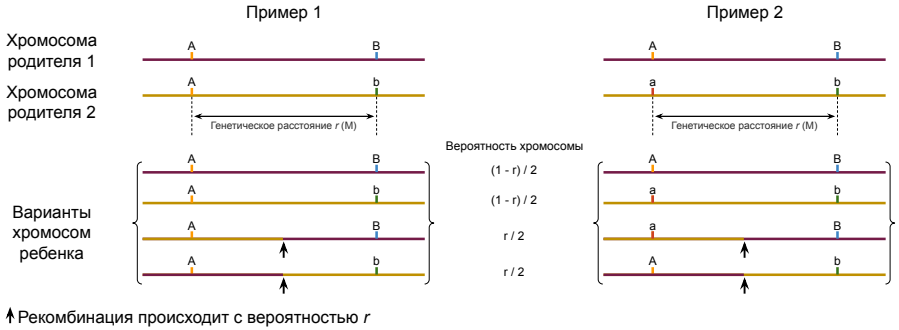


Рисунок 27 – Пример передачи гаплотипов между поколениями и их вероятности

Основываясь на идеи аллель-частотного спектра, был предложен двухлокусный гаплотип-частотный спектр — совместное распределение частот двухлокусных гаплотипов у популяций [95]. Для простоты рассмотрим пример такой статистики для одной популяции. Для каждой пары локусов, находящихся на фиксированном генетическом расстоянии r, вычисляются абсолютные частоты встречаемости η_{AB}, η_{Ab}, η_{aB} трех гаплотипов AB, Ab и aB соответственно. Заметим, что частота η_{ab} четвертого варианта гаплотипа ab однозначно определяется частотами остальных. Тогда двухлокусный гаплотип-частотный спектр — это трехмерный тензор $\Phi \in \mathbb{N}^{n \times n \times n}$, где n равно числу рассмотренных хромосом в популяции. Каждый элемент спектра $\Phi[x_1, x_2, x_3]$ равен числу пар локусов, где гаплотипы AB, Ab, aB, ab имеют частоты x_1, x_2, x_3 и $(n - x_1 - x_2 - x_3)$. Пример двухлокусного гаплотип-частотного спектра, построенного для данных 10 диплоидных особей одной популяции, представлен на рисунке 28. Такие двух локусные гаплотип-частотные спектры, построенные для набора генетических расстояний, используют в качестве данных для вывода демографической истории популяций [95].

Другой важной характеристикой неравномерного сцепления генов является коэффициент неравномерного сцепления генов, который определяется как ковариация относительных частот аллелей A и B:

$$D = f_{AB} - f_A f_B = f_{AB} f_{ab} - f_{Ab} f_{aB},$$

где f_A, f_B — относительные частоты аллелей A и B соответственно. Для вывода демографической истории используют и другие статистики неравномерного сцепления генов, например, статистику z, которая определяется следующим образом:

$$z = (1 - 2f_A)(1 - 2f_B).$$

Величина z наибольшая, когда A и B являются редкими аллелями, и положительная, если аллели A и B либо обе минорные, либо обе мажорные. Еще одна

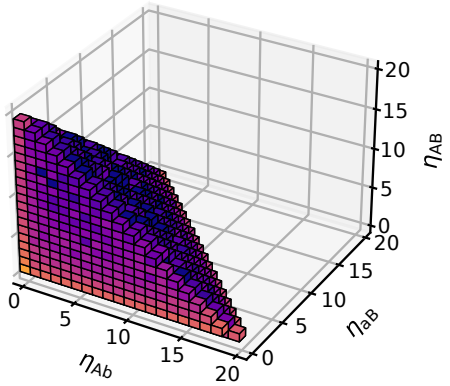


Рисунок 28 – Пример двухлокусного гаплотип-частотного спектра, построенного для данных 10 диплоидных особей одной популяции.

используемая статистика — совместная гетерозиготность π_2 среди пар локусов:

$$\pi_2 = f_A(1 - f_A)f_B(1 - f_B).$$

Статистика π_2 пропорциональна вероятности того, что, если мы случайно выберем четыре гаплотипа в популяции, то первая пара будет отличаться в первом локусе, а вторая — во втором.

Для вывода демографической истории используют различные варианты обозначенных статистик, например, D^2, Dz и π_2, которые были предложены в [47]. В качестве данных используют кривую зависимости статистики от расстояния между локусами. Для каждого генетического расстояния из фиксированного набора вычисляется статистика для пар локусов, находящихся на этом расстоянии и строится кривая изменения. Примеры кривых для D^2 и Dz, построенные для данных одной популяции представлены на рисунке 29. Линии соответствуют среднему значению статистики, а область вокруг отображает дисперсию. Кривая D^2 имеет убывающий характер, так как с увеличением расстояния между локусами вероятность случайного сопряжения (рекомбинации) между ними увеличивается, а, следовательно, среднее значение ковариации стремится к нулю.

1.4.3. Математические модели эволюции, методы дифференциального исчисления, численные методы и программные комплексы для вычисления правдоподобия

В популяционной генетике существует множество моделей, которые используются для изучения эволюционных процессов, происходящих в популяции. Каждая из этих моделей представляет собой абстрактную систему, описывающую основные свойства популяции, такие как ее размер, структура, скорость

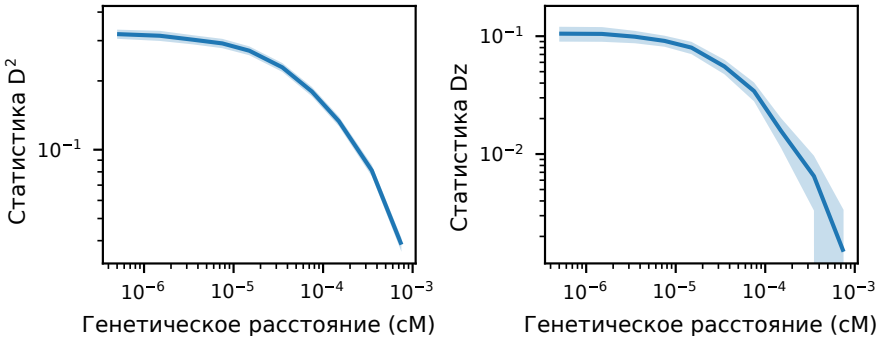


Рисунок 29 – Пример кривых зависимостей значений разных статистик от генетического расстояния между локусами.

размножения и другое. Эти модели лежат в основе существующих методов вычисления правдоподобия демографической истории популяций и данных, поэтому сначала рассмотрим несколько используемых классических моделей популяционной генетики. При этом подробно рассмотрим их свойства, а затем опишем применение этих моделей для вычисления функции правдоподобия $f_{\mathcal{M}}(\theta, \mathfrak{D})$ в дальнейшем.

Стохастическая эволюционная *модель Райта-Фишера* описывает изменение частот аллелей между поколениями. Она была предложена независимо Р. Фишером в 1922 году [96] и С. Райтом в 1931 году [97], а затем расширена М. Кимурой в 1955 году [98]. Рассмотрим самую простую модель Райта-Фишера без миграции и отбора.

Пусть задана диплоидная популяция постоянного размера N, которую будем рассматривать как популяцию $2N$ гаплоидных особей. Предположим, что поколения не пересекаются, а новое поколение формируется путем случайного выбора с повторениями особей предыдущего поколения. Рассмотрим генетический локус, в котором встречаются только две аллели A и a. Каждое поколение этой популяции содержит $2N$ копий рассматриваемого локуса. Рассмотрим величину η^t, равную числу аллелей A в поколении t, которая является биномиальной случайной величиной $\eta^t \sim Bin(n, p)$ с числом испытаний равным $n = 2N$, и вероятностью успеха $p = \frac{\eta^{t-1}}{2N}$. Тогда согласно изменение η^t между поколениями является Марковской цепью с матрицей переходов:

$$P_{ij} = P(\eta^t = i | \eta^{t-1} = j) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}.$$

Ожидаемая частота аллели A остается постоянной для разных поколений и равна начальной частоте $\mathbb{E}[\eta^t] = \eta^0$, тогда как дисперсия на поколение составляет $Var[\eta^t] = 2N \cdot f^0(1 - f^0)$, где $f^t = \frac{\eta^t}{2N}$ — относительная частота аллели A [99]. Вероятность того, что аллель A в конечном итоге станет фиксированной равна

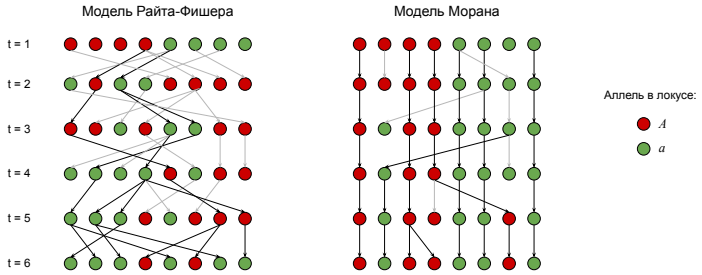


Рисунок 30 – Примеры изменения частот аллелей в модели Райта-Фишера и модели Морана

начальной относительной частоте f^0. В частности, вероятность фиксации новой мутации, присутствующей в единственном экземпляре, равна $\frac{1}{2N}$.

Пример изменения частот аллелей в модели Райта-Фишера представлен на рисунке 30. Популяция состоит из $2N = 8$ особей, в первом поколении $t = 1$ половина из которых содержит аллель A, а другая половина аллель a. В течение шести поколений новые особи выбираются случайным образом с повторениями из предыдущего и в последней популяции при $t = 6$ остается только две особи с аллелью A. Стрелочки отображают выбор особи для копирования в новое поколение, черные стрелочки отображают передачу генетической информации от первого поколения к последнему.

Модель Райта-Фишера хорошо описывает, например, популяции однолетних растений или оленей в северной части штата Нью-Йорк [100], так как поколения этих видов почти не пересекаются. Однако для многих других видов, например, человека, дрозофилы или дрожжей, такое упрощение не применимо. Существует альтернативная *модель Морана* (1958) [101], в которой поколениям разрешено пересекаться. В этой модели в моменты времени $t = 0, 1, 2, ...$ выбираются две особи случайным образом с заменой. Это может быть одна и та же особь, а могут быть и разные. Первая выбранная особь размножается — копирует себя, а вторая умирает. Таким образом, размер популяции не изменяется. Если одну и ту же особь выбрать дважды, то она размножится, а затем умрет.

Рассмотрим простейший случай, когда нет ни отбора, ни мутации. Как и для модели Райта-Фишера, предположим, что популяция состоит из $2N$ гаплоидных особей, для каждой из которых будем рассматривать единственный локус с двумя аллелями и a. Предположим, что в момент времени $t - 1$ число аллелей в популяции равно $\eta^{t-1} = j \in [1, 2, \ldots, 2N]$. Вероятность выбрать особь с аллелью A для размножения или смерти равна $f^{t-1} = \frac{\eta^{t-1}}{2N}$. Величина η^t может принимать одно из трех значений: $\eta^{t-1} + 1$, η^{t-1} и $\eta^{t-1} - 1$. Вероятность того, что η^t увеличится, равна вероятности того, что аллель будет выбрана для гибели, умноженной на вероятность того, что аллель будет выбрана для размно-

жения. Используя аналогичные рассуждения для двух других значений, можно определить следующую матрицу переходов:

$$P(\eta^t = i|\eta^{t-1} = j) = \begin{cases} f^{t-1}(1-f^{t-1}), & \text{если } j = i+1, \\ f^{t-1}(1-f^{t-1}), & \text{если } j = i-1, \\ (f^{t-1})^2 + (1-f^{t-1})^2, & \text{если } j = i, \\ 0, & \text{в противном случае,} \end{cases}$$

где $f^t = \frac{\eta^t}{2N}$ — относительная частота аллели A. Ожидаемая частота $\mathbb{E}[\eta^t]$ остается постоянной для разных поколений и равна $\mathbb{E}[\eta^t] = \eta^0$, тогда как дисперсия в каждый момент времени составляет $Var[\eta^t] = 2 \cdot f^0(1-f^0)$, где $f^t = \frac{\eta^t}{2N}$ — относительная частота аллели A [99].

Пример изменения частот аллелей в модели Морана представлен на рисунке 30. Популяция состоит из $2N = 8$ особей, в первом поколении $t = 1$ половина из которых содержит аллель A, а другая половина аллель a. На каждом из следующих пяти поколений выбирается две особи: одна для смерти и одна для размножения. В итоге на шестом поколении $t = 6$ остается четыре особи с аллелью A. Стрелочки отображают выбор особи для копирования в новое поколение, черные стрелочки отображают передачу генетической информации от первого поколения к последнему.

Кроме пересекающихся поколений, модель Морана отличается от модели Райта-Фишера генетическим разнообразием [101]. Это приводит к тому, что эффективный размер популяции вычисляется по-разному для каждой модели. Было доказано, что коэффициент гетерозиготности, который используется для измерения генетического разнообразия, уменьшается в два раза быстрее для модели Морана, чем для модели Райта-Фишера при одинаковом размере популяции $2N$ [101]. В результате диплоидный эффективный размер популяции определяется следующим образом:

$$N_e = \begin{cases} N, & \text{для модели Райта-Фишера,} \\ \frac{1}{2}N, & \text{для модели Морана.} \end{cases}$$

Модели Райта-Фишера и Морана связаны между собой *процессом коалисценции Кингсмана*, предложенным в 1982 году [102]. Модель коалисценции описывает процесс, того как нескольких последовательностей ДНК из выборки сходятся к общему предку в прошлом. Если устремить размер популяции к бесконечности $N \to \infty$ в моделях Райта-Фишера и Морана, то обе сойдутся к процессу коалисценции [99], что приводит к тому, что модели дают схожие результаты.

Для моделей Райта-Фишера и Морана был рассмотрен простейший случай эволюции без отбора и мутации. В более общем случае каждая из этих моделей может быть расширена и включать в себя мутации, отбор, существование нескольких полов, изменения численности популяций и тому подобное.

Другая модель — *модель бесконечного числа сайтов* позволяет описать процесс мутации для генетической последовательности [103]. Предположим, что геном состоит из бесконечного числа сайтов, каждый из которых может содержать либо мутировавшую — приобретенную аллель, либо предковую. Модель описывает, что длина последовательности бесконечна, а мутации возникают независимо друг от друга всегда на разных сайтах с вероятностью μ. Обратные мутации из приобретенной аллели в предковую, не допускаются. Предположим, что число новых мутировавших сайтов, приобретаемое каждой гаметой при размножении — это случайная Пуассоновская величина со средним ν, тогда вероятность того, что хотя бы один сайт мутировал равна $1 - e^{-\nu} \approx \nu$. Определим величину θ:

$$\theta = 4\nu N_e,$$

которая равна среднему числу новых мутировавших сайтов для популяции эффективного размера N_e на одно поколение [94].

В реальности, генетические последовательности не бесконечны, и величину θ поэтому можно оценить следующим образом [45]:

$$\theta = 4 \cdot N_e \cdot \mu \cdot L = N_e \cdot \theta_0,$$

где μ — оценка средней скорости мутации, равной вероятности возникновения мутации на позиции за одно поколение, L — длина последовательности, а $\theta_0 = 4u = 4 \cdot \mu \cdot L$.

Существует множество **методов для вывода демографической истории популяций по генетическим данным**. Они используют различные статистики данных, примеры которых представлены в разделе 1.4.2, и предполагают применение описанных моделей популяционной генетики. Представим некоторые из них, которые использованы в данной диссертации.

Один из наиболее популярных методов для вывода демографической истории популяций является *метод аппроксимации диффузией*, реализованный в программном обеспечении ∂a∂i [45]. Метод предполагает модель Райта-Фишера и бесконечного числа сайтов. На основе входных генетических данных $\mathfrak{D}$ строится аллель-частотный спектр $A^{\mathfrak{D}}$. В данном методе вычисление значения правдоподобия демографической истории и данных происходит в три шага. На первом из них происходит построение *уравнений диффузии* для демографической истории, а также поиск их решений численными методами. Пусть $\phi(x_1, \ldots, x_P, t)$ — плотность распределения числа приобретенных мутаций с относительными частотами $x_1, \ldots, x_P$ для популяций $1, 2 \ldots, P$ соответственно в момент времени t. Используя предположение о модели Райта-Фишера и модели бесконечного числа сайтов, можно записать уравнение диффузии с решением $\phi(x_1, \ldots, x_P, t)$ для каждого интервала времени демографической истории, когда размеры популяций константные [77]. Например, рассмотрим интервал времени длиной t поколений с постоянными размерами популяций $N_1, \ldots, N_P$ и миграциями $m_{i,j}$.

Тогда уравнение диффузии будет иметь вид:

$$\frac{\partial\phi(x;t)}{\partial\tau} = \frac{1}{2}\sum_{i=1,..,P}\frac{\partial^2}{\partial x^2}\frac{x_i(1-x_i)}{N_i}\phi(x;t) - \sum_{i=1,..,P}\frac{\partial}{\partial x}\sum_{j=1,...P} m_{i,j}(x_i - x_j)\phi(x;t).$$

Уравнение диффузии строится для каждого интервала времени демографической истории, в течении которого размеры популяций остаются постоянными. В случае непрерывной функции $g^j(t),\ j \in [1,2,\ldots,P]$ размера популяции, она равномерно аппроксимируется кусочно-постоянной функцией. Внешний вид уравнений диффузии определяется данными значениями параметров θ модели $\mathcal{M}$ демографической истории. Затем происходит последовательное *численное решение уравнений диффузии*. Конечной целью первого шага является вычисление величины $\phi(x;T)$, где T — суммарное время всех временных интервалов демографической истории.

Затем происходит второй шаг — вычисление ожидаемого аллель-частотного спектра для модели демографической истории $\mathcal{M}$ с использованием полученной величины $\phi(x) = \phi(x;T)$:

$$A^{\mathcal{M}}[d_1,d_2,\ldots,d_P] = \int_0^1\cdots\int_0^1 \prod_{i=1,2,\ldots,P}\binom{n_i}{d_i}x_i^{d_i}(1-x_i)^{n_i-d_i}\phi(x_1,x_2,\ldots,x_P)dx_i.$$

Различные *численные методы* используются на первом и втором шагах вычисления правдоподобия. Для решения уравнений диффузии используется метод Чанга-Купера, предложенного для решения уравнения Фоккера-Планка [86]. Метод Чанга-Купера позволяет получить решение уравнения на заданной сетке G путем построения и решения системы линейных алгебраических уравнений. Размерность уравнений диффузии равна числу популяций, поэтому в случае более двух популяций применяется метод переменных направлений [104, 105], который позволяет свести систему к набору тридиагональных систем, которые можно эффективно решить методом прогонки [106]. Для интегрирования, необходимого для вычисления ожидаемого аллель-частотного спектра, применяется метод трапеций [107].

На последнем третьем шаге вычисляется значение правдоподобия, исходя из предположения, что каждый элемент наблюдаемого аллель-частотного спектра $A^{\mathfrak{D}}[d_1,\ldots,d_P]$ является независимой Пуассоновской случайной величиной со средним $A^{\mathcal{M}}[d_1,\ldots,d_P]$:

$$\mathcal{L}(\theta|\mathfrak{D}) = \mathcal{L}(\mathcal{M}(\theta)|\mathfrak{D}) = \prod_{i=1,...P}\prod_{d_i=1,..n_i}\frac{e^{-A^{\mathcal{M}}[d_1,\ldots,d_P]}A^{\mathcal{M}}[d_1,\ldots,d_P]^{A^{\mathfrak{D}}[d_1,\ldots,d_P]}}{A^{\mathfrak{D}}[d_1,\ldots,d_P]!}$$

В качестве целевой функции оптимизации ∂а∂i использует логарифм правдоподобия:

$$f_{\mathcal{M}}^{\partial\text{a}\partial\text{i}}(\theta) = \log\mathcal{L}(\theta|\mathfrak{D}).$$

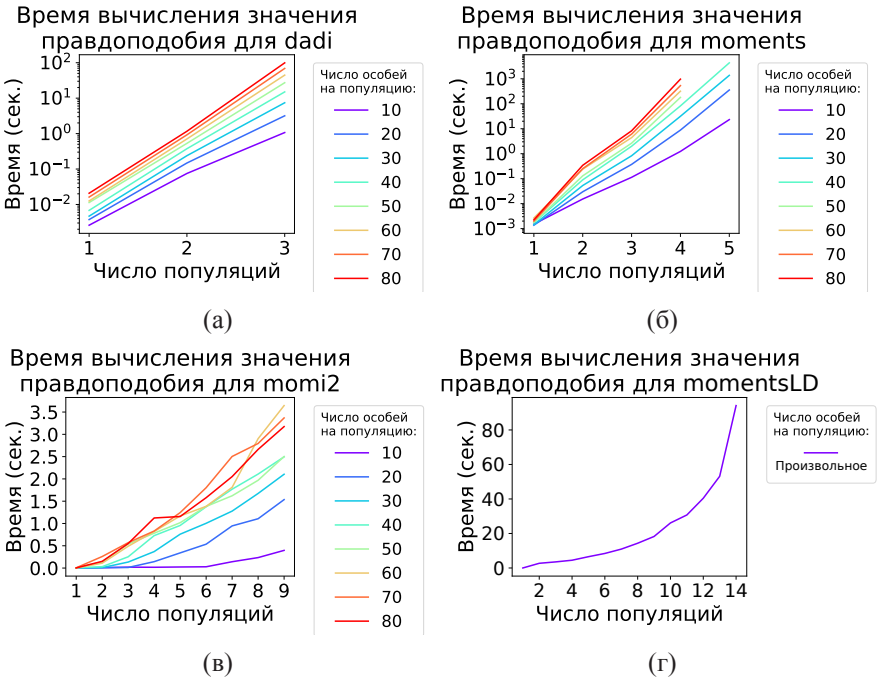


Рисунок 31 – Время вычисления значения правдоподобия для ∂a∂i, *moments*, *momentsLD* и *momi2*

Для вычисления $f_{\mathcal{M}}^{\partial a \partial i}(\theta)$ ∂a∂i требуется размер сетки G для численного решения уравнений диффузии. Для большей точности ∂a∂i принимает на вход три размера $pts = \{G_1, G_2, G_3\}$, вычисляет ожидаемые аллель-частотные спектры для каждого G_i, $i \in \{1, 2, 3\}$, а затем использует аппроксимацию Ричардсона [108] для генерации более точного спектра на втором шаге вычисления значения правдоподобия [45].

Начальная версия ∂a∂i была представлена в статье 2009 года [45]. Метод аппроксимации диффузией был ограничен тремя популяциями из-за вычислительной сложности, которая экспоненциальная от числа P популяций и равна $O(G^P)$, где G — число точек в сетке численного решения уравнения. Зависимость времени вычисления значения правдоподобия в ∂a∂i при разном числе популяций и размеров данных представлены на рисунке 31а. В 2020 году метод был расширен для учета коэффициентов инбридинга [51], а в 2021 году в ∂a∂i включили возможность использования GPU, что позволило добавить вывод демографической истории для четырех и пяти популяций [109].

В 2019 году был представлен *метод моментов* для вывода демографической истории популяций по аллель-частотному спектру $A^{\mathfrak{D}}$, реализованный

в программном обеспечении *moments* [55]. Этот метод основан на тех же математических моделях: на модели Райта-Фишера и бесконечного числа сайтов, как и метод аппроксимации диффузией в *∂a∂i*. Однако в методе, используемом в *moments*, прямое решение уравнения диффузии заменено линейной системой обычных дифференциальных уравнений для вычисления $\phi(x;T)$. Для решения этой системы *moments* использует *численный метод Кранка-Николсона* [110], а также *метод переменных направлений* [90] в случае более трех популяций.

Целевая функция оптимизации в *moments* также является логарифмом Пуассоновского правдоподобия, как и в *∂a∂i*:

$$f_{\mathcal{M}}^{moments}(\theta) = f_{\mathcal{M}}^{\partial a \partial i}(\theta) = \log \mathcal{L}(\theta|\mathfrak{D}).$$

Было продемонстрировано, что вычисление правдоподобия в *moments* требует меньше времени, однако менее точно, чем в *∂a∂i* [55]. В следствие скорости, *moments* поддерживает до пяти популяций, однако вывод демографической истории пяти популяций остается вычислительно сложной задачей. На рисунке 31б представлен график среднего времени вычисления правдоподобия с использованием *moments* в зависимости от числа популяций и размера данных.

Еще один популярный метод вывода демографической истории, использующий модель Морана и модель бесконечного числа сайтов, реализован в программном обеспечении *momi2* [46]. Модель Морана позволяет поколениям пересекаться, однако в пределе сходится к тому же процессу коалисценции, что и модель Райта-Фишера, это приводит к похожим результатам. В статье 2017 года [82] впервые был реализован метод непрерывной по времени модели Морана [100] в первой версии *momi*. Однако исходно метод не поддерживал ни непрерывные, ни единичные миграции. В 2019 году авторы расширили возможности и включили вывод единичных миграций [46]. Как и *∂a∂i* и *moments*, *momi2* вычисляет ожидаемый аллель-частотный спектр $A^{\mathcal{M}}$ по заданной модели $\mathcal{M}$ демографической истории с параметрами θ. Используя модель Пуассоновского случайного поля для числа мутировавших сайтов, предложена следующая функция для вычисления логарифма правдоподобия:

$$f_{\mathcal{M}}^{momi2}(\theta) = \sum_{\overline{d}} A^{\mathfrak{D}}[\overline{d}] \log\left(\sum_{\overline{d}} A^{\mathcal{M}}[\overline{d}]\right) - \sum_{\overline{d}} A^{\mathcal{M}}[\overline{d}] + \sum_{\overline{d}} A^{\mathfrak{D}}[\overline{d}] \log\left(\frac{A^{\mathcal{M}}[\overline{d}]}{\sum_{\overline{p}} A^{\mathcal{M}}[\overline{p}]}\right),$$

где $\overline{d} = (d_1, d_2, \ldots, d_P)$. Метод вычисления ожидаемого аллель-частотного спектра $A^{\mathcal{M}}$, а, следовательно, и метод вычисления правдоподобия, реализованный в *momi2*, значительно быстрее, чем в *∂a∂i* и *moments*. Это позволяет поддерживать большее число популяций. Например, в статье [46] была выведена демографическая история девяти популяций с использованием *momi2*. На рисунке 31в показана зависимость времени вычисления правдоподобия с использованием *momi2* от разного числа популяций и размеров данных.

Одновременно с развитием методов, использующих аллель-частотный спектр, происходило развитие методов, которые учитывали неравновесное сцепление генов. Например, в 2017 году метод аппроксимацией диффузией в *∂a∂i*

был модифицирован для использования двухлокусного гаплотип-частотного спектра, представленного в разделе 1.4.2 [95]. Однако этот метод применим только для вывода демографической истории одной популяции и не получил широкого применения. В 2019 и 2020 годах был разработан новый *метод моментов для множества двухлокусных статистик*, он был реализован как подмодуль *momentsLD* основного программного обеспечения *moments* [47]. В качестве предположений в методе используется модель Райта-Фишера с не пересекающимися поколениями. Для вычисления правдоподобия *momentsLD* вычисляет моменты для фиксированного множества двухлокусных статистик, используя уравнения Хила-Робертсона [79]. Например, в случае одной популяции для второго момента коэффициента неравномерного сцепления генов D, а также первых моментов статистики z и совместной гетерозиготности π_2 можно записать следующее рекурсивное уравнение Хила-Робертсона:

$$\mathbf{y} = \begin{pmatrix} \mathbb{E}[D^2] \\ \mathbb{E}[Dz] \\ \mathbb{E}[\pi_2] \end{pmatrix}$$

$$\mathbf{y}_{t+1} - \mathbf{y}_t = \left(\frac{1}{2g(t)} \begin{pmatrix} -3 & 1 & 1 \\ 4 & -5 & 0 \\ 0 & 1 & -2 \end{pmatrix} + r \begin{pmatrix} -2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right) \mathbf{y}_t,$$

где $g(t)$ — функция изменения численности популяции, r — вероятность рекомбинации. Это уравнение обобщается для более высоких моментов, а также для большего числа популяций с использованием таких статистик как $\mathbb{E}[D_i D_j]$, где D_i и D_j — коэффициенты неравномерного сцепления генов для популяций i иj соответственно [47]. Рекурсивное уравнение в *momentsLD* аппроксимируется дифференциальным, которое решается с применением *численного метода Кранка-Николсона* [110].

Для вычисления правдоподобия в *momentsLD* требуются набор генетических расстояний $r_1, r_2, \ldots, r_n$, каждое из которых определяет вероятность рекомбинации локусов, находящихся на этом расстоянии. Для последовательности пар $[r_i, r_{i+1}]$, $i \in \{1, 2, \ldots, n-1\}$ среди данных выбираются все пары локусов (pos_1^i, pos_2^i), которые располагаются на генетическом расстоянии r таком, что $r \in [r_i, r_{i+1}]$. Для выбранных пар вычисляется средние и дисперсии набора двухлокусных статистик, определенный числом популяций и уравнением Хила-Робертсона. Обозначим набор статистик для интервала $[r_i, r_{i+1}]$ за $\boldsymbol{\nu}_i^{\mathfrak{D}}$, средние значения, вычисленные для пар локусов, находящиеся на генетическом расстоянии $r \in [r_i, r_{i+1}]$, за $\hat{\boldsymbol{\nu}}_i^{\mathfrak{D}}$, а матрицу ковариаций как $\Sigma_i^{\mathfrak{D}}$. Пусть $\boldsymbol{\nu}_i^{\mathcal{M}}$ — ожидаемые статистики вычисленные по данной модели $\mathcal{M}$ демографической истории с параметрами θ с использованием уравнений Хилла-Робертсона. Значение правдоподобия для интервала $[r_i, r_{i+1}]$ вычисляется как вероятность наблюдать данные $\boldsymbol{\nu}_i^{\mathfrak{D}}$ при условии, что они распределены нормально со средним $\hat{\boldsymbol{\nu}}_i^{\mathfrak{D}}$ и матрицей

ковариаций $\Sigma_i^{\mathfrak{D}}$:

$$\mathcal{L}(\theta|\hat{\boldsymbol{\nu}}_i^{\mathfrak{D}}) = \mathcal{N}(\hat{\boldsymbol{\nu}}_i^{\mathfrak{D}}, \boldsymbol{\nu}_i^{\mathcal{M}}, \Sigma_i^{\mathfrak{D}}),$$

где $\mathcal{N}$ — плотность многомерного нормального распределения. Итоговое значение правдоподобия вычисляется как произведение величин $\mathcal{L}(\theta|\hat{\boldsymbol{\nu}}_i^{\mathfrak{D}})$, полученных для каждого интервала $[r_i, r_{i+1}],\ i \in \{1, 2, \ldots, n-1\}$:

$$\mathcal{L}(\theta) = \prod_{i=1,2,\ldots,n-1} \mathcal{L}(\theta|\hat{\boldsymbol{\nu}}_i^{\mathfrak{D}}).$$

Для численной стабильности в качестве целевой функции оптимизации *momentsLD* использует логарифм правдоподобия:

$$f_{\mathcal{M}}^{momentsLD}(\theta) = \log \mathcal{L}(\theta).$$

На рисунке 31г показана зависимость времени вычисления правдоподобия с использованием *momentsLD* от разного числа популяций. Заметим, что сложность метода является экспоненциальной, как в случае *∂a∂i* и *moments*, однако, в отличие от них, не зависит от размера данных.

1.5. Методы оптимизации для настройки параметров модели демографической истории популяций по генетическим данным

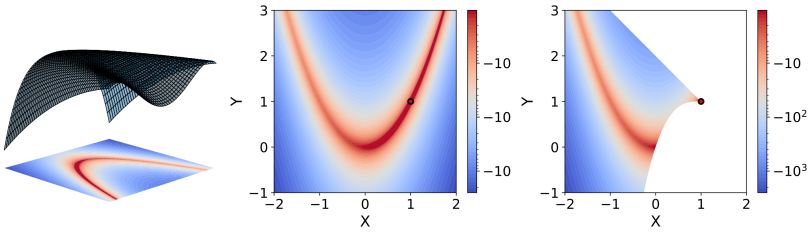
Методы оптимизации — это класс методов, которые решают задачи поиска экстремумов функций. Их суть состоит в поиске оптимального набора значений для функции f, которая описывает некоторую систему или процесс. Эта функция называется *целевой функцией* и может иметь множество входных параметров $\theta = \{\theta_i\}_{i=1}^N$, и задача оптимизации состоит в том, чтобы найти набор значений, который максимизирует или минимизирует эту функцию. Будем решать задачу максимизации целевой функции f:

$$\theta : f(\theta) \to \max .$$

Заметим, что задача минимизации функции g эквивалентна задаче максимизации функции $f(\theta) = -g(\theta)$.

Оптимизационные задачи могут быть классифицированы по нескольким критериям, например, по наличию ограничений (условные и безусловные), по числу переменных (одномерные и многомерные) и типу переменных (дискретные и непрерывные), по типу целевой функции (линейные и нелинейные) и другим параметрам.

Дискретные задачи оптимизации отличаются тем, что переменные принимают дискретные значения — набор из конечного или счётного числа возможных значений: $\exists i : \quad \theta_i \in \mathcal{D}$. *Непрерывные задачи оптимизации* имеют переменные, которые могут принимать любые значения из непрерывного интервала: $\theta \in \mathbb{R}^N$. *Одномерные задачи оптимизации* имеют только одну переменную ($N = 1$), в то



(а) (б) (в)

Рисунок 32 – Отрицательная функция Розенброка [111]

время как *многомерные* имеют несколько переменных ($N > 1$), которые нужно оптимизировать.

Кроме того, существуют задачи оптимизации с ограничениями, когда функционал f, который нужно оптимизировать, зависит не только от переменных θ, но и от дополнительных ограничений $\theta \in \Sigma$, которые должны быть удовлетворены. Такие задачи называются *задачами условной оптимизации*:

$$\begin{aligned} \theta: \quad & f(\theta) \to \max, \\ \text{при условии} \quad & \theta \in \Sigma. \end{aligned}$$

Безусловные задачи оптимизации не имеют ограничений на допустимые решения. Например, задача максимизации функции:

$$\theta : -(\theta - 1)^2 \to \max$$

является безусловной задачей оптимизации и имеет решение $\theta = 1$. При добавлении следующего неравенства:

$$\begin{aligned} \theta: \quad & -(\theta - 1)^2 \to \max, \\ \text{при условии} \quad & \theta \leqslant 0. \end{aligned}$$

задача преобразуется в условную оптимизацию с решением $\theta = 0$.

Линейные задачи оптимизации имеют линейную целевую функцию $f(\theta) = \sum_{i=1}^{N} c_i \cdot \theta_i = c_1\theta_1 + c_2\theta_2 + \cdots + c_N\theta_N$, а *нелинейные задачи* — нелинейную целевую функцию f.

На рисунке 32 представлена иллюстрация отрицательной функции Розенброка [111]. Функция Розенброка — это математическая функция, которая часто используется для тестирования оптимизационных методов. Функция была предложена в 1960 году Х. Розенброком, как пример нелинейной задачи оптимизации [111]. Эта функция определена на множестве двух и более переменных и имеет вид:

$$f(\theta) = f(x, y) = (a - x)^2 + b \cdot (y - x^2)^2 \quad \to \min,$$

где a и b — произвольные параметры. Глобальный минимум этой функции равен $f(1,1) = 0$, он находится внутри длинной узкой плоской области параболической формы. Найти эту область обычно не представляет сложности для методов, однако поиск глобального оптимума вызывает трудности, так как функция имеет множество локальных оптимумов.

На рисунке 32 представлена отрицательная функция Розенброка, которая задана следующей формулой:

$$f(\theta) = f(x,y) = -(1-x)^2 - 100 \cdot (y - x^2)^2 \quad \to \max .$$

На рисунке 32а изображено трехмерное представление поверхности, заданной функцией Розенброка, а также ее двухмерная проекция в виде контурного графика. Рисунок 32б иллюстрирует контурный график при постановке безусловной задачи оптимизации. Рисунок 32в представляет контурный график условной задачи оптимизации при следующих ограничениях:

$$\begin{aligned} (x-1)^3 - y + 1 &\leqslant 0, \\ x + y - 2 &\leqslant 0. \end{aligned}$$

Исключенная область из области определения θ изображена белым цветом на рисунке 32в. Точка оптимума функции Розенброка $\theta = (1,1)$ изображена красным цветом на рисунке.

Методы оптимизации могут быть детерминированными или стохастическими, локальными или глобальными, прямыми или с использованием градиента. *Детерминированные методы* используют стратегии пошагового улучшения решения, основанные на информации о производной функции, а *стохастические методы* ищут экстремумы, используя случайные процессы. *Локальные методы* находят экстремумы только в небольшой области пространства параметров, в то время как *глобальные методы* ищут экстремумы на всем пространстве параметров. *Прямые методы оптимизации* позволяют найти экстремум функции без использования градиента ∇f.

Классические методы оптимизации включают в себя методы, которые не используют машинное обучение и основаны на математических принципах и эвристических методах. Часто они являются методами локальной оптимизации. Некоторые из них включают в себя метод Ньютона, метод сопряженных градиентов, методы наискорейшего спуска и многие другие. Каждый метод имеет свои сильные и слабые стороны и может быть эффективным в зависимости от характеристик задачи.

Среди методов локальной оптимизации можно выделить несколько наиболее используемых. Например, метод BFGS (Broyden-Fletcher-Goldfarb-Shanno) [62–65] был разработан в 1970-х годах и назван в честь фамилий его создателей

— Ч. Бройдена, Р. Флетчера, Д. Голдфарба и Д. Шанно. Этот метод — итерационный метод численной оптимизации, который применяется для поиска экстремума нелинейных многомерных функций. Он относится к классу квазиньютоновских методов, которые основываются на приближенном вычислении гессиана целевой функции. Метод BFGS обеспечивает эффективное приближение гессиана функции и позволяет находить ее экстремумы, используя только значения функции и ее градиента. Он демонстрирует быструю сходимость и считается одним из наиболее эффективных методов оптимизации для большинства задач. Существует модифицированные версии метода: L-BFGS (Limited-memory BFGS) и L-BFGS-B (Limited-memory BFGS with bounds) — для решения безусловной и условных задач оптимизации соответственно при ограниченном размере доступной памяти [87].

Метод Нелдера-Мида [66], также известный как симплекс-метод, — это итерационный метод оптимизации без использования производных, который используется для поиска экстремума нелинейных, многомерных функций. Метод основывается на представлении функции как набора вершин симплекса, где каждая вершина представляет собой точку в пространстве параметров функции. Метод начинается с определения начального симплекса — набора вершин в пространстве параметров функции. Затем происходит итеративный процесс, в котором симплекс изменяется и перемещается по направлению к оптимальному значению функции. На каждой итерации происходит оценка значений функции в вершинах симплекса и выбор следующего симплекса на основе тех, которые дали наименьшие значения функции.

Метод Пауэлла [67] — это метод безусловной оптимизации, который разработал М. Пауэлл в 1964 году. Идея метода Пауэлла заключается в том, чтобы использовать направления, соответствующие осям координат и поворотам вокруг этих осей, для приближенного нахождения минимума функции. На каждой итерации метод определяет направление, в котором следует совершить шаг, путем решения подзадачи одномерной оптимизации. Затем он пересчитывает направления осей координат, чтобы учесть выполненный шаг, и повторяет процесс до достижения заданного критерия останова. Метод Пауэлла хорошо справляется с оптимизацией многомерных функций, не имеющих ограничений на переменные. Он также может использоваться для решения условной задачи, если ограничения параметров могут быть учтены при определении направлений осей координат.

Рисунок 33 демонстрирует использование классических методов оптимизации для поиска оптимума целевой функции которая задана следующей формулой:

$$f(\theta) = f(x,y) = 3(1-x)^2 \cdot e^{-(x^2)} - (y+1)^2 - 10\left(\frac{x}{5} - x^3 - y^5\right) \cdot e^{-x^2-y^2} - \frac{1}{3} \cdot e^{-(x+1)^2-y^2}$$

На рисунке 33а показано трехмерное представление поверхности, заданной функцией, ее двухмерная проекция в виде контурного графика, а также пока-

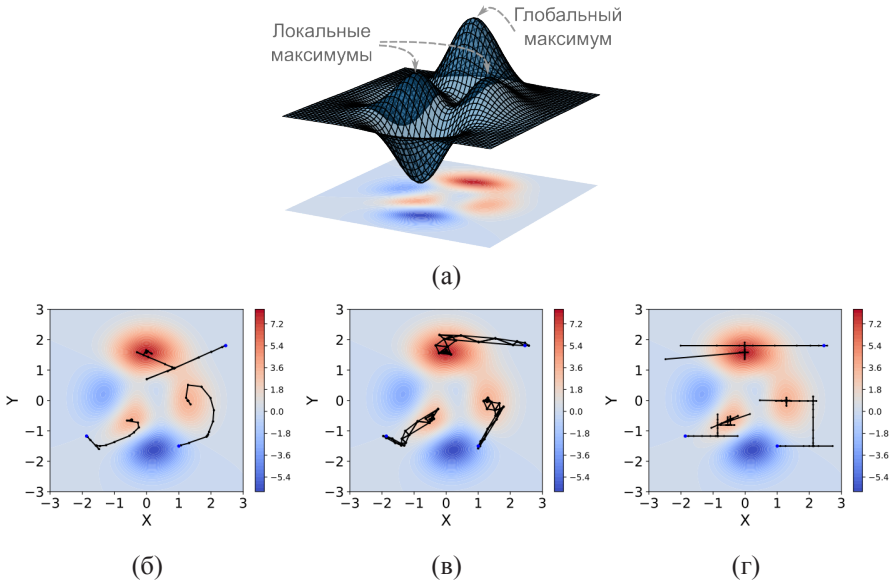


Рисунок 33 – Примеры работы методов локальной оптимизации при поиске оптимума функции, изображенной на рисунке (а): (б) метод BFGS, (в) метод Нелдера-Мида, (г) метод Пауэла.

заны глобальный и два локальных максимума. Три разных метода оптимизации — BFGS, метод Нелдера-Мида и метод Пауэлла — запущены из трех начальных точек, обозначенных синим цветом, и представлены соответственно на рисунках 33б, 33в и 33г.

На каждой итерации метод BFGS вычисляет градиент ∇f и делает шаг в сторону локального оптимума, именно это поведение можно увидеть на рисунке 33б). Метод Нелдера-Мида использует понятие симплекса — набора вершин. На каждой итерации метода формируется симплекс, образованный из $N + 1$ точек в N-мерном пространстве. Для двумерной задачи оптимизации, как на рисунке 33в, симплекс на каждой итерации состоит из трех вершин, образующих треугольники. Метод Пауэла использует направления базисных векторов и производит вычисление целевой функции f вдоль этих направлений. Рисунок 33г демонстрирует направления, используемые в методе Пауэла.

Классические методы оптимизации применяются для поиска минимумов и максимумов заданной целевой функции f. Однако, при оптимизации сложных и многомерных функций может быть множество локальных минимумов, что приводит к тому, что классические методы могут оказаться неэффективными и застрять в этих оптимумах. Для решения этой проблемы были разработаны глобальные методы оптимизации, которые могут искать глобальный экстремум

функции. Среди таких методов наиболее популярными являются методы, основанные на принципах эволюции, имитации отжига и методах роя частиц.

Для решения задачи настройки параметров модели $\mathcal{M}$ демографической истории по данным $\mathfrak{D}$ используются различные методы оптимизации, преимущественно, методы локальной оптимизации. Каждое из программных решений *дaдi*, *moments* и *momentsLD* включают в себя набор следующих методов, описанных ранее:

- Метод BFGS,
- Метод L-BFGS-B,
- Метод Нелдера-Мида,
- Метод Пауэлла.

Чтобы использовать эти методы, необходимо задать начальное приближенное решение, а затем метод осуществляет поиск локального оптимума целевой функции f. Однако для поиска глобального оптимума каждый метод следует запускать несколько раз для разных начальных приближений. Несмотря на то, что эти методы широко используются, ни выбор метода оптимизации, ни число его запусков, ни способ генерации начальных точек не стандартизованы и остаются на выбор пользователя [1, 45, 55]. Это приводит к потенциальным ошибкам, пониженной эффективности методов и ненадежности результатов [1].

В отличие от других программных решений, *momi2* реализует аналитическое вычисление градиента и, как следствие, имеет эффективный метод оптимизации — усеченный метод Ньютона (Truncated Newton Constrained) — для поиска оптимальных параметров модели демографической истории. Этот метод эффективен для оптимизации нелинейных функций с большим числом независимых непрерывных переменных с ограничениями. Усеченный метод Ньютона состоит в многократном применении итеративного метода Ньютона для поиска точки, в которой градиент целевой функции равен нулю. Однако метод усечен — выполняется только ограниченное число итераций.

В 2017 году был представлен новый метод оптимизации для поиска параметров модели демографической истории популяций в программном обеспечении *dadi pipeline* [49]. Как следует из названия, *dadi pipeline* был разработан как оболочка для *дaдi*, однако в 2019 году этот же метод был реализован для *moments* и получил название *moments pipeline* [50]. Метод оптимизации, реализованный в *dadi pipeline* и в *moments pipeline*, использует несколько последовательных запусков метода Нелдера-Мида. В качестве начального приближенного решения для первого запуска используются случайно сгенерированные параметры. Для каждого последующего запуска метода Нелдера-Мида используются текущие лучшие параметры, измененные случайным образом. С увеличением числа запусков изменение параметров ослабевает, что приводит к сходимости метода.

После предварительной нерецензируемой публикации статьи диссертанта [1], в которой были описаны проблемы методов локального оптимизации и предложено решение, авторы *дaдi* включили первый метод глобальной оптимизации BOBYQA (Bound Optimization BY Quadratic Approximation) [88], реализо-

ванный в библиотеке NLopt [112]. Метод BOBYQA решает задачу условной оптимизации без вычисления градиента. Метод имеет набор решений $\{\theta_1, \ldots, \theta_m\}$ заданного размера m для целевой функции f. На каждой итерации происходит построение квадратичной аппроксимации Q для целевой функции f по точкам $\{\theta_1, \ldots, \theta_m\}$. На основе квадратичной аппроксимации Q выбирается новая точка $\bar{\theta}$. Если значение $f(\bar{\theta})$ оказывается лучше значения текущего оптимума θ^*, то $\bar{\theta}$ заменяет θ^* в поддерживаемом наборе точек. Этот метод был впервые применен для поиска параметров модели демографической истории для популяций с инбридингом в статье 2020 года [51].

Можно выделить следующие недостатки использования существующих программных средств и методов для настройки значений параметров модели демографической истории:

- пользователю требуется задавать и проверять каждую модель вручную с использованием выбранной библиотеки;
- методы оптимизации ограничены выводом значений только для непрерывных параметров;
- требуют вовлечения пользователя для задания начальных значений параметров или выбора числа запусков локальной оптимизации;
- использование методов локальной оптимизации не гарантирует нахождение глобального оптимума.

1.6. Методы перебора моделей демографической истории

Для корректного вывода демографической истории популяций необходимо выбрать модель, которая максимально точно описывает процессы эволюции, произошедшие в прошлом с популяциями. Однако при выборе модели необходимо учитывать возможность переобучения, когда модель слишком сложна и точно подстраивается под имеющиеся данные, но плохо обобщает и не учитывает некоторые важные особенности демографической истории. Поэтому необходимо выбирать модель, которая достаточно проста, чтобы избежать переобучения, и при этом имеет хорошую обобщающую способность. Существуют различные методы выбора и поиска наилучшей модели демографической истории. Классический подход использует перебор моделей и выбор наилучшей на основе определенного критерия, такого, например, как информационный критерий Акаике. Однако он имеет ряд недостатков. Во-первых, пространство возможных моделей часто слишком велико, и исследователь обычно рассматривает только его часть исходя из своего опыта и ожиданий и может пропустить наилучшую модель. Во-вторых, если метод оптимизации, используемый для поиска оптимальных параметров моделей, оказывается недостаточно эффективным, то неверная модель может быть выбрана в качестве наилучшей. До текущей работы диссертанта не существовало автоматического метода поиска наилучшей модели демографической истории популяций.

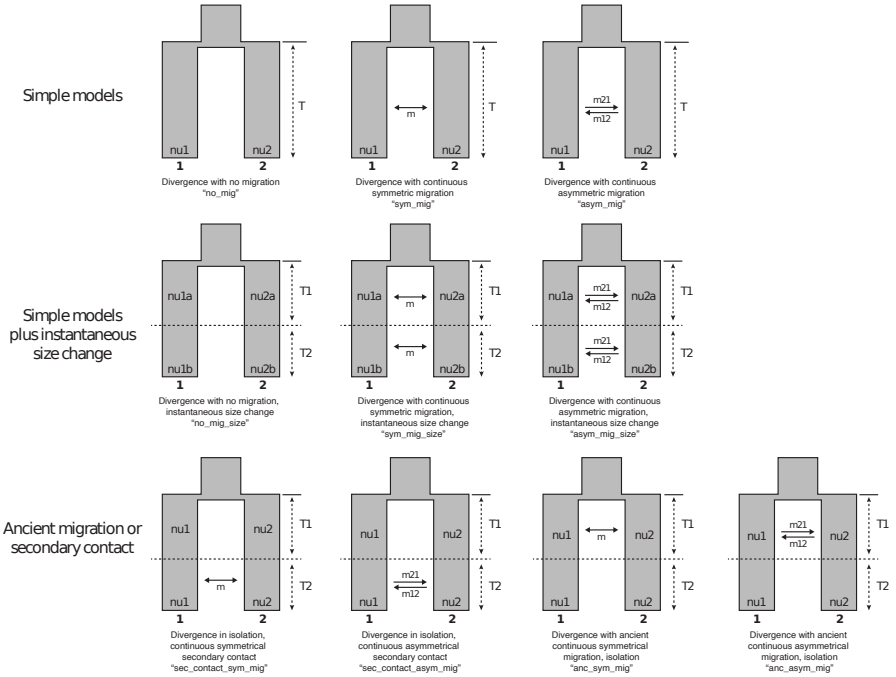


Рисунок 34 – Пример некоторых моделей из каталога *dadi pipeline*. Источник: [49]

Существует каталог моделей для демографической истории двух и трех популяций в программных обеспечениях *dadi pipeline* и *moments pipeline*. Всего в каталоге представлено 32 модели для двух популяций и 33 модели для трех популяций. Модели отличаются наличием миграций, структурой разделения популяций, типами изменения численности и числом эпох. На рисунке 34 изображены примеры 10 моделей из каталога. Рекомендуется тестировать только некоторое подмножество моделей из каталога, так как некоторые из них были созданы для конкретных проектов и имеют смысл только в определенных биологических контекстах [49].

Для выбора наилучшей модели *dadi pipeline* и *moments pipeline* используют информационный критерий AIC, описанный в разделе 1.3.3. Однако, AIC применим только в случае, если вариабельные позиции, использованные для построения аллель-частотного спектра, независимы. В противном случае требуются другие метрики сравнения моделей, которые не реализованы в этих программных решениях. В работе [49] *dadi pipeline* использовался для вывода параметров и сравнения двенадцати моделей демографической истории для данных двух популяций лягушки *Scotobleps gabonicus*.

В статье [54] представлено программное обеспечение для перебора и поиска наилучшей модели демографической истории двух популяций с использованием *moments*. Каталог насчитывает восемь основных моделей и 100 дополнительных моделей. Как и в методе выбора модели в *dadi pipeline* и *moments pipeline* предполагается независимость вариабельных позиций, которые были использованы для построения аллель-частотного спектра, что позволяет применять информационный критерий Акаике для выбора наилучшей модели. Используя процедуру бутстрапа [113], строится новое множество наборов данных для исходных генетических данных. Размер этого множества выбирается пользователем и обычно не очень большой [54]. Для каждой модели из каталога и построенного набора данных осуществляется поиск оптимальных параметров и вычисление значения AIC. Для выбора лучшей модели используется медиана значений AIC, полученных для множества наборов данных. Во второй версии программного обеспечения, представленного в работе [54], в качестве метода оптимизации для поиска параметров моделей был использован метод, основанный на генетическом алгоритме, который был разработан диссертантом в этой работе.

Существующие программные решения, обеспечивающие перебор моделей демографической истории, имеют следующие недостатки:

- перебор моделей ограничен каталогом;
- сравнение моделей выполняется с использованием информационного критерия Акаике и предполагает независимость данных;
- единственный метод автоматического перебора позволяет автоматически перебрать модели только для демографической истории двух популяций.

Функции правдоподобия, которые используются для поиска параметров модели демографической истории, обычно являются произведением нескольких вероятностей независимых событий. Например, для *∂a∂i* и *moments* функция правдоподобия предполагает независимость элементов аллель-частотного спектра, а, следовательно, и независимость вариабельных позиций генетических данных. Если предположение о независимости выполнено, то выбор модели может быть осуществлен с помощью теста отношения правдоподобия (LRT), информационного критерия Акаике или байесовского информационного критерия. Однако если события зависимы, эти подходы будут ошибочно отдавать предпочтение моделям с большим числом параметров [68]. Этих погрешностей можно избежать, выполнив оценку максимального правдоподобия на множестве наборов данных, построенного процедурой бутстрапа [113], однако это требует значительных вычислительных затрат. В статье [69] предложена корректировка информационного критерия Акаике и теста отношения правдоподобия в случае зависимостей в данных. Для описания этих методов рассмотрим следующие матрицы J и H для функции правдоподобия $\mathcal{L}$ и параметров θ:

$$J(\theta) = E_\theta \left\{ \frac{\partial \mathcal{L}(\theta|\mathfrak{D})}{\partial \theta} \left(\frac{\partial \mathcal{L}(\theta|\mathfrak{D})}{\partial \theta} \right)^T \right\},$$

$$H(\theta) = E_\theta \left\{ -\frac{\partial^2}{\partial\theta\partial\theta^T}\mathcal{L}(\theta|\mathfrak{D}) \right\}.$$

Матрица $J(\theta)$ является матрицей ковариации, а матрица $H(\theta)$ — гессианом. Авторы статьи [69] предлагают использовать обычную аппроксимацию матрицы Гессиана [114] для вычисления $H(\theta)$ и метод, представленный в [115], для оценки матрицы ковариаций $J(\theta)$. Этот подход реализован в программных пакетах *∂a∂*i и *moments*. Однако для оценки J требуется использование целого набора данных, полученного с помощью процедуры бутстрапа из исходных данных. При этом процедура бутстрапа должна быть выполнена для независимых участков генома, которые могут включать позиции, мутации в которых взаимосвязаны, однако мутации на разных участках должны быть независимыми. Для генерации новых данных с помощью процедуры бутстрапа происходит случайный выбор с повторениями выделенных независимых участков.

CLAIC (Composite Likelihood Akaike Information Criterion) — это модификация критерия AIC для случая зависимых данных [69, 116]. Он вычисляется по формуле:

$$\text{CLAIC}(\mathcal{M}, \mathfrak{D}) = 2\dot{t}r(J(\theta^*)H^{-1}(\theta^*))-2 \cdot \log(\mathcal{L}(\theta^*|\mathfrak{D})),$$

где θ^* — значения параметров модели $\mathcal{M}$, которые обеспечивают максимальное значение правдоподобия $\mathcal{L}(\theta|\mathfrak{D})$ для данных $\mathfrak{D}$. Чем меньше значение CLAIC, тем лучше модель соответствует данным.

Тест отношения правдоподобия используется для сравнения двух моделей $\mathcal{M}_{full}$ и $\mathcal{M}_{nested}$, где одна из них включает в себя другую модель $\theta_{full} = \theta_{nested} \cup \psi$. В разделе 1.3.3 была описана статистика λ_{LRT}, которая используется для этого сравнения. В случае зависимых данных была предложена корректировка этой статистики [69, 117]:

$$\lambda_{LRT}^{adj} = \frac{\lambda_{LRT}}{\mu(\theta_{full})},$$

где $\mu(\theta_{full}) = tr(J(\theta_\psi)H(\theta_\psi)^{-1}/d)$. Здесь $H(\theta_\psi)$ и $J(\theta_\psi)$ обозначают подмножества матриц $H(\theta_{full})$ и $J(\theta_{full})$, соответствующие параметрам ψ во вложенной модели $\mathcal{M}_{nested}$, которые были зафиксированы на значениях $\psi_i = C_i$, $i = 1, 2, \ldots d$. Для определения, является ли вложенная модель $\mathcal{M}_{nested}$ лучше более сложной модели $\mathcal{M}_{full}$, проверяется гипотеза о том, что статистика λ_{LRT}^{adj} имеет распределение хи-квадрат χ^2.

Выводы по граве 1

1. Существующие методы вывода демографической истории по генетическим данным решают задачу поиска параметров заданной модели демографической истории, обеспечивающих максимальное значение правдоподобия для генетических данных.

2. Используемые модели демографической истории имеют только непрерывные параметры и фиксированные динамики изменения численности популяций (константный, линейный или экспоненциальный законы изменения численности).
3. При использовании нескольких программных решений требуется задавать одну и ту же модель для каждого из них отдельно, так как существующие решения имеют различные интерфейсы спецификации моделей, которые нельзя переиспользовать.
4. Методы вычисления правдоподобия являются методами численного имитационного моделирования.
5. Для настройки параметров используются методы локальной оптимизации, которые не гарантируют нахождения глобального оптимума.
6. Существующие методы перебора моделей используют для сравнения и выбора наилучшей модели информационный критерий Акаике, который предполагает независимость данных.
7. На момент начала исследований не существовало метода автоматического перебора моделей. Единственный альтернативный метод, имеющий ряд существенных ограничений, появился после публикации статьи диссертанта [1].

Глава 2. Расширенный класс моделей демографической истории популяций и методы настройки параметров моделей по генетическим данным

В данной главе приведено описание разработанного расширенного класса моделей демографической истории, а также методов настройки параметров моделей по генетическим данным.

Раздел 2.1 содержит описание моделей расширенного класса, которые могут включать не только непрерывные, но и дискретные параметры динамики изменения численности популяций.

Метод, основанный на комбинации генетического алгоритма и локального поиска, для настройки параметров модели демографической истории по генетическим данным описан в разделе 2.2. Гиперпараметры метода были настроены автоматически для эффективного решения поставленной задачи.

В разделе 2.3 представлен метод настройки параметров в условиях сложновычислимой функции, основанный на комбинации ансамблевой байесовской оптимизации и локального поиска. Разработанный метод имеет настроенные гиперпараметры и эффективен для вывода демографической истории четырех и пяти популяций.

Разделы 2.4 и 2.5 содержат описание и результаты экспериментальных исследований разработанных методов, основанных на генетическом алгоритме и байесовской оптимизации, для настройки параметров моделей, включая модели расширенного класса.

2.1. Расширенный класс моделей демографической истории популяций

Для упрощения работы пользователя-биоинформатика был разработан расширенный класс моделей.

Эти модели включают новый тип параметров для вывода — динамики изменения численности. При этом закон изменения численности в модели теперь может быть не фиксирован, а задан дискретным параметром, и его значение можно найти методом оптимизации.

Приведем пример модели с параметром нового типа. Пусть модель на рисунке 14а имеет дополнительный параметр `Dyn` — динамика изменения второй популяции после разделения. При разных значениях этого параметра численность второй популяции будет либо константная, либо будет иметь линейный или экспоненциальный закон изменения. Изображение предложенной модели, а также демографические истории при разных значениях параметра `Dyn` показаны на рисунке 35.

В качестве прототипа расширенного класса моделей был выбран первый класс, в котором модели описываются временными интервалами и разделениями популяций. Этот класс имеет больше преимуществ по сравнению со вторым

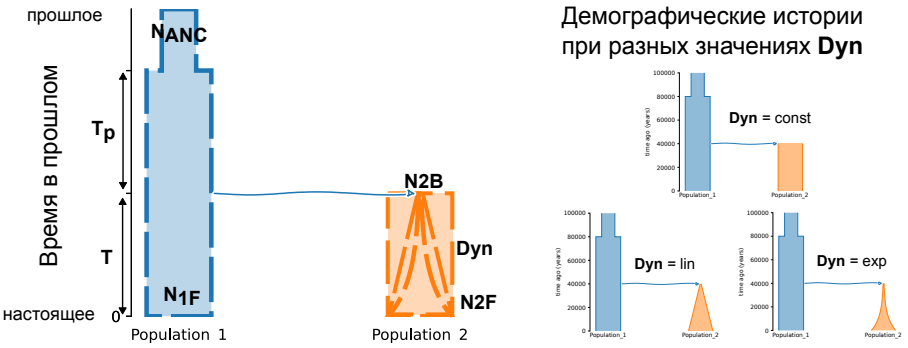


Рисунок 35 – Пример разработанной расширенной модели демографической истории с параметрами и соответствующие ей демографические истории при разных значениях параметра `Dyn`

классом, так как модели в нем могут описывать линейное изменение численности.

Определение 21. Динамическими характеристиками $\chi_{dyn}(\mathcal{I})$ временного интервала $\mathcal{I} = \langle p, T, \mathfrak{N}^{start}, \mathfrak{N}^{end}, \mathfrak{d} \rangle$ называется множество $\{d_1, \ldots, d_p\}$, которое соответствует набору динамик временного интервала.

Определение 22. Модель расширенного класса для демографической истории P популяций — параметрическая модель для демографической истории P популяций, которая представляется в виде шестерки $\langle \Theta, \Theta_d, \mathcal{E}, \mathfrak{F}, \mathfrak{F}_d \rangle$, где $\Theta \subset \mathbb{R}_+^{k_1}$ — множество значений непрерывных параметров модели, $\Theta_d \subset \{0, 1, 2\}^{k_2}$ — множество значений дискретных параметров динамики, $\mathcal{E} = \{E_i\}_{i=1}^{K}$, $E_i \in \mathcal{I} \cup \mathcal{A} \cup \mathcal{S}$ — последовательность элементов временных интервалов, единичных миграций и разделений, $\mathfrak{F} : \Theta \to \bigcup \chi(E_i)$ — отображение непрерывных параметров модели в характеристики элементов, $\mathfrak{F}_d : \Theta_d \to \bigcup \chi_{dyn}(I_i)$, $I_i \in \mathcal{E} \cap I$ — отображение дискретных параметров динамики в динамические характеристики элементов временных интервалов.

Рисунок 36 изображает представление разработанной модели расширенного класса в виде шестерки $M = \langle \Theta, \Theta_d, \mathcal{E}, \mathfrak{F}, \mathfrak{F}_d \rangle$.

Таким образом, при использовании предлагаемой модели появляется возможность запуска метода оптимизации для определения оптимальных законов изменения численности для любого временного интервала (в нашем примере только для второй популяции). Это позволяет пользователю не перебирать разные модели с разными динамиками вручную, как это требуется для *∂a∂i*, *moments*, *momentsLD* и *momi2*, а сделать это автоматически.

Реализация моделей расширенного класса была проведена в два этапа. На первом этапе были спроектированы и реализованы классы для параметров-переменных моделей. Затем были спроектированы и реализованы классы мо-

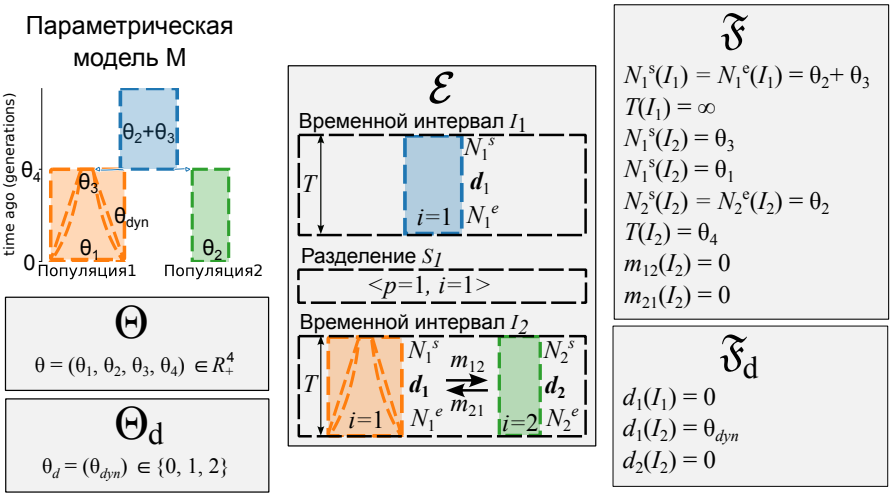


Рисунок 36 – Пример модели $M = \langle \Theta, \Theta_d, \mathcal{E}, \mathfrak{F}, \mathfrak{F}_d \rangle$ расширенного класса

делей. В результате были реализованы два модуля на языке программирования Python: `variables` и `models`.

Модуль `variables` содержит классы параметров моделей. Структура этих классов представлена на рисунке 37. Абстрактный класс `Variable` представляет произвольный параметр с областью определения `domain` и процедурой `resample()` генерации случайного значения. У каждого параметра обязательно присутствует имя `name`. От `Variable` наследуются три абстрактных класса: 1) абстрактный класс `ContnuousVariable` непрерывных параметров, 2) абстрактный класс `DiscreteVariable` дискретных параметров, 3) абстрактный класс `DemographicVariable` параметров моделей демографической истории. Модуль содержит шесть не абстрактных классов, которые соответствуют следующим параметрам моделей демографической истории:

- класс `PopulationSizeVariable` — параметры численности популяции;
- класс `TimeVariable` — параметры времени;
- класс `FractionVariable` — параметры доли, например, параметр доли особей, совершивших единичную миграцию или параметр доли численности, которая формирует новую популяцию при разделении;
- класс `GrowthRateVariable` — параметры степени экспоненциального изменения, которые обычно используются в моделях второго класса;
- класс `MigrationVariable` — параметры темпа непрерывной миграции;

– класс DynamicVariable — параметры динамики изменения численности популяций.

Каждый из этих классов является наследником абстрактного класса DemographicVariable и одного из классов ContnuousVariable или DiscreteVariable.

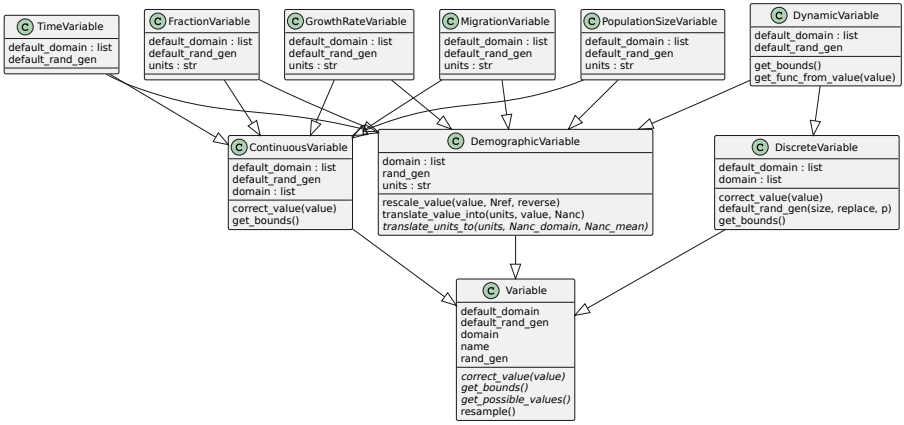


Рисунок 37 – Структура классов разработанного модуля variables

Модуль models содержит множество классов для разных моделей. Структура этих классов представлена на рисунке 38. Основным классом является абстрактный класс Model, который описывает произвольный объект модели с параметрами. Основными процедурами класса являются процедура add_variable добавления параметров класса Variable и процедура fix_variable фиксирования значения указанного параметра. Можно выделить три группы классов, которые наследуются от Model.

Первая группа содержит реализацию классов для комбинаций переменных (общий абстрактный класс VariablesCombination). Он включает в себя бинарные (абстрактный класс BinaryOperation) и унарные (абстрактный класс UnaryOperation) операции для параметров класса Variable. Среди бинарных операций реализованы операции сложения, деления, умножения, возведения в степень и вычитания параметров. Были реализованы две унарные операции: логарифмирование параметра и возведение экспоненты в степень параметра.

Вторая группа классов в модуле models — элементы событий (общий абстрактный класс Event) в моделях демографической истории. Она включает:

– класс Epoch, соответствующий элементам временных интервалов в моделях расширенного класса. Он содержит характеристики продолжительности (time_arg, начальной и конечной численности популяций

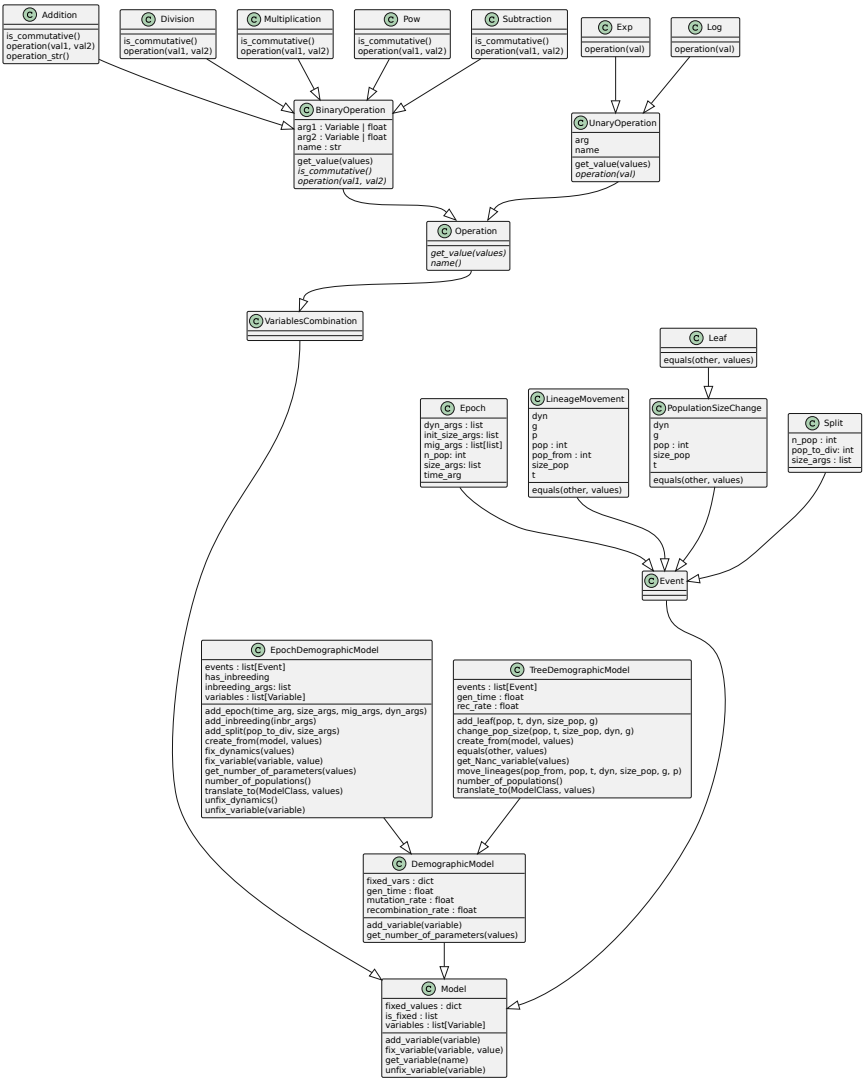


Рисунок 38 – Структура классов разработанного модуля `models`

(`init_size_args` и `size_args`), темпов непрерывных миграций (`mig_args`);

– класс `Split`, соответствующий элементам разделения в моделях расширенного класса;

– класс `PopulationSizeChange`, соответствующий событиям изменения численности в моделях второго класса;
– класс `LineageMovement`, соответствующий событиям единичной миграции в моделях второго класса;

Каждая характеристика описанных классов может быть:

а) параметром — объектом класса `Variable`;
б) комбинацией параметров — объектом класса `VariableCombination`;
в) константой — объектом типа `str` для характеристик динамики или объектом типа `float` в случае других характеристик.

Третья группа классов модуля `models` — классы моделей демографической истории. Объекты абстрактного класса `DemographicModel` содержат основные характеристики рассматриваемых популяций: среднее время одного поколения (`gen_time`), вероятность мутации одной позиции генома на одно поколение (`mutation_rate`), вероятность рекомбинации позиций генома, находящихся на расстоянии в один миллион пар оснований (`recombination_rate`). У этого класса реализованы два наследника: класс `EpochDemographicModel` и класс `TreeDemographicModel`.

Класс `EpochDemographicModel` **реализует разработанные модели расширенного класса**. Заметим, что он также реализует модели первого класса, так как первый класс вложен в расширенный класс моделей. Объект класса `EpochDemographicModel` имеет список `events` элементов временных интервалов — объекты класса `Epoch`, и разделений — объекты класса `Split`. Были реализованы четыре основные процедуры: 1) процедура `add_epoch` добавления элемента временного интервала, 2) процедура `add_split` добавления элемента разделения, 3) процедура `add_pulse_migration` добавления элемента единичной миграции, 4) процедура `add_inbreeding` добавления характеристик инбридинга популяций. Пример задания модели расширенного класса, представленной на рисунке 35, показан на рисунке 39.

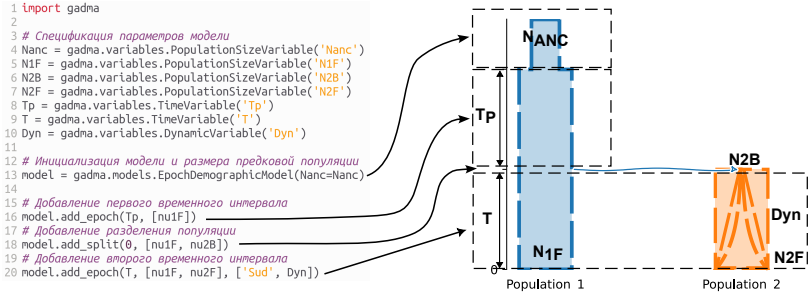


Рисунок 39 – Пример задания расширенной модели

Класс `TreeDemographicModel` реализует модели второго класса. Он содержит список `events` событий изменения численности — объектов клас-

са `PopulationSplit`, и событий единичной миграции — объектов класса `LineageMovement`.

Классы `EpochDemographicModel` и `TreeDemographicModel` были реализованы вместе с процедурой `translate_to`, которая переводит модель из одного класса в другой при заданных значениях параметров. Таким образом, реализованную модель расширенного класса при фиксированных значениях параметров можно перевести в модель второго класса для использования в библиотеке *momi2*.

2.2. Метод на основе комбинации генетического алгоритма и локального поиска для настройки параметров моделей демографической истории популяций по генетическим данным

Эволюционные алгоритмы — это класс алгоритмов оптимизации, основанных на биологической эволюции и ее принципах. Они используют процессы, которые происходят в природе, такие как мутация, скрещивание и естественный отбор, для решения задач оптимизации.

Один из наиболее известных и широко используемых эволюционных алгоритмов — *генетический алгоритм* [118]. Он моделирует процесс естественного отбора в биологии и применяется для поиска наилучшего решения в пространстве поиска. Генетический алгоритм начинается с определения первого *поколения* X^{init} размера N_{gen}, которое состоит из N_{gen} *особей*, представляющих потенциальные решения задачи. Каждая особь в генетическом алгоритме представляет собой вектор $x_i = (x_{i,1}, \cdots, x_{i,N})$ параметров целевой функции f, и этот вектор x_i называется геномом. Каждый параметр $x_{i,j}$ этого вектора — ген, и он соответствует определенной характеристике решения, которую нужно оптимизировать. Первое поколение $X^{\text{init}} = \{x_i^0\}_{i=1}^K$ особей в генетическом алгоритме может быть сгенерировано случайным образом. Затем происходит итерационный процесс, в котором каждое новое поколение X_t создается на основе предыдущего X_{t-1}, используя операторы генетических операций таких как мутация, скрещивание и отбор.

Оператор *мутации* в генетическом алгоритме — это случайное изменение генов особи (решения) с некоторой вероятностью. Мутация может произойти в любом месте генома особи, и это может привести к изменению соответствующей характеристики решения. Существует несколько способов реализации мутации в генетическом алгоритме, но наиболее распространенный — это замена значения случайно выбранного гена $x_{i,j}$ на другое случайно выбранное значение из допустимого диапазона. Оператор *скрещивания* в генетическом алгоритме представляет собой процесс создания новых особей путем комбинирования генетического материала двух родительских особей. Самый простой и широко используемый вид скрещивания — это одноточечное скрещивание (single-point crossover), когда случайно выбирается точка раздела между генами в родительских геномах, и гены до этой точки копируются из первого родителя, а после

этой точки — из второго. Другой вид скрещивания — равномерное (uniform), обозначает выбор каждого гена равновероятно между соответствующими генами родителей. Процесс *отбора* в генетическом алгоритме заключается в выборе наиболее приспособленных особей для создания следующего поколения. Особи оцениваются по значению *функции приспособленности* (fitness), которая отражает качество решения, представленного данным геномом. Часто значение приспособленности определяется целевой функцией f. В результате каждой итерации, находится новое поколение решений, которое более близко к оптимальному решению.

Генетический алгоритм применяется для решения различных задач биоинформатики, например, для поиска генов и аннотации генома [119], выравнивания нескольких последовательностей [120], моделирования структуры белков [121], поиска новых молекул лекарств [122] и многого другого.

Также существуют методы, которые сочетают в себе несколько различных подходов, они называются *комбинированными методами оптимизации*. Например, генетический алгоритм может быть применен в сочетании с методом локального поиска для улучшения точности оптимизации. В работе [123] такой комбинированный метод был разработан для поиска связей между мутациями в генах и болезнями.

2.2.1. Разработка метода на основе комбинации генетического алгоритма и локального поиска

В этом разделе описан разработанный метод для настройки параметров модели демографической истории популяций по генетическим данным.

Это комбинированный метод, основанный на совместном использовании генетического алгоритма и метода локальной оптимизации, для решения задачи настройки параметров θ модели $\mathcal{M}$ демографической истории популяций по генетическим данным. Метод представляет последовательный запуск генетического алгоритма и метода локальной оптимизации. Генетический алгоритм был модифицирован для оптимизации как непрерывных параметров целевой функции, так и дискретных, что позволяет применять его для настройки параметров разработанных моделей расширенного класса. В качестве методов локальной оптимизации в разработанном комбинированном методе предлагается применять методы BFGS, L-BFGS-B, Пауэлла или Нелдера-Мида, которые используются в существующих методах настройки параметров моделей первого и второго класса [45]. Эти методы позволяют настраивать только непрерывные параметры, поэтому при их применении значения дискретных параметров моделей расширенного класса фиксируются. Таким образом, разработанный комбинированный метод, основанный на генетическом алгоритме, сначала производит настройку всех параметров заданной модели с использованием генетического алгоритма, а затем проводит дополнительную настройку непрерывных параметров с помощью выбранного метода локальной оптимизации.

Приведем подробное описание примененного генетического алгоритма. Описание рассматриваемых методов локальной оптимизации было дано ранее в разделе 1.5.

Геномом особи в генетическом алгоритме является вектор значений параметров θ заданной модели $\mathcal{M}$, а функция приспособленности равна целевой функцией $f_{\mathcal{M}}(\theta)$, которая является логарифмом правдоподобия $\log(\mathcal{L}(\theta|\mathfrak{D}))$.

На первом этапе генетический алгоритм генерирует поколение особей, случайным образом, если оно не было задано заранее. Для формирования нового поколения отбираются наиболее приспособленные особи — с наибольшими значениями функции приспособленности, среди набора мутировавших, скрещенных и случайных особей. Выбор особей для применения операторов мутации или кроссовера является случайным, но вероятность выбора прямо пропорциональна значению приспособленности: чем лучше приспособленность вектора параметров, тем больше вероятность того, что он будет выбран. Генетический алгоритм останавливается, когда он больше не может получить лучшее решение по значению функции приспособленности за несколько итераций.

Схема алгоритма метода, основанного на комбинации генетического алгоритма и локального поиска, представлена на рисунке 40.

Псевдокод разработанного генетического алгоритма представлен в листинге 1. На вход алгоритм принимает следующие параметры: целевая функция f, начальный набор решений X^{init}, число лучших решений N_{E} предыдущего поколения в новом, число мутировавших решений N_{M}, число потомков скрещенных решений N_{C}, число случайно-сгенерированных решений N_{R} в новом поколении, максимальное число итераций без улучшения $T^{\max}_{\text{noImpr}}$, сила миграции μ_{s}, степень миграции μ_{r} и константы $C_{\mu_{\text{s}}}, C_{\mu_{\text{r}}}$, определяющие их изменения.

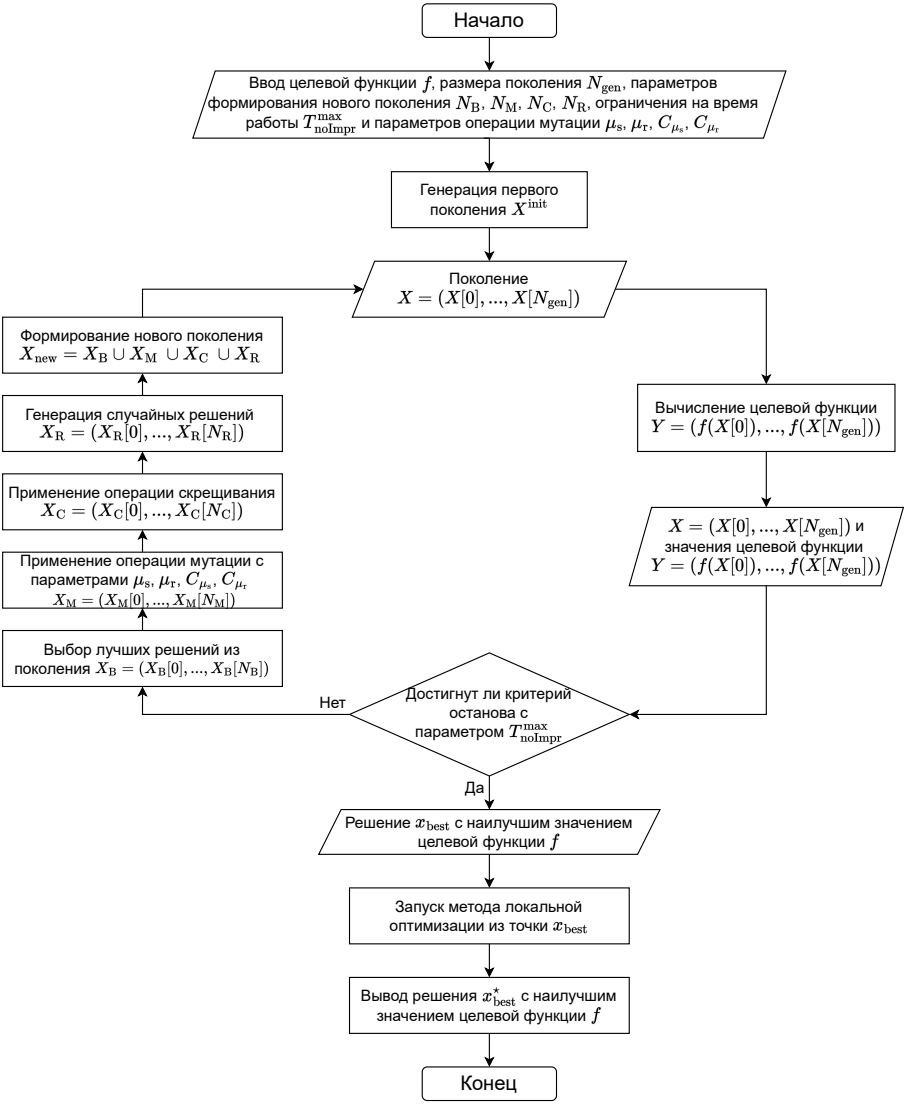


Рисунок 40 – Схема алгоритма метода, основанного на комбинации генетического алгоритма и локального поиска

Листинг 1 – Псевдокод разработанного генетического алгоритма

```
function GeneticAlgorithm(f, X^init, N_E, N_M, N_C, N_R, T^max_noImpr, μ_s, μ_r, C_μs, C_μr)
    N_gen ← (N_B + N_M + N_C + N_R)                          ▷ Размер поколения
    Y^init ← {f(θ), θ ∈ X^init}
    X, Y ← Selection(X^init, Y^init, N_gen)                  ▷ Первое поколение решений
    T_noImpr ← 0                                             ▷ Счетчик числа итераций без улучшения
    while T_noImpr ⩽ T^max_noImpr do
        X_new ← [ ]                                          ▷ Строим новое поколение решений
        ω ← { Y[i] / Σ_j Y[j] }_{i=1}^{N_gen}                ▷ Вероятности выбора решений
        for l ← 1..N_B do                                    ▷ Добавляем лучшие модели
            X_new.Add(X[l])
        for l ← 1..N_M do                                    ▷ Добавляем мутированные решения
            j ← DiscreteRandom({i}_{i=1}^N, ω)
            X_new.Add(Mutate(X[j], μ_s, μ_r))
        for l ← 1..N_C do                                    ▷ Добавляем скрещенные решения
            j_1 ← DiscreteRandom({i}_{i=1}^N, ω)
            j_2 ← DiscreteRandom({i}_{i=1}^N, ω)
            X_new.Add(CrossOver(X[j_1], X[j_2]))
        for k ← 1..N_R do                                    ▷ Добавляем случайные решения
            X_new.Add(GenerateRandomParameters( ))
        Y^new ← {f(x), x ∈ X^new}
        X_new ←Selection(X_new, Y_new, N_gen)
        if f(X_new[0]) > f(X[0]) then                        ▷ Обновление счетчика
            T_noImpr ← 0
            b_improved ← True
            b_improvedByMutation ← (X_new[0].lastOperation == "Mutation")
        else
            T_noImpr ← T_noImpr + 1
            b_improved ← False
            b_improvedByMutation ← False
        X ← X_new,  Y ← Y_new
                                 ▷ Обновление адаптивной силы и степени мутации
        μ_s ← UpdateValue(b_improvedByMutation, μ_s, C_μs)
        μ_r ← UpdateValue(b_improved, μ_r, C_μr)
    return X[0]
```

Мутация особи в генетическом алгоритме равнозначна процессу изменения значений отдельных генов. На рисунке 41 показано применение операции мутации для изменения одного гена — параметра заданной модели демографической истории. Разработанный оператор определяется двумя константами: *силой* μ_s и *степенью* μ_r мутации. Число изменяемых генов-параметров выбирается из биномиального распределения $\sim B(n, \mu_s)$ со средним значением, равным силе мутации. Гены, которые мутируют, выбираются с вероятностью, прямо пропорциональной их весам ω, которые вначале равны — выбор равновероятен, а затем каждый вес может быть увеличен, если произошла мутация соответствующего гена, которая привела к улучшению модели.

Разработанный генетический алгоритм позволяет работать как с непрерывными параметрами, так и с дискретными. При мутации гена, который соответствует непрерывному параметру, мера того насколько его значение изменится определяется знаком: -1 или $+1$ равновероятно, и абсолютной мерой λ, которая случайным образом генерируется из усеченного на отрезке $[0, 1]$ нормального распределения $\lambda \sim \hat{N}(\mu_r, \sigma^2, 0, 1)$ со средним значением, равным силе мутаций и дисперсией σ^2. Дисперсия σ^2 определяется специальной процедурой SigmaBasedOnThreeSigmaRule, описание которой приведено далее.Если происходит мутация гена, который соответствует дискретному параметру, то его значение изменяется равновероятно на любое другое возможное. Псевдокод для процедуры мутации разработанного генетического алгоритма в случае, когда параметры модели непрерывны, представлен в листинге 2.

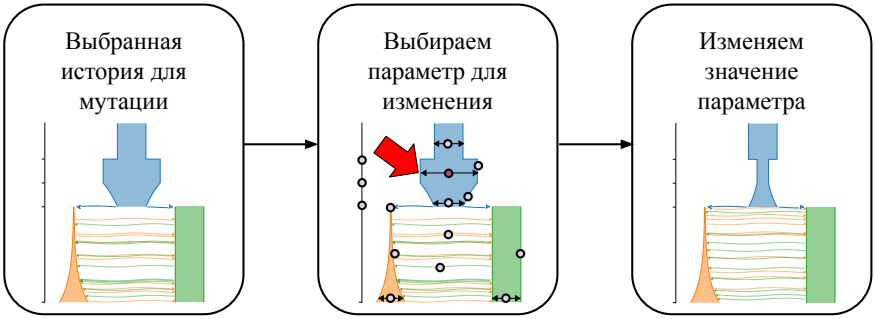


Рисунок 41 – Пример применения оператора мутации разработанного генетического алгоритма

Листинг 2 – Псевдокод оператора мутации разработанного генетического алгоритма. На вход подается целевая функция f, особь x с геномом x.params в виде вектора значений параметров длины n, сила мутации μ_s и степень мутации μ_r

```
function Mutate(f, x, μ_s, μ_r)
    if x.weights == None then
        x.weights ← {1}^n                          ▷ Задаем веса, если они не были заданы раньше
    k ← BinomialRandom(μ_s, n)                     ▷ Число параметров для изменения
    ω ← { x.weights[i] / Σ_{j=1}^{n} x.weights[j] }_{i=1}^{n}   ▷ Вероятность изменения параметра прямо
    пропорциональна его весу
    inds ← DiscreteRandom({i}_{i=1}^{n}, p = ω, size = k)   ▷ Индексы параметров
    x_mut ← x
    for i ∈ inds do                                ▷ Изменяем каждый параметр
        s ← UniformRandom({−1, +1})
        σ ← SigmaBasedOnThreeSigmaRule(μ_r, 0.0, 1.0)
        λ ← TruncNormRandom(μ_r, σ, 0.0, 1.0)
        x_mut.params[i] ← (1 + s · λ) · x_mut.params[i]
    if f(x_mut) < f(x) then                        ▷ Если произошло улучшение f
        for i ∈ inds do                            ▷ Обновляем веса
            x_mut.weights[i] ← x_mut.weights[i] + 1
    x_mut.lastOperation ← "Mutation"
    return x_mut
```

На начальных итерациях генетического алгоритма сильные мутации большого числа генов гораздо эффективнее слабых мутаций небольшого числа генов, тогда как при приближении к оптимальному решению все наоборот. Поэтому были разработаны **адаптивные степень и сила мутации**, которые могут изменяться в процессе работы алгоритма. Есть несколько способов сделать параметр алгоритма адаптивным, одним из самых популярных является алгоритм одной пятой [124]. Рассмотрим его применение к степени мутации μ_r: на каждой итерации, если у нас произошло событие «успеха» — улучшение целевой функции, то умножаем степень мутации μ_r на константу $C_{\mu_\text{r}} \in [1, 2]$. Если лучшее решение осталось прежним, то значение μ_r делится на корень четвертой степени из C_{μ_r}, уменьшая меру изменения генов при применении мутации. В случае силы мутации μ_s был предложен аналогичный подход, отличающийся тем, что для «успеха» дополнительно требуется проверить условие того, что новое лучшее решение было получено с помощью мутации.

Таким образом, частые обновления текущей наилучшей особи с помощью оператора мутации приводят к увеличению числа изменяющихся генов, а также к увеличению меры их изменения. При приближении к оптимальному решению и сходимости алгоритма происходит уменьшение частоты обновлений, числа генов и меры их изменения уменьшаются, что приводит к более точному поиску.

Псевдокод обновления адаптивных силы и степени мутации по алгоритму «одной пятой» представлен в листинге 3.

Листинг 3 – Процедура обновления силы и степени мутации по алгоритму «одной пятой». На вход алгоритм принимает значение μ, которое может быть μ_s или $\mu_г$, и соответствующую ему константу C_μ — C_{μ_s} и $C_{\mu_г}$ соответственно

```
1: function UpdateValue(success, μ, C_μ)
2:     if success == True then                    ▷ Правило «одной пятой»
3:         μ ← c · μ
4:     else
5:         μ ← 1 / c^0.25 · μ
       return μ
```

Следующим оператором разработанного генетического алгоритма является **оператор скрещивания**. При его применении выбираются две особи из текущего поколения, которые называются *родителями*, и на основе их геномов строится геном их *потомка*. В разработанном методе родители выбираются случайным образом по тому же правилу, что и особи для применения оператора мутации — вероятность выбора прямо пропорциональна значению функции приспособленности. Каждый ген в геноме потомка выбирается случайным образом с равной вероятностью от одного или второго родителя — был применен равномерный вид скрещивания. На рисунке 42 показано применение операции скрещивания для выбранных родителей. Псевдокод процедуры кроссовера представлен в листинге 4.

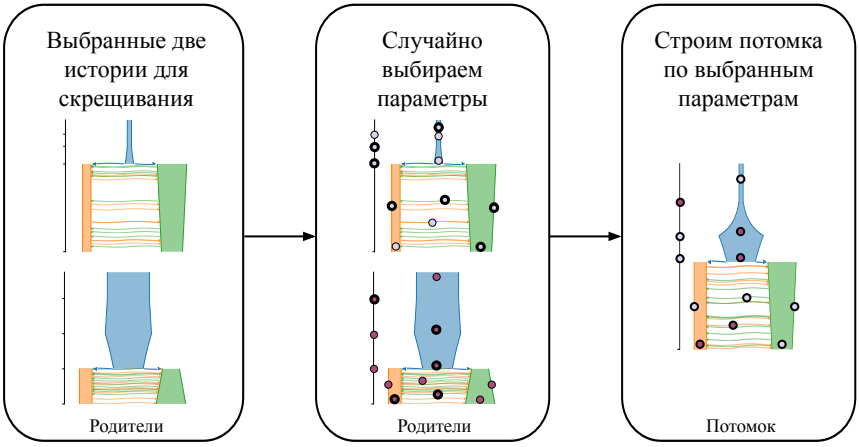


Рисунок 42 – Пример применения оператора кроссовера разработанного генетического алгоритма

Листинг 4 – Псевдокод оператора скрещивания разработанного генетического алгоритма. На вход подаются две выбранные особи-родителя x_1 и x_2, геном каждой из которых представлен в виде вектора значений параметров длины n

```
1:  function CrossOver(x_1, x_2)
2:      x ← CreateNewSolution( )
3:      x.params ← [ ]
4:      for i ← 1..n do                    ▷ Берем нужные параметры от родителей
5:          p ← UniformRandom({1, 2})
6:          if p == 1 then
7:              x.params.Add(x_1.params[i])
8:          else                                              ▷ Если p == 2
9:              x.params.Add(x_2.params[i])
10:     x.lastOperation ← "CrossOver"
11:     return x
```

Правило трех сигм и алгоритм выбора дисперсии усеченного нормального распределения. При применении оператора мутации для изменения генома особи используется усеченное нормальное распределение $\hat{N}(\mu, \sigma^2, a, b)$. Оно определяется четырьмя параметрами: математическим ожиданием μ, дисперсией σ^2, границами области значений $[a, b]$. Математическое ожидание этого распределения равно параметру силы мутации $\mu_{\text{т}}$, а выбор дисперсии σ^2 выполняется согласно специально разработанной процедуре. Опишем разработанную процедуру выбора среднеквадратичного отклонения σ, а, следовательно, и дисперсии σ^2, при заданном математическом ожидании μ. Идея основана на *правиле трех сигм* [125].

В соответствии с этим правилом вероятность того, что абсолютное отклонение нормальной случайной величины от математического ожидания будет меньше трех среднеквадратичных отклонений, равна $\sim 0{,}997$:

$$P(|\xi - \mu| < 3\sigma) \approx 0{,}997, \quad \xi \sim N(\mu, \sigma^2).$$

Данное правило можно переформулировать следующим образом: при сэмплировании случайной величины из распределения $N(\mu, \sigma^2)$ с вероятностью $99{,}7\%$ значение будет лежать в интервале $[\mu - 3\sigma, \mu + 3\sigma]$. Рисунок 43 иллюстрирует правило трех сигм. На рисунке представлен график плотности нормального распределения $N(0, 1)$, на котором наглядно продемонстрированы несколько интервалов области значений и вероятности того, что выборка случайной величины будет принадлежать этим интервалам. Синие области соответствуют интервалам, которые имеют ширину, равную среднеквадратичному отклонению σ. Правило трех сигм утверждает, что около 99.7% значений случайной величины лежат в пределе трех среднеквадратичных отклонений σ от математического ожидания μ.

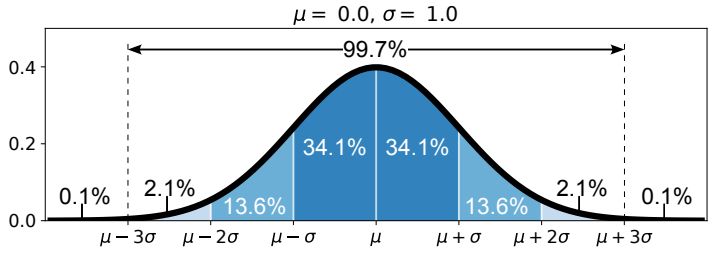


Рисунок 43 – Плотность нормального распределения $N(\mu, \sigma^2)$ и иллюстрация правила трех сигм

На основе правила трех сигм был разработан метод выбора среднеквадратичного отклонения σ по заданным значениям среднего μ, нижней и верхней границ допустимых значений случайной величины. Пусть требуется построить распределение случайной величины, которое имеет математическое ожидание μ и область определения, равную отрезку $[a, b]$. Рассмотрим нормальное распределение $\hat{N}(\mu, \sigma^2, a, b)$, усеченное на отрезок $[a, b]$. Оно определяется плотностью нормального распределения $N(\mu, \sigma^2)$, значения которой равны нулю за пределами указанного интервала, и которая умножена на константу таким образом, чтобы ее интеграл был равен единице. Выберем среднеквадратичное отклонение σ, как треть расстояния от среднего μ до границ отрезка $[a, b]$. Тогда согласно правилу трех сигм значения случайной величины будут лежать на отрезке с вероятностью 99.7%. В случае если отрезок не симметричен относительно среднего μ, среднеквадратичное отклонение выбирается как треть наибольшего расстояния от среднего до границ отрезка.

Псевдокод процедуры SigmaBasedOnThreeSigmaRule выбора среднеквадратичного отклонения для усеченного нормального распределения на отрезке $[a, b]$ при известном математическом ожидании $\mu \in [a, b]$ представлен в листинге 5. Примеры плотностей усеченного нормального распределения на отрезке $[0, 1]$, построенных для разных значений среднего μ_r и среднеквадратичного отклонения σ, выбранного с помощью разработанного метода, основанного на правиле трех сигм, приведены на рисунке 44.

Листинг 5 – Процедура выбора среднеквадратичного отклонения для усеченного нормального распределения на отрезке $[a, b]$ при известном математическом ожидании $\mu \in [a, b]$ по правилу трех сигм

```
1: function SigmaBasedOnThreeSigmaRule(μ, a, b)
2:     σ ← max{(μ−a),(b−μ)} / 3
3:     return σ
```

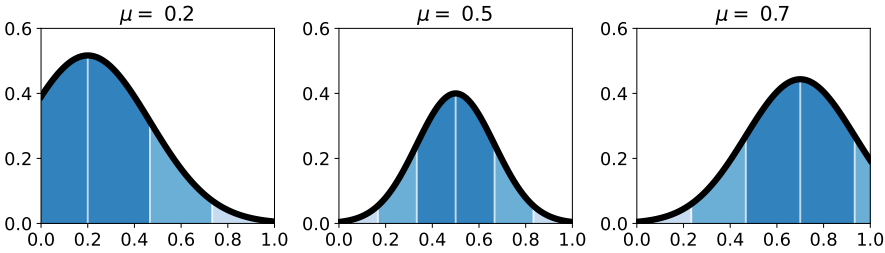


Рисунок 44 – Примеры плотности усеченного нормального распределения $\hat{N}(\mu, \sigma^2, 0, 1)$ на отрезке $[0, 1]$, у которого среднеквадратичное отклонение σ выбрано в соответствии с разработанным методом

2.2.2. Реализация разработанного метода, основанного на комбинации генетического алгоритма и локального поиска

Для реализации разработанного метода, основанного на комбинации генетического алгоритма и локального поиска, был реализован модуль `optimizers` на языке программирования Python. Этот модуль содержит различные методы оптимизации, которые применимы для решения задачи оптимизации произвольной функции f. Задача настройки параметров моделей демографической истории рассматривается, как задача оптимизации функции $f_{\mathcal{M}}(\theta) = \log(\mathcal{L}(\theta|\mathfrak{D}))$, которую можно решать с использованием методов рассматриваемого модуля. Структура классов реализованного модуля `optimizers` представлена на рисунке 45.

Основной абстрактный класс `Optimizer` соответствует произвольному методу оптимизации. У объекта этого класса задан атрибут `maximize`, который определяет какая задача решается: минимизации или максимизации целевой функции. Абстрактная процедура `optimize` запускает процесс поиска оптимальных значений параметров `variables` целевой функции `f`, который может быть ограничен `maxeval` — максимальным числом вычислений. От этого абстрактного класса наследуются:

- абстрактный класс `ContinuousOptimizer` методов решения задач оптимизации непрерывных параметров;
- абстрактный класс `UnconstrainedOptimizer` методов решения задач безусловной оптимизации — без ограничений на область значений параметров;
- абстрактный класс `ConstrainedOptimizer` методов решения задач условной оптимизации — с ограничениями на область значений параметров;
- абстрактный класс `GlobalOptimizer` методов глобальной оптимизации;

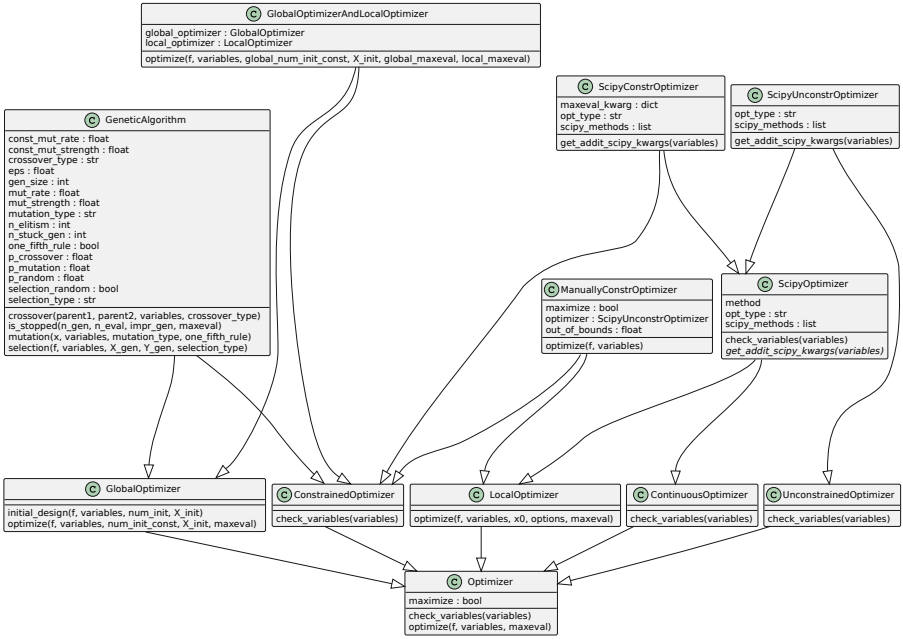


Рисунок 45 – Структура классов разработанного модуля `optimizers`

– абстрактный класс `LocalOptimizer` методов локальной оптимизации.

Рассмотрим реализацию методов локальной оптимизации. Для этого была использована библиотека SciPy [91], которая предоставляет множество методов, используемых в существующих программных решениях вывода демографической истории. Для сохранения предложенного интерфейса с процедурой `optimize` были реализованы классы-обертки `ScipyOptimizer`, `ScipyConstrOptimizer` и `ScipyUnconstrOptimizer`, которые соответствуют абстрактному классу произвольного метода, классу методов безусловной и условной оптимизации, представленных в библиотеке SciPy, соответственно. Дополнительный класс `ManuallyConstrOptimizer` позволяет использовать методы безусловной оптимизации из библиотеки SciPy для решения задач условной оптимизации. Например, метод BFGS, который является методом безусловной оптимизации, был реализован, как объект класса `ManuallyConstrOptimizer` для применения его для решения условной задачи вывода демографической истории. Были реализованы следующие методы, доступные в библиотеке SciPy:

– метод BFGS — объект класса `ManuallyConstrOptimizer`;
– метод L-BFGS-B — объект класса `ScipyConstrOptimizer`;

– метод Пауэлла — объект класса `ManuallyConstrOptimizer`;
– метод Нелдера-Мида — объект класса `ManuallyConstrOptimizer`.

Реализация методов глобальной оптимизации включает два метода: генетический алгоритм (класс `GeneticAlgorithm`) и разработанный комбинированный метод (класс `GlobalOptimizerAndLocalOptimizer`). Генетический алгоритм был реализован без использования сторонних библиотек. Он имеет процедуру `crossover` операции скрещивания, процедуру `mutation` операции мутации и процедуру `selection` отбора — вычисление целевой функции на множестве особей и выбор наиболее приспособленных. Дополнительная процедура `is_stopped` соответствует критерию останова.

Разработанный комбинированный метод реализован, как объект класса `GlobalOptimizerAndLocalOptimizer`. Этот класс позволяет комбинировать любой реализованный метод глобальной оптимизации с методом локальной оптимизации для последовательного запуска. Реализация разработанного метода на основе комбинации генетического алгоритма и метода BFGS локальной оптимизации с помощью модуля `optimizers` показана на рисунке 46.

```
import gadma

ga = gadma.optimizers.get_global_optimizer("Genetic_algorithm")
ls = gadma.optimizers.get_local_optimizer("BFGS")

optimizer = gadma.optimizers.GlobalOptimizerAndLocalOptimizer(
        global_optimizer=ga,
        local_optimizer=ls
)

```

Рисунок 46 – Реализация разработанного комбинированного метода с помощью модуля `optimizers`

Модуль `optimizers` можно использовать для решения произвольной задачи оптимизации. Пример решения задачи поиска минимума функции Розенброка [111] (рисунок 32) представлен на рисунке 47. Заметим, что программный код использует модуль `variables`, который был предложен для реализации моделей демографической истории и описан в разделе 2.1. Вывод, полученный путем запуска программного кода поиска минимума функции Розенброка, представлен на рисунке 48.

```python
import gadma
from scipy.optimize import rosen

# Рассмотрим функцию Розенброка и найдем ее минимум
f = rosen

# На вход она принимает вектор x параметров
x = [1.2, 1.5, 11]
f(x)

# Создадим переменные функции с областью значений
var1 = gadma.variables.ContinuousVariable(name='var1', domain=[-1, 2])
var2 = gadma.variables.ContinuousVariable(name='var2', domain=[0, 10])
var3 = gadma.variables.ContinuousVariable(name='var3', domain=[0.01, 100])
variables = [var1, var2, var3]

# Создадим объект генетического алгоритма
ga = gadma.optimizers.get_global_optimizer("Genetic_algorithm")
ga.maximize = False

# Создадим объект метода Нелдера-Мида
nm = gadma.optimizers.get_local_optimizer("Nelder-Mead")
nm.maximize = False

# Создаем объект комбинированного метода
optimizer = gadma.optimizers.GlobalOptimizerAndLocalOptimizer(ga, nm)

# Запускаем оптимизацию для нахождения точки минимума
res = optimizer.optimize(
        f,
        variables,
        verbose=1,
        global_maxeval=102,
        local_maxeval=None,
)
```

Рисунок 47 – Применение разработанного комбинированного метода на основе генетического алгоритма, реализованного с помощью модуля `optimizers`, для поиска точки минимума функции Розенброка

```
--Start global optimization Genetic_algorithm--
Generation #0.
Current generation of solutions:
N       Value of fitness function       Solution
0        1702.867066    (var1=-8.40e-01,        var2=2.95768,   var3=12.19485)  r
1        3481.693449    (var1=1.17017,  var2=6.42679,   var3=44.29422)  r
2        3619.723356    (var1=1.80943,  var2=3.27259,   var3=16.72141)  r
...
49       873076.902020  (var1=1.14757,  var2=2.52083,   var3=99.78527)  r
Current mean mutation rate:      0.200000
Current mean number of params to change during mutation:          1

--Best solution by value of fitness function--
Value of fitness: 1702.8670662220752
Solution:               (var1=-8.40e-01,        var2=2.95768,   var3=12.19485)  r

...

Generation #5.
Current generation of solutions:
N       Value of fitness function       Solution
0        1201.420029    (var1=-8.40e-01,        var2=3.1181,    var3=12.19485)  c
1        1702.867066    (var1=-8.40e-01,        var2=2.95768,   var3=12.19485)  r
2        1716.993700    (var1=-8.21e-01,        var2=2.95768,   var3=12.19485)  m
3        1896.320118    (var1=-1.00e+00,        var2=4.23198,   var3=15.01605)  m
4        2095.132556    (var1=-8.40e-01,        var2=4.23198,   var3=15.01605)  m
5        11236.211961   (var1=-3.50e-01,        var2=9.78427,   var3=100.0)     mm
6        17876.525753   (var1=1.17637,  var2=5.58294,   var3=18.48368)  c
7        119548.247744  (var1=1.89707,  var2=3.1181,    var3=44.29422)  c
8        529830.730429  (var1=-9.49e-01,        var2=8.88778,   var3=6.64731)   r
9        841461.009411  (var1=-1.66e-01,        var2=1.45411,   var3=93.83444)  r
Current mean mutation rate:      0.200000
Current mean number of params to change during mutation:          1

--Best solution by value of fitness function--
Value of fitness: 1201.420029007254
Solution:               (var1=-8.40e-01,        var2=3.1181,    var3=12.19485)  c

Mean time:      0.002 sec.

--Finish global optimization Genetic_algorithm--
--Start local optimization optimize_fmin--
0       1201.420029007254       (var1=-8.40e-01,        var2=3.1181,    var3=12.19485)
1       1201.420029007254       (var1=-8.40e-01,        var2=3.1181,    var3=12.19485)
2       1167.2248836071785      (var1=-8.82e-01,        var2=3.1181,    var3=12.19485)
...
326     1.109015909590664e-09   (var1=1.00001,  var2=1.00002,   var3=1.00004)
--Finish local optimization optimize_fmin--
Result:
   status: 0
 success: True
 message: GLOBAL OPTIMIZATION: MAXIMUM NUMBER OF FUNCTION EVALUATIONS ACHIEVED; LOCAL
OPTIMIZATION:
       x: [1.00000855 1.00001704 1.00003682]
       y: 1.109015909590664e-09
  n_eval: 621
```

Рисунок 48 – Пример вывода программы, представленной на рисунке 47

2.2.3. Настройка гиперпараметров разработанного генетического алгоритма

Гиперпараметр — это параметр алгоритма. Производительность любого алгоритма зависит от значений его гиперпараметров — от его *конфигурации*. Одним из классических примеров является темп (learning rate) обучения нейронной сети. Проблема нахождения гиперпараметров, обеспечивающих наилучшую сходимость алгоритма на нескольких экземплярах задачи, называется *задачей конфигурации алгоритма*. Экземплярами задачи могут быть различные постановки этой задачи для разных данных. *Задачу конфигурации алгоритма* можно сформулировать следующим образом: для алгоритма A, набора экземпляров задачи I и *функции стоимости* c найти набор гиперпараметров для A, который является наилучшим относительно c на всем множестве I. Обычно функция стоимости c основана либо на времени, необходимом для решения проблемы, либо на оценке качества решения, достигнутом в рамках фиксированного бюджета. SMAC [126, 127] — это программное обеспечение, реализующее метод, основанный на байесовской оптимизации, для решения задачи конфигурации алгоритма.

Метод байесовской оптимизации основан на использовании *суррогатной модели* для аппроксимации целевой функции f и *функции выбора* α для поиска многообещающих точек на каждой итерации. На каждой итерации происходит аппроксимация суррогатной моделью по имеющемуся набору точек $\theta_1, y_1, \theta_2, y_2, \ldots, \theta_n, y_n$, где $y_i = f(\theta_i)$. Затем по аппроксимации строится функция выбора α и происходит поиск точки ее максимума, которая затем выбирается как новая точка для вычисления целевой функции. Подробное описание метода байесовской оптимизации приведено в разделе 2.3.

В качестве суррогатной модели байесовская оптимизация в SMAC использует случайный лес, а в качестве функции выбора функцию ожидаемого улучшения (Expected Improvement, EI):

$$\alpha_{\mathrm{EI}}(\theta) = \mathbb{E}_\theta\big[\max(0, \hat{f}(\theta)) - \max_{1 \leqslant i \leqslant n}(y_i)\big],$$

где $\hat{f}$ — построенная аппроксимация целевой функции f. В отличие от классической байесовской оптимизации, SMAC реализует многокритериальную оптимизацию, так как рассматривает несколько экземпляров задачи. Таким образом, цель состоит в поиске решения, которое будет наилучшим для всех экземпляров задачи в соответствии с заданной функцией стоимости c. Данная многокритериальность может быть решена различными способами. SMAC сводит эту задачу к задаче обычной однокритериальной оптимизации, используя в качестве целевой функции среднее значение функций стоимости на экземплярах задач. Еще одним расширением метода является «интенсификация» — процедура сравнения конфигураций гиперпараметров и выбора наилучшей. Метод «интенсификации», реализованный в SMAC, рассматривает конфигурацию как новую лучшую, если она лучше, чем текущая лучшая на наборе пар из экземпляров задачи

и генераторов случайных чисел (random seed). Учитывание генераторов случайных чисел позволяет проводить честные сравнения двух конфигураций на одном экземпляре задачи в случае, если рассматриваемый алгоритм не детерминированный. Используемый набор пар расширяется каждый раз, когда тестируемая конфигурация оказывается хуже текущей лучшей, с которой и происходит сравнение. Таким образом, этот набор всегда растет и требуется все больше и больше сравнений, чтобы превзойти текущую лучшую конфигурацию.

Из изложенного следует, что SMAC — это итеративный метод, который на каждой итерации поддерживает наилучшую конфигурацию гиперпараметров оптимизируемого алгоритма. Новая многообещающая конфигурация выбирается и сравнивается с текущей лучшей в рамках процедуры «интенсификации» на каждой итерации. В результате в качестве ответа SMAC предоставляет конфигурацию гиперпараметров для алгоритма, которая имеет наилучшее по среднему значение функции стоимости. SMAC был использован для поиска оптимальных значений гиперпараметров разработанного генетического алгоритма для поиска демографической истории популяций.

Метод оптимизации, такой как генетический алгоритм, имеет несколько гиперпараметров, и их число изменяется в зависимости от конкретной реализации. Генетический алгоритм, представленный в разделе 2.2, имеет десять гиперпараметров. Каждая особь в поколении представляется в виде массива значений параметров θ модели $\mathcal{M}$ демографической истории. Начальное множество решений, называемое поколением, формируется процедурой *начального дизайна*, которая генерирует множество случайных параметров. Размер этого набора определяется значением гиперпараметра `n_init_const`: число решений в начальном поколении равно числу параметров модели, умноженному на `n_init_const`. Размер каждого поколения в генетическом алгоритме равен значению гиперпараметра `gen_size`. Новое поколение строится итеративно с помощью мутации, скрещивания и отбора лучших по значению вероятностных моделей. Доли наиболее адаптированных, мутировавших, скрещенных и случайных моделей, формирующих новое поколение, определяются `p_elitism`, `p_mutation`, `p_crossover`, `p_random` гиперпараметрами соответственно. Особым случаем является процесс адаптивной мутации: дополнительно содержащий параметры `mutation_strength` и `mutation_rate`, которые определяют, сколько параметров модели и насколько сильно будут изменяться их значения Каждый из этих двух гиперпараметров изменяется в процессе работы генетического алгоритма по правилу одной пятой: чем ближе к оптимуму, тем меньше изменений в процессе мутации. Константы правила одной пятой (`const_mutation_strength`, `const_mutation_rate`) также являются гиперпараметрами генетического алгоритма. Всего было выделено десять гиперпараметров генетического алгоритма: два имеют целочисленные дискретные значения, а восемь — численные непрерывные. Краткое описание каждого гиперпараметра представлено в таблице 2.

Таблица 2 – Краткое описание и обозначения гиперпараметров разработанного генетического алгоритма.

ID гиперпараметра	Обозначение в псевдокоде	Описание гиперпараметра
`gen_size`	N_{GEN}	Число решений в одном поколении генетического алгоритма
`n_init_const`	—	Константа, определяющая число случайных решений в случайно сгенерированном поколении
`p_elitism`	$N_{\mathrm{B}}/N_{\mathrm{gen}}$	Доля лучших решений, перенесенных в новое поколение
`p_mutation`	$N_{\mathrm{M}}/N_{\mathrm{gen}}$	Доля мутировавших решений в новом поколении
`p_crossover`	$N_{\mathrm{C}}/N_{\mathrm{gen}}$	Доля решений, образованных кросинговером, в новом поколении
`p_random`	$N_{\mathrm{R}}/N_{\mathrm{gen}}$	Доля случайно сгенерированных решений в новом поколении
`mutation_strength`	μ_{s}	Начальная сила мутации
`const_mutation_strength`	$C_{\mu_{\mathrm{s}}}$	Константа для изменения `mutation_strength` во время генетического алгоритма согласно правилу одной пятой
`mutation_rate`	μ_{r}	Начальная степень мутации
`const_mutation_rate`	$C_{\mu_{\mathrm{r}}}$	Константа для изменения `mutation_rate` во время генетического алгоритма согласно правилу одной пятой

Начальные значения гиперпараметров в первой версии разработанного генетического алгоритма были настроены вручную в ходе вывода демографической истории двух популяций современных людей для модели и данных из [45] (набор данных `2_YRI_CEU_6_Gut`). Их значения представлены в таблице 3.

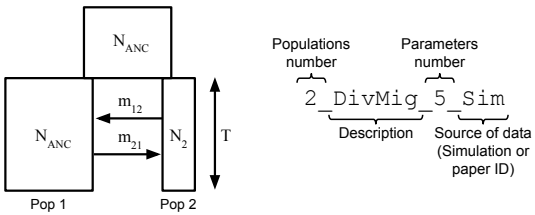


Рисунок 49 – Пример описание структурированного формата имени набора данных из пакета `deminf_data v1.0.0`

Таблица 3 – Начальные значения и область значений, используемые для оптимизации гиперпараметров разработанного генетического алгоритма. Первые два гиперпараметра `gen_size` и `n_init_const` являются целыми числами и имеют дискретную область значений. Остальные восемь гиперпараметров являются непрерывными.

ID гиперпараметра	Начальное значение	Область значений
`gen_size`	10	$\{10, 50, 100\}$
`n_init_const`	10	$\{5, 10, 20\}$
`p_elitism`	0.2	$[0, 1]$
`p_mutation`	0.3	$[0, 1]$
`p_crossover`	0.3	$[0, 1]$
`p_random`	0.2	$[0, 1]$
`mutation_strength`	0.2	$[0, 1]$
`const_mutation_strength`	1.01	$[1, 2]$
`mutation_rate`	0.2	$[0, 1]$
`const_mutation_rate`	1.02	$[1, 2]$

Все используемые наборы данных были взяты из пакета Python `deminf_data v1.0.0`, который доступен на платформе GitHub по ссылке: `https://github.com/noscode/demographic_inference_data`. Каждый набор данных имеет структурированное имя (рисунок 49), которое представляет собой последовательность: а) номер популяции, б) краткое описание демографической модели, в) число параметров и г) информация об источнике данных.

Метод SMAC был использован для настройки гиперпараметров разработанного генетического алгоритма. Наборы данных были разделены на две группы: обучающие и тестовые. Оптимизация выполнялась с использованием метода *moments* для вычисления правдоподобия на четырех обучающих наборах данных, которые представляли экземпляры задачи. Тестовые наборы данных были использованы для проверки производительности новых конфигураций после оптимизации с помощью SMAC.

Четыре набора данных были выбраны в качестве обучающих: три набора имели симулированные данные (`2_BotDivMig_8_Sim`, `2_DivMig_5_Sim`, `2_ExpDivNoMig_5_Sim`) и один (`2_YRI_CEU_6_Gut`) содержал реальные генетические данные для современных человеческих популяций из работы [45]. Обучающие набора данных имели одинаковое число популяций и размер генетических данных.

Тестовые набора данных являются более разнообразными: два набора данных (`1_Bot_4_Sim`, `1_AraTha_4_Hub`) для одной популяции с симулированными и реальными данными из [128], два набора (`2_ButAll_3_McC`, `2_ButSynB2_5_McC`) для двух популяций бабочек из [129]; один на-

бор с симулированными генетическими данными для трех популяций (`3_DivMig_8_Sim`); и один набор данных (`2_YRI_CEU_struct_11_Nos`) с генетическими данными для двух популяций современного человека из [45] и моделью расширенного класса.

Функция стоимости c была определена как лучшее значение логарифма правдоподобия, достигнутое генетическим алгоритмом за фиксированное число вычислений. Для остановки было использовано число вычислений, которое в 200 раз превышает число параметров модели в наборе данных. Полный запуск генетического алгоритма может требовать как меньше, так и больше вычислений целевой функции в зависимости от набора данных, так как он имеет критерий остановки, основанный на сходимости оптимизации.

Во время первого запуска SMAC, все десять гиперпараметров (Таблица 4) были оптимизированы. Затем два дискретных гиперпараметра (`gen_size` и `n_init_const`) были зафиксированы в конфигурации, и запущены дополнительные пять попыток оптимизации гиперпараметров. Во время второй попытки оптимизации с помощью SMAC значения дискретных гиперпараметров `gen_size` и `n_init_const` были установлены на значения по умолчанию (10 и 10 соответственно). Для третьей и четвертой попыток были протестированы альтернативные значения гиперпараметра `n_init_const` — 5 и 20. Затем размер поколения в генетическом алгоритме (`gen_size`) был увеличен до 50, а гиперпараметр начального дизайна (`n_init_const`) был протестирован на значениях 10 и 20. Значение `n_init_const` равное пяти было исключено, так как оно приводит к небольшому числу решений для первого поколения, которое имеет размер 50.

Было выполнено шесть попыток оптимизации гиперпараметров при помощи SMAC. При первой попытке SMAC не смог найти конфигурацию гиперпараметров, превосходящую исходную. В результате было получено пять новых конфигураций, которые были сравнены между собой для выбора наилучшей. Полученные конфигурации представлены в таблице 4. Для оценки эффективности новых конфигураций и выбора наилучшей были проведены полные запуски генетического алгоритма. Для каждой конфигурации и набора данных генетический алгоритм был запущен 128 раз с использованием двух методов вычисления правдоподобия: *moments* и *momi2*. Для остановки полных запусков генетического алгоритма использовался тот же или эквивалентный критерии, что и в исходной версии метода. Например, генетический алгоритм с конфигурацией, где размер поколения (`gen_size`) равен 10, был остановлен после 100 поколений без улучшения значения правдоподобия. Для конфигураций с `gen_size` равным 50 использовался эквивалентный критерий остановки — отсутствие улучшений в течение 20 поколений.

Сравнение новых конфигураций было выполнено следующим образом. Сначала были измерены темпы ускорения, которые представляют собой долю вычислений целевой функции, которая экономится при использовании новой конфигурации вместо исходной. Например, если конфигурация по умолчанию в

Таблица 4 – Значения гиперпараметров генетического алгоритма после каждой попытки оптимизации с помощью SMAC

ID Гиперпараметра	Номер попытки					
	1 (по умолчанию)	2	3	4	5	6
gen_size	10	10*	10*	10*	50*	50*
n_init_const	10	10*	5*	20*	10*	20*
p_elitism	0.20	0.30	0.30	0.40	0.40	0.40
p_mutation	0.30	0.20	0.20	0.10	0.08	0.10
p_crossover	0.30	0.30	0.30	0.30	0.42	0.46
p_random	0.20	0.20	0.20	0.20	0.10	0.04
mutation_strength	0.200	0.776	0.370	0.534	0.833	0.528
const_mutation_strength	1.010	1.302	1.290	1.648	1.199	1.492
mutation_rate	0.200	0.273	0.886	0.882	0.595	0.345
const_mutation_rate	1.020	1.475	1.942	1.417	1.645	1.472

*Значения были зафиксированы во время оптимизации гиперпараметров с помощью SMAC.

среднем выполняла X вычислений целевой функции для набора данных, а новая конфигурация в среднем занимает Y вычислений, то темп ускорения будет равен $\frac{X-Y}{X}$. Затем, каждая новая конфигурация была сравнена с исходной с помощью сравнения финальных значений правдоподобия. Для каждого набора данных новая конфигурация считается лучше, чем исходная, если медиана и оба квартиля из 128 значений правдоподобия выше, чем для конфигурации по умолчанию. Если медиана и оба квартиля ниже, то производительность новой конфигурации на наборе данных считается хуже. Во всех остальных случаях, конфигурации считаются несравнимыми. Целью исследования является выбор такой конфигурации, которая работает быстрее и лучше, чем конфигурация по умолчанию на максимально возможном числе наборов данных, и работает хуже на минимальном числе наборов данных.

Темпы ускорения, вычисленные для новых конфигураций, представлены в таблице 5. В среднем все новые конфигурации требуют меньшего числа вычислений, чем исходный генетический алгоритм. Конфигурации, полученные в ходе попыток три, пять и шесть, являются самыми быстрыми. Однако их финальные правдоподобия хуже, чем у конфигурации по умолчанию на большинстве наборов данных (рисунок 50). Генетический алгоритм с конфигурациями, полученными в ходе попыток два и четыре, демонстрирует лучшую эффективность с точки зрения финального правдоподобия среди новых конфигураций. Более того, генетический алгоритм с гиперпараметрами из попытки два имеет лучшие значения правдоподобия для *moments*, в то время как конфигурация из попытки четыре имеет лучшие результаты для *momi2*. Поскольку при оптимизации гиперпараметров использовался *moments* движок, конфигурация номер два была выбрана как новые гиперпараметры для генетического алгоритма. Рисунок 51 суммирует улучшение генетического алгоритма с новыми гиперпараметрами по сравнению с исходной версией. Новая конфигурация экономит около 10% вычис-

лений и обеспечивает лучшие результаты в среднем по сравнению с исходным генетическим алгоритмом.

Таблица 5 – Темпы ускорения генетического алгоритма при использовании новых конфигураций по сравнению с исходной версией

	Номер попытки				
	2	3	4	5	6
moments	6%	30%	11%	28%	51%
momi2	16%	35%	15%	37%	55%
Среднее	10%	32%	13%	32%	53%

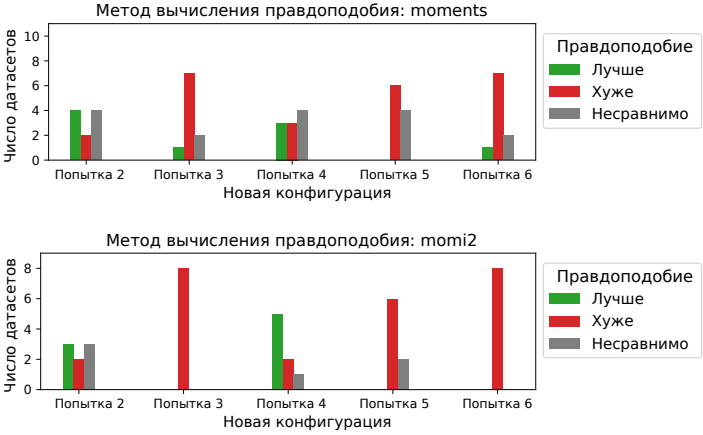


Рисунок 50 – Сравнение значений правдоподобия, полученных с помощью новых конфигураций и исходной конфигурации генетического алгоритма

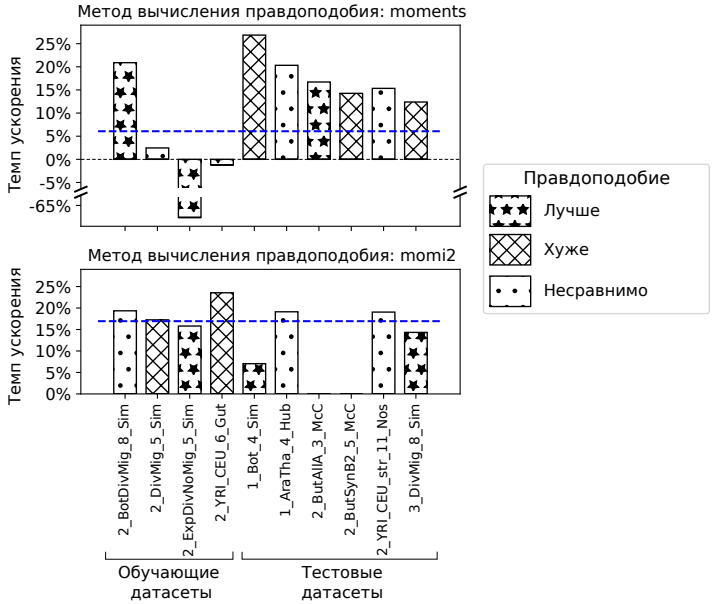


Рисунок 51 – Сравнение генетического алгоритма с настроенными гиперпараметрами с исходной версией на различных наборах данных

2.3. Метод на основе комбинации байесовской оптимизации и локального поиска для настройки параметров модели демографической истории популяций по генетическим данным

Байесовская оптимизация является одним из наиболее популярных методов для оптимизации сложных функций, которые не имеют заданной структуры, например, градиента, и требуют значительных вычислительных затрат [130–132]. Она направлена на минимизацию целевой функции $f : \Theta \to \mathbb{R}$, где для любого $\theta \in \Theta$ можно получить зашумленное наблюдение $y(\theta)$ для значения $f(\theta)$. Целью оптимизации является быстрая и точная сходимость к глобальному оптимуму в рамках заданного временного бюджета.

Процедура байесовской оптимизации включает две основные компоненты: аппроксимацию целевой функции с использованием *суррогатной модели* и выбор следующей точки для вычисления целевой функции с помощью *функции выбора*. Для суррогатной модели часто применяется регрессия на основе гауссовского процесса (подробнее об этом будет рассказано далее). Гауссовские процессы имеют ряд преимуществ, включая способность работать с небольшим объемом данных и возможность количественной оценки неопределенности, связанной с аппроксимацией функции. Кроме гауссовских процессов, в байесовской оптимизации иногда используются также случайные леса [126]. Например, они

были применены в методе оптимизации гиперпараметров SMAC. Тем не менее гауссовский процесс является наиболее распространенным выбором, и именно он был выбран в качестве суррогатной модели для байесовской оптимизации в данной работе. Поэтому в данном разделе описание метода сразу предполагает использование гауссовского процесса.

2.3.1. Разработка метода на основе комбинации байесовской оптимизации и локального поиска

Процедура байесовской оптимизации начинается с построения начального множества решений и вычисления целевой функции на них. Эта процедура называется *начальным дизайном*. После этого выбирается априорный гауссовский процесс.Выбор априорного процесса может быть сделан на основе известных свойств целевой функции или с помощью процедуры кросс-валидации, которая подробно описана далее.

На каждой итерации байесовской оптимизации строится регрессия на основе Гауссовского процесса на текущих точках $(\theta_1, y_1), \dots (\theta_n, y_n)$. Затем на основе регрессии строится функция выбора, выбирается новая точка θ_*, как точка ее максимума:

$$\theta_* = \arg\max_{\theta} \alpha(\theta),$$

и происходит вычисление целевой функции в ней: $y_{n+1} = f(\theta^*)$. Затем начинается новая итерация байесовской оптимизации. Функция выбора в общем случае может быть произвольной, однако для обеспечения эффективности метода ее вычисление и оптимизация должны требовать меньше вычислительных ресурсов, чем вычисление целевой функции.

Наиболее часто используемые функции выбора [133]:

$$\alpha_{\text{EI}}(\theta) = \mathbb{E}\big[\max\big(0, \min_{1 \leqslant i \leqslant n}(y_i) - \hat{f}(\theta)\big)\big], \tag{1}$$

$$\alpha_{\text{PI}}(\theta) = \mathbb{P}\big[\hat{f}(\theta) < \min_{1 \leqslant i \leqslant n}(y_i)\big], \tag{2}$$

$$\alpha_{\text{logEI}}(\theta) = \mathbb{E}\big[\max(0, \min_{1 \leqslant i \leqslant n}(e^{y_i}) - e^{\hat{f}(\theta)})\big], \tag{3}$$

где $\hat{f}$ — аппроксимация целевой функции f суррогатной моделью. Функция выбора α_{EI} (EI, Expected Improvement) оценивает ожидаемое улучшение — максимально возможное ожидаемое улучшение целевой функции в точке. Функция выбора α_{PI} (PI, Probability of Improvement) моделирует вероятность *любого* улучшения по сравнению с текущим оптимумом. Функция выбора $\alpha_{\texttt{logEI}}(\theta)$, была предложена в [134] и является аналогом функции выбора `EI`, когда аппроксимация $\hat{f}$ суррогатной моделью использует логарифмированные значения целевой функции f. Это означает, что регрессия строится на точках $(\theta_i, \log y_i)$, вместо точек (θ_i, y_i). Логарифмирование требует, чтобы целевая функция принимала только положительные значения — $y_j > 0$. Целевая функция f, используемая в данной

работе — отрицательный логарифм правдоподобия, который всегда имеет положительное значение, поэтому функцию выбора $\alpha_{\texttt{logEI}}(\theta)$ можно использовать. Для суррогатной модели гауссовского процесса рассматриваемые функции выбора представляются в аналитическом виде и могут быть эффективно оптимизированы с использованием алгоритмов градиентного спуска с перезапусками. Далее в тексте функции выбора (1), (2) и (3) будут обозначаться, как `EI`, `PI` и `LogEI` соответственно.

Пример итерации байесовской оптимизации изображен на рисунке 52. Голубая линия соответствует предсказанию регрессии на основе гауссовского процесса, голубая область вокруг — доверительный интервал предсказания. Красная линия соответствует функции выбора, пунктирная черная линия — точке ее максимума. Рисунок 52а изображает начало итерации, когда происходит построение регрессии на основе гауссовского процесса по черным точкам, где была вычислена целевая функция. Затем происходит поиск точки максимума функции выбора, изображенной красной линией на рисунке 52б, и вычисление целевой функции в ней — красная точка. Наконец, рисунок 52в демонстрирует как текущее множество точек обновляется и новая регрессия строится на его основе.

Псевдокод процедуры байесовской оптимизации представлен в листинге 6. Алгоритм принимает на вход целевую функцию f, которая является функцией правдоподобия, начальное множество точек X^{init}, суррогатную модель гауссовского процесса GP, функцию выбора α и максимальное число итераций T^{max}.

Для решения задачи настройки параметров моделей демографической истории по генетическим данным был разработан комбинированный метод байесовской оптимизации и локального поиска. Как и в случае с генетическим алгоритмом, разработанный метод последовательно запускает сначала байесовскую оптимизацию для настройки всех параметров, а затем метод локальной оптимизации для дополнительной настройки непрерывных параметров. Схема алгоритма разработанного метода представлена на рисунке 53.

Для выбора функции выбора и гиперпараметров суррогатной модели в байесовской оптимизации были проведены экспериментальные исследования для выявления эффективности различных конфигураций. Был также разработан и протестирован метод автоматического выбора функции ковариации суррогатной модели гауссовского процесса. В результате было предложено

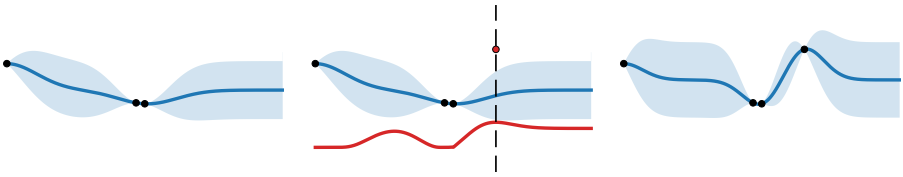


Рисунок 52 – Фрагмент байесовской оптимизации

Листинг 6 – Псевдокод метода байесовской оптимизации

```
1:  function BayesianOptimization(f, X^init, GP, α, T^max)
2:      Y^init ← {f(θ), θ ∈ X^init}
3:      X, Y ← X^init, Y^init
4:      T_cur ← 1
5:      while T_cur ⩽ T^max do
6:          GP ← GaussianProcessRegression(GP, X, Y)
7:          α_GP ← GetAcquisitionFunctionForGP(GP)
8:          θ* ← arg max_θ α_GP(θ)
9:          X ← X ∪ {θ*}
10:         Y ← Y ∪ {f(θ*)}
11:     return θ_best : arg max_{θ∈X} f(θ)
```

применение ансамблевого метода байесовской оптимизации в составе разработанного комбинированного метода. Этот метод включает разработанный метод автоматического выбора функции ковариации, а также ансамбль из функций выбора. На каждой итерации байесовской оптимизации функция выбора выбирается случайным образом из этого ансамбля. Подробное описание настройки гиперпараметров байесовской оптимизации приведено в разделе 2.3.3.

Регрессия на основе гауссовского процесса. С математической точки зрения, *гауссовский процесс* — это семейство $\{\hat{f}_\theta\}_{\theta\in\times}$ случайных величин $\hat{f}_\theta$, проиндексированных множеством $\Theta \subset \mathbb{R}^d$ и распределенных согласно многомерному гауссовскому распределению. Следовательно, для любого конечного множества индексов $\theta_1, \theta_2, \ldots, \theta_k$ множества Θ случайная величина:

$$\hat{f}_{\theta_1,\ldots,\theta_k} = (\hat{f}_{\theta_1}, \ldots, \hat{f}_{\theta_k})$$

имеет многомерное нормальное распределение $\hat{f}_{\theta_1,\ldots,\theta_k} \sim \mathcal{N}_k(\mu,\, K^2)$.

Гауссовский процесс или, строго говоря, его распределение определяется парой детерминистических функций: функцией среднего $m : \Theta \to \mathbb{R}$ и функцией ковариации $k : \Theta \times \Theta \to \mathbb{R}$:

$$m(\theta) = \mathbb{E}(\hat{f}_\theta), \qquad k(\theta, \theta') = Cov(\hat{f}_\theta, \hat{f}_{\theta'}). \tag{4}$$

Функция ковариации k также называется *ядром* гауссовского процесса.

Было доказано, что любая пара функций m и k, где k — положительно определенная функция, определяет гауссовский процесс [133], из этого следует стандартное обозначение:

$$\hat{f} \sim GP(m, k). \tag{5}$$

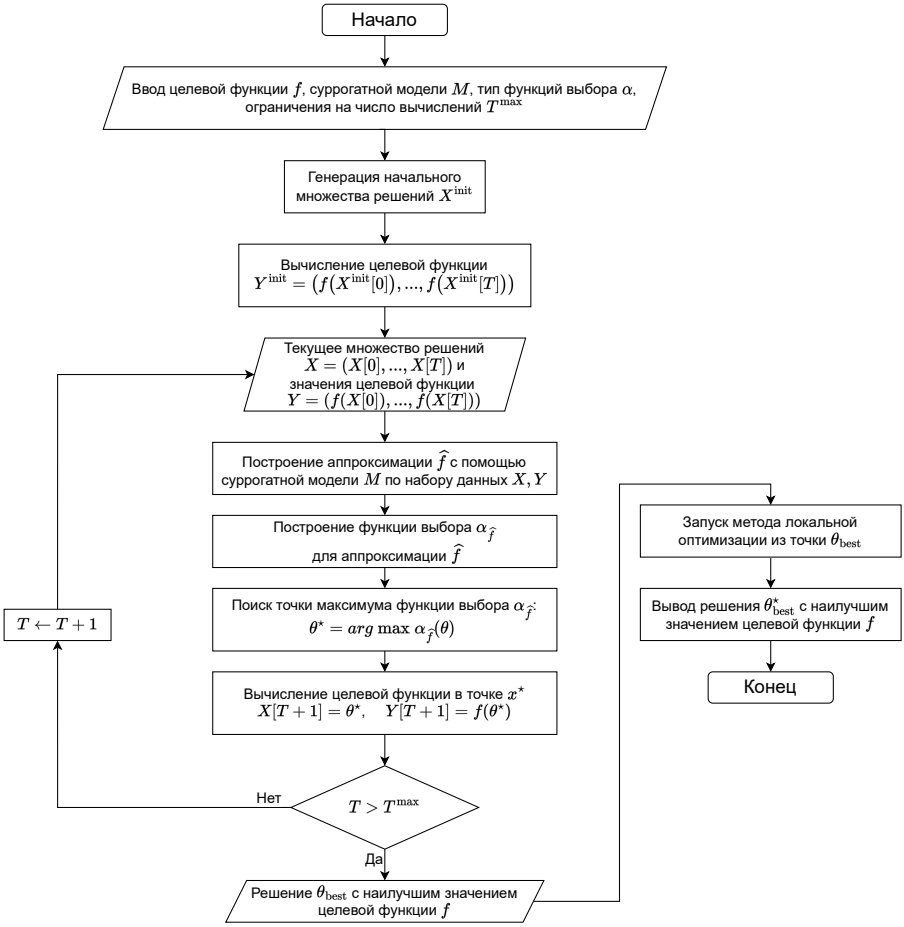


Рисунок 53 – Схема алгоритма разработанного комбинированного метода, основанного на баейсовской оптимизации и методе локального поиска

Для процедуры регрессии на основе гауссовского процесса требуется выбрать *параметры априорного распределения* $\hat{f}_0 \sim GP(m, k)$. На практике функция среднего m обычно предполагается равной нулю или другой константе [133]. В разработанной байесовской оптимизации для настройки параметров моделей демографической истории используется $m \equiv 0$, что позволяет сосредоточиться больше на выборе ковариационной функции k.

Наиболее распространенное семейство ковариационных функций для гауссовского процесса — *семейство ядер Матерна* [133, 135]. В контексте данной работы понятие ядра равнозначно понятию ковариационной функции гауссовского процесса. Семейство ядер Матерна параметризовано тремя гиперпараметрами: *гладкость* ν, *масштаб* κ и *дисперсия* σ^2. Гладкость определяет степень дифференцируемости гауссовского процесса, масштаб и дисперсия масштабируют оси значений параметров и целевой функции. Предполагая нулевой априорный вектор средних, общий вид ядер Матерна выглядит следующим образом [136]:

$$k_{\nu,\kappa,\sigma^2}(t,t') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|t-t'\|}{\kappa} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{\|t-t'\|}{\kappa} \right), \tag{6}$$

где K_ν — это функция Бесселя второго порядка, а Γ — гамма-функция.

Семейство ядер Матерна, заданное формулой (6), часто делят на подсемейства, соответствующие выбору гладкости ν. Наиболее используемыми являются подсемейства для $\nu \in \{1/2, 3/2, 5/2, \infty\}$, что приводит к получению следующих формул:

$$k_{1/2,\kappa,\sigma^2}(t,t') = \sigma^2 \exp\left(-\frac{u}{\kappa}\right), \tag{7}$$

$$k_{3/2,\kappa,\sigma^2}(t,t') = \sigma^2 \left(1 + \frac{\sqrt{3}u}{\kappa}\right) \exp\left(-\frac{\sqrt{3}u}{\kappa}\right), \tag{8}$$

$$k_{5/2,\kappa,\sigma^2}(t,t') = \sigma^2 \left(1 + \frac{\sqrt{5}u}{\kappa} + \frac{5u^2}{3\kappa^2}\right) \exp\left(-\frac{\sqrt{5}u}{\kappa}\right), \tag{9}$$

$$k_{\infty,\kappa,\sigma^2}(t,t') = \sigma^2 \exp\left(-\frac{u^2}{2\kappa^2}\right), \tag{10}$$

где $u = \|t - t'\|$. Первое ядро $k_{1/2,\kappa,\sigma^2}$ обычно называется *экспоненциальным*, последнее $k_{\infty,\kappa,\sigma^2}$ — *гауссовским* ядром или ядром *RBF*.

Аппроксимирующая часть предложенной байесовской оптимизации использует *регрессию на основе гауссовского процесса*. Имея априорный гауссовский процесс $\hat{f}_0$ и набор точек $(\theta_1, y_1), \ldots, (\theta_n, y_n)$, можно построить апостериорный гауссовский процесс $\hat{f}$, используя формулу Байеса:

$$p(\hat{f}) \stackrel{\text{def}}{=} p(\hat{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\hat{f}_0)p(\hat{f}_0)}{p(\boldsymbol{y})}, \tag{11}$$

где $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. В общем случае апостериорное распределение не всегда является гауссовским процессом. Однако, если предположить, что каждое наблюдение y_i зашумлено $y_i = f(\theta_i) + \varepsilon$, где $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, то $\hat{f}$ всегда является гауссовским процессом [133]. Более того, $\hat{f} \sim GP(\hat{m}, \hat{k})$ с функциями $\hat{m}$ и $\hat{k}$, полученными следующим образом (для простоты предполагаем, что $m \equiv 0$):

$$\hat{m}(\cdot) = \mathbf{K}_{\cdot,\boldsymbol{\theta}}(\mathbf{K}_{\boldsymbol{\theta},\boldsymbol{\theta}} + \sigma_\varepsilon^2\mathbf{I})^{-1}\boldsymbol{y}, \tag{12}$$

$$\hat{k}(\cdot,\cdot') = k(\cdot,\cdot') - \mathbf{K}_{\cdot,\boldsymbol{\theta}}(\mathbf{K}_{\boldsymbol{\theta},\boldsymbol{\theta}} + \sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{K}_{\boldsymbol{t},\cdot'}, \tag{13}$$

где для пары векторов $\boldsymbol{a} \in \Theta^l, \boldsymbol{b} \in \Theta^s$ символ $\mathbf{K}_{\boldsymbol{a},\boldsymbol{b}}$ обозначает $l \times s$ матрицу, определенную как $(\mathbf{K}_{\boldsymbol{a},\boldsymbol{b}})_{ij} = k(\boldsymbol{a}_i, \boldsymbol{b}_j)$; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$, а $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$.

Регрессия на основе гауссовского процесса требует выбора функции среднего и ковариационной функции — ядра. Регрессия выполняется в два этапа. На первом из них происходит поиск оптимальных гиперпараметров $\hat{\lambda}$ ковариационной функции k и дисперсии шума $\hat{\sigma_\varepsilon}^2$, дающих максимальное значение правдоподобия $\log p_{\lambda,\sigma_\varepsilon^2}(\boldsymbol{y})$ [133]:

$$\begin{aligned}\hat{\lambda}, \hat{\sigma_\varepsilon}^2 &= \arg\max_{\lambda,\sigma_\varepsilon^2} \log p_{\lambda,\sigma_\varepsilon^2}(\boldsymbol{y}) = \\ &= \arg\max_{\lambda,\sigma_\varepsilon^2} \Bigg\{ -\frac{1}{2}\boldsymbol{y}^\top(\mathbf{K}_{\lambda,\boldsymbol{\theta},\boldsymbol{\theta}} + \sigma_\varepsilon^2\mathbf{I})^{-1}\boldsymbol{y} - \\ &\qquad - \frac{1}{2}\log\left(\det\left(\mathbf{K}_{\lambda,\boldsymbol{\theta},\boldsymbol{\theta}} + \sigma_\varepsilon^2\mathbf{I}\right)\right) - \frac{n}{2}\log(2\pi)\Bigg\}.\end{aligned}$$

Для решения этой оптимизационной задачи обычно используют градиентный спуск с перезапусками.

Вторым этапом регрессии на основе гауссовского процесса является вычисление разложения Холецкого [137] для обращения положительно определенной матрицы $(\mathbf{K}_{\hat{\lambda},\boldsymbol{\theta},\boldsymbol{\theta}} + \hat{\sigma}_\varepsilon^2\mathbf{I})$, что эквивалентно решению следующей линейной системы:

$$(\mathbf{K}_{\hat{\lambda},\boldsymbol{\theta},\boldsymbol{\theta}} + \hat{\sigma}_\varepsilon^2\mathbf{I})\,\boldsymbol{a} = \boldsymbol{b}. \tag{14}$$

После этого апостериорное среднее $\hat{m}(x)$ и функцию ковариации $\hat{k}(x,x)$ можно вычислить по формулам 12 и 13.

Задача оптимизации, включающая решение линейной системы и вычисление определителя, а также последующий шаг вычисления разложения Холецкого имеют вычислительную сложность $\mathrm{O}(n^3)$, где n — размер точек, где была вычислена целевая функция. Поиск $\hat{\lambda}, \hat{\sigma_n}^2$ требуется на каждой итерации процедуры оптимизации, что делает ее узким местом вычислительной сложности метода регрессии. Это не позволяет использовать этот простой подход при больших n. Однако вычисление функций среднего и ковариации имеют оценки $\mathrm{O}(n)$ и $\mathrm{O}(n^2)$ соответственно и легко распараллеливаемы.

Кросс-валидация для выбора параметров априорного гауссовского процесса. Параметры априорного гауссовского процесса такие, как функции

среднего и ковариации, часто выбираются вручную, основываясь на знаниях свойств аппроксимируемой функции. Однако есть способ выполнить этот выбор автоматически с использованием процедуры *кросс-валидации*. Для фиксированного конечного набора значений параметров эта процедура позволяет выбрать наилучшие значения с помощью метрики $L_{\text{LOO}-\text{CV}}$ качества регрессии. Процедура кросс-валидации была использована при настройке гиперпараметров комбинированного метода, основанного на байесовской оптимизации и локальном поиске.

Пусть имеются данные $\{(\theta_i, y_i)\}_{i=1}^n$ для построения регрессии. Для вычисления метрики $L_{\text{LOO}-\text{CV}}$ последовательно перебираются точки θ_i и для каждой происходит построение регрессии гауссовского процесса $\hat{f}_{-i}$ по данным $\boldsymbol{\theta}_{-i}, \boldsymbol{y}_{-i}$ с исключенной точкой θ_i. Для построенной регрессии вычисляется значение функции $Q(y_i, \hat{f}_{-i}(\theta_i))$, которая оценивает качество предсказания $\hat{f}_{-i}(\theta_i)$ в исключенной точке θ_i. Метрика $L_{\text{LOO}-\text{CV}}$ регрессии гауссовского процесса с заданными параметрами определяется, как усреднение значений $Q(y_i, \hat{f}_{-i}(\theta_i))$:

$$L_{\text{LOO}-\text{CV}} = \frac{1}{n}\sum_{i=1}^{n} Q(y_i, \hat{f}_{-i}(\theta_i)).$$

При этом чем больше значение $L_{\text{LOO}-\text{CV}}$, тем лучше качество регрессии.

Функция качества Q может быть определена разными способами. Например, если предположить, что $\hat{f}_{-i}(\theta_i) = N(\mu, \sigma^2)$, то Q может быть определена следующим образом:

$$Q(y_i, \hat{f}_{-i}(t_i)) = -\mathbb{E}(y_i - \hat{f}_{-i}(t_i))^2 = -(y_i - \mu)^2 - \sigma^2.$$

Однако метрика $L_{\text{LOO}-\text{CV}}$ с такой функцией качества будет отдавать предпочтение тем регрессиям, которые более уверены в своем прогнозе: меньшее значение σ^2 напрямую приводит к большему значению Q. В связи с этим была использована другая функция оценки качества предсказания, которая способна лучше сбалансировать качество предсказания со степенью уверенности [133]:

$$Q(y_i, \hat{f}_{-i}(t_i)) = \log p_{\hat{f}_{-i}(t_i)}(y_i) = -\frac{(y_i - \mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2). \qquad (15)$$

В случаях, когда используется функция выбора `LogEI`, значение y_i в формуле (15) заменяется на $\log y_i$.

2.3.2. Реализация разработанного метода, основанного на комбинации байесовской оптимизации и локального поиска

Для реализации разработанного метода, основанного на комбинации байесовской оптимизации и локального поиска, был расширен модуль `optimizers` добавлением классов для баейсовской оптимизации. Модуль `optimizers` был

разработан для реализации аналогичного комбинированного метода на основе генетического алгоритма. Его описание доступно в разделе 2.2.2.

Выделим несколько готовых к использованию библиотек, реализующих байесовскую оптимизацию: GPyOpt [138], BOTorch [139] и SMAC [126, 140]. Они включают реализацию различных суррогатных моделей, например, гауссовских процессов, и функций выбора, которые необходимы при разработке и реализации байесовской оптимизации. Библиотека GPyOpt была одной из широко применяемых библиотек, однако она не поддерживается с 2020 года. Библиотека BOTorch была создана в 2020 году и все еще продолжает активно развиваться. Библиотека SMAC, предложенная в 2011 году, успела хорошо зарекомендовать себя в ряде приложений [141–143] и, что важно, реализует функцию выбора α_{logEI} [134], которая используется в данной работе. Поэтому для реализации байесовской оптимизации была использована библиотека SMAC v0.13.1.

Два класса `SMACBayesianOptimizer` `SMACBOEnsemble` были разработаны и интегрированы в модуль `optimizers` для реализации байесовской оптимизации. На рисунке 54 изображена обновленная структура класса, добавленные классы выделены более жирными линиями.

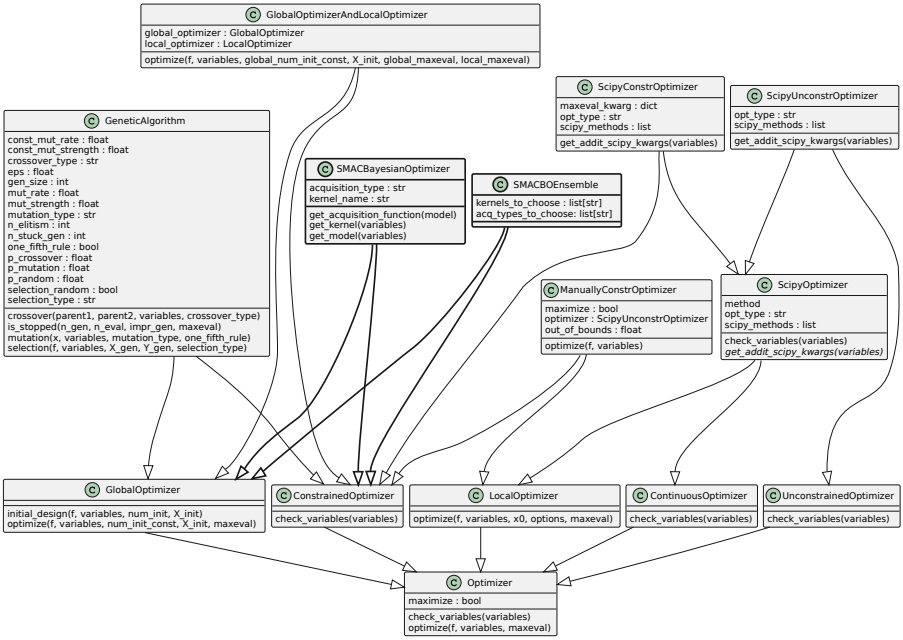


Рисунок 54 – Обновленная структура классов модуля `optimizers`

Класс `SMACBayesianOptimizer` реализует байесовскую оптимизацию с гауссовским процессом в качестве суррогатной модели. Он требует выбран-

ных функции ковариации (атрибут `kernel_type`) и функции выбора (атрибут `acquision_type`). Пример создания байесовской оптимизации с ковариационной функцией, заданной ядром Матерна с гладкостью $\nu = 5/2$, и функцией выбора α_{EI}, представлен на рисунке 55. Дополнительно, класс `SMACBayesianOptimizer` реализует метод автоматического выбора ковариационной функции, что применяется в случае, если атрибут `kernel_type` объекта выбран равным `"Auto"`. Пример создания байесовской оптимизации с функцией выбора α_{PI} и функции ковариации, которая выбирается автоматически, также представлен на рисунке 55.

```python
import gadma

# Классическая байесовская оптимизация
bo1 = gadma.optimizers.SMACBayesianOptimizer(
        kernel="Matern52",
        acquisition_type='EI'
)

# Байесовская оптимизация с автоматическим выбором ядра
bo2 = gadma.optimizers.SMACBayesianOptimizer(
        kernel="Auto",
        acquisition_type='PI'
)
```

Рисунок 55 – Обновленная структура классов модуля `optimizers`

Класс `SMACBOEnsemble` реализует ансамблевую байесовскую оптимизацию. Атрибут `kernels_to_choose` является списком доступных функций ковариации гауссовского процесса для автоматического выбора. По умолчанию метод выбирает из двух функций: ядро Матерна с гладкостью $\nu = 5/2$ и ядро RBF. Атрибут `acq_types_to_choose` является списком функций выбора, включенных в ансамбль, по умолчанию рассматриваются две функции выбора: α_{PI} и α_{LogEI} Подробное описание причин выбора этих функций будет приведено в разделе 2.3.3.

Как и в случае с генетическим алгоритмом, **разработанный комбинированный метод** на основе байесовской оптимизации и локального поиска реализован, как объект класса `GlobalOptimizerAndLocalOptimizer`. Однако, в качестве метода глобальной оптимизации используется ансамблевая байесовская оптимизация — объект класса `SMACBOEnsemble`. Пример реализации разработанного метода представлен на рисунке 56, где дополнительно продемонстрировано применение этого метода для поиска точки минимума функции Розенброка [111]. Пример вывода этой программы представлен на рисунке 57.

```python
import gadma
from scipy.optimize import rosen

# Рассмотрим функцию Розенброка и найдем ее минимум
f = rosen

# На вход она принимает вектор x параметров
x = [1.2, 1.5, 11]
f(x)

# Создадим переменные функции с областью значений
var1 = gadma.variables.ContinuousVariable(name='var1', domain=[-1, 2])
var2 = gadma.variables.ContinuousVariable(name='var2', domain=[0, 10])
var3 = gadma.variables.ContinuousVariable(name='var3', domain=[0.01, 100])
variables = [var1, var2, var3]

# Создадим объект байесовской оптимизации
bo = gadma.optimizers.get_global_optimizer("SMAC_BO_ensemble")
bo.maximize = False

# Создадим объект метода Нелдера-Мида
nm = gadma.optimizers.get_local_optimizer("Nelder-Mead")
nm.maximize = False

# Создаем объект комбинированного метода
optimizer = gadma.optimizers.GlobalOptimizerAndLocalOptimizer(bo, nm)

# Запускаем оптимизацию для нахождения точки минимума
res = optimizer.optimize(
        f,
        variables,
        verbose=1,
        global_maxeval=60,
        local_maxeval=None,
)
```

Рисунок 56 – Применение разработанного комбинированного метода на основе байесовской оптимизации, реализованного с помощью модуля `optimizers`, для поиска точки минимума функции Розенброка

```
--Start global optimization SMAC_BO_combination--
For usual Y:
Kernel was chosen automatically: rbf
For log_transformed Y:
Kernel was chosen automatically: rbf

====================== Iteration 00000 ======================
Initial design:
Fitness function        Parameters
3762.0510755675377      [-0.12328106715341847, 5.072845372791751, 29.177917221313436]
4179.375970808465       [1.1684463396319509, 0.3150880903372544, 6.477831241230526]
5066.367450516042       [-0.9223545173951715, 7.105528015711249, 53.824925445914886]
...
940204.1621961601       [-0.4304586405087313, 0.725223659084494, 97.48846211953487]

************************************************************
Current optimum: 3762.051
On parameters: [-0.12328106715341847, 5.072845372791751, 29.177917221313436]
************************************************************


GP was optimized: True
Current number of points: 50
====================== Iteration 00001 ======================
Got points:
Fitness function        Parameters
1396.9304561528957      [0.35906106399224824, 3.831588417100318, 15.099758341423227]    rbf_PI

Time for GP training: 0.10845613479614258
Time for GP prediction: 0.0005297660827636719
Time for acq. optim.: 0.2217860221862793
Time of evaluation: 6.151199340820312e-05
Total time of iteration: 0.331268310546875
============================================================

************************************************************
Current optimum: 1396.930
On parameters: [0.35906106399224824, 3.831588417100318, 15.099758341423227]     rbf_PI
************************************************************

====================== Iteration 00010 ======================
Got points:
Fitness function        Parameters
75.38737821509756       [1.0910989455948985, 0.36505533702949733, 0.3948097659903245]   rbf_PI

Time for GP training: 0.20359444618225098
Time for GP prediction: 0.0005099773406982422
Time for acq. optim.: 0.1580350399017334
Time of evaluation: 5.984306335449219e-05
Total time of iteration: 0.3625755310058594
============================================================

************************************************************
Current optimum: 48.810
On parameters: [1.9994403876780895, 3.3552342097499537, 11.158602899769543]     rbf_logEI
************************************************************

--Finish global optimization SMAC_BO_combination--
Result:
   status: 0
 success: True
       x: [ 1.99944039  3.35523421 11.1586029 ] rbf_logEI
       y: 48.8101632168688 94
  n_eval: 60
  n_iter: 10

--Start local optimization optimize_fmin--
0       48.810163216868894      (var1=1.99944,  var2=3.35523,   var3=11.1586)
1       48.810163216868894      (var1=1.99944,  var2=3.35523,   var3=11.1586)
2       186.87983765238184      (var1=1.99944,  var2=3.523,     var3=11.1586)
...
155     2.1807501873289364e-09  (var1=1.0,      var2=0.99999,   var3=0.99998)
--Finish local optimization optimize_fmin--
Result:
   status: 0
 success: True
 message: GLOBAL OPTIMIZATION: ; LOCAL OPTIMIZATION:
       x: [0.99999685 0.99998943 0.99998041]
       y: 2.1807501873289364e-09
  n_eval: 336
  n_iter: 165
```

Рисунок 57 – Пример вывода программы, представленной на рисунке 56

2.3.3. Настройка гиперпараметров байесовской оптимизации и разработка ансамблевого метода

Для решения задачи поиска параметров θ модели $\mathcal{M}$ демографической истории популяций на основе генетических данных $\mathfrak{D}$ было разработано несколько методов, основанных на байесовской оптимизации. Для априорного гауссовского процесса ранее было предложено использовать функцию среднего $m(\theta) \equiv 0$. В таких предположениях метод байесовской оптимизации включает два неизвестных гиперпараметра: 1) функцию ковариации гауссовского процесса и 2) функцию выбора. Сначала была выполнена процедура кросс-валидации для оценки качества регрессии гауссовского процесса, построенной на $2\,000$ точках при разных ковариационных функциях. Затем были проведены экспериментальные исследования различных методов байесовской оптимизации на множестве наборов данных. Методы были сравнены с использованием графиков сходимости, и были определены гиперпараметры, обеспечивающие наилучшую сходимость. Сначала был рассмотрен метод классической байесовской оптимизации, который описан в разделе 2.3. Была выявлена эффективность этого метода при различных конфигурациях двух рассматриваемых гиперпараметров. Затем был разработан метод байесовской оптимизации с автоматическим выбором ядра гауссовского процесса и были сравнены его конфигурации с разными функциями выбора. Наконец, на основе проведенных экспериментальных исследований был разработан ансамблевый метод с гиперпараметрами, который показал наилучшую производительность. Было продемонстрировано, что ансамблевый метод является наиболее эффективным среди всех разработанных методов байесовской оптимизации.

Для вычисления правдоподобия был использован метод, реализованный в библиотеке *moments*. Для экспериментальных исследований было использовано одиннадцать наборов данных из пакета `deminf_data v1.0.0`, который ранее использовался для настройки гиперпараметров генетического алгоритма (раздел 2.2.3) и доступен по ссылке `https://github.com/noscode/demographic_inference_data`. Каждый набор данных содержит генетические данные в виде аллель-частотного спектра, модель демографической истории и границы для параметров модели. Названия наборов данных имеют структуру, которая была приведена ранее на рисунке 49. Список использованных наборов данных вместе с соответствующим временем вычисления правдоподобия представлен на рисунке 58. Ось ординат представлена в логарифмической шкале. Диаграммы размаха для времени вычисления правдоподобия, представленные на рисунке, окрашены в соответствии с их медианными значениями.

Сравнение методов байесовской оптимизации в этом разделе было проведено на первых одиннадцати наборах данных. Среди них присутствует один набор данных для одной популяции, четыре набора для двух популяций, два набора для трех популяций, три набора для четырех популяций и один набор данных для пяти популяций. Два последних набора данных (`4_YRI_CEU_CHB_JPT_17_Jou` и `5_YRI_CEU_CHB_JPT_KHV_21_Jou`)

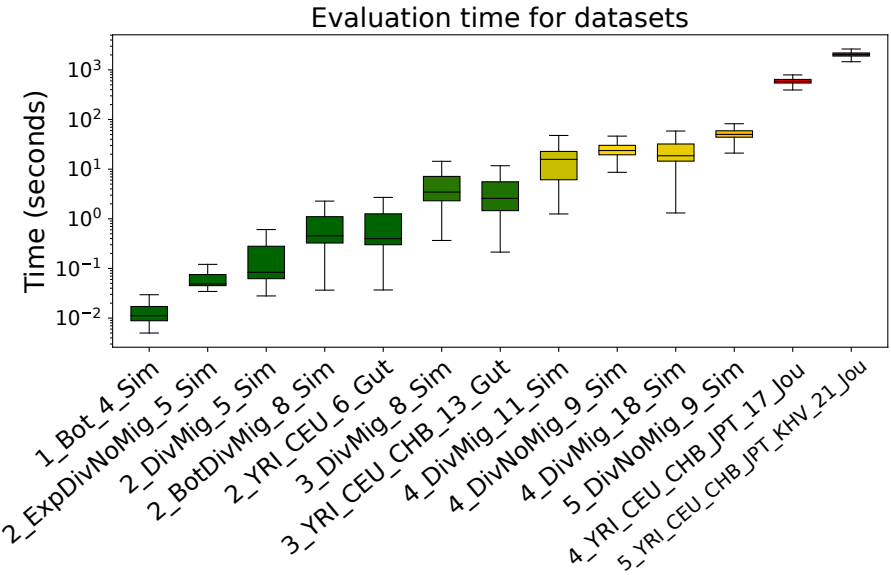


Рисунок 58 – Время вычисления логарифма правдоподобия с помощью *moments* для тестируемых наборов данных из пакета `deminf_data v1.0.0`

были использованы для сравнения ансамблевого метода байесовской оптимизации с генетическим алгоритмом и будут описаны далее.

Эффективность метода байесовской оптимизации зависит от выбора гиперпараметров — ковариационной функции гауссовского процесса и функции выбора. Были рассмотрены три функции выбора:

а) Функция ожидаемого улучшения α_{EI} (Expected Improvement, `EI`),
б) Функция вероятности улучшения α_{PI} (Probability of Improvement, `PI`),
в) Функция ожидаемого улучшения для логарифмированных значений целевой функции α_{logEI} (Log Expected Improvement, `LogEI`).

Были рассмотрены четыре функции ковариации гауссовского процесса:

а) Экспоненциальное ядро $k_{1/2}$ (`Exponential`).
б) Ядро Матерна $k_{3/2}$ с гладкостью $\nu = 3/2$ (`Matern32`).
в) Ядро Матерна $k_{5/2}$ с гладкостью $\nu = 5/2$ (`Matern52`).
г) Гауссовское ядро или ядро RBF k_∞ (`RBF`).

Одним из популярных способов выбора ядра гауссовского процесса для байесовской оптимизации является процедура кросс-валидации (LOO-CV), описанная в разделе 2.3.1.Четыре ядра гауссовского процесса были сравнены на одиннадцати наборах данных с использованием метрики L_{LOO-CV}, полученной по 2 000 случайным точкам. Так как в функции выбора `LogEI` значения целевой функции логарифмированы, то метрика L_{LOO-CV} была вычислена для двух

гауссовских процессов: без применения логарифма и с применением логарифма к целевой функции.

Результаты полученных метрик представлены в работе [3] в таблицах S1 и S2. Они показывают, что экспоненциальная функция ковариации `Exponential` имеет худшие значения метрики, а функция ковариации `Matern52` — лучшие значения L_{LOO-CV} на большинстве наборов данных. Однако были получены исключения, например, метрика L_{LOO-CV} на наборе данных `4_DivMig_18_Sim` имела лучшее значение в случае экспоненциальной функции ковариации `Exponential`. Несмотря на то, что функция `Matern52` во многих случаях имеет наилучшее значение метрики, на некоторых наборах данных она оказывается одной из наихудших, поэтому нельзя выделить явного чемпиона среди тестируемых функций ковариации.

Затем были проведены экспериментальные исследования для классической байесовской оптимизации. Двенадцать конфигураций с разными ядрами и функциями выбора были сравнены на одиннадцати наборах данных. Каждая конфигурация была обозначена, как `(функция выбора)+(функция ковариации)` — обозначение `PI+RBF` соответствует классической байесовской оптимизации с функцией выбора α_{PI} и с гауссовским процессом, имеющим ядро RBF. Для каждого набора данных и конфигурации гиперпараметров было проведено 64 независимых запуска. Полученные графики сходимости для первых 200 итераций могут быть найдены в работе [3] на рисунках S1 – S3. Примеры двух графиков представлены на рисунке 59. Каждая итерация соответствует одному вычислению целевой функции правдоподобия $f^{moments}$. Сплошные линии отображают медианы выборок из 64 для каждой конфигурации, а закрашенные области соответствуют диапазону между первой и третьей квартилями. Серая область на рисунках отображает процедуру начального дизайна — генерацию начальных точек случайным образом перед запуском байесовской оптимизации. Метки в легендах отсортированы в соответствии с медианными значениями на последней итерации.

Кандидаты конфигураций, которые имеют либо ковариационную функцию `Exponential`, либо функцию выбора `EI`, показали худшую сходимость для всех одиннадцати наборов данных. Ковариационная функция `Matern52` продемонстрировала лучшую эффективность по сравнению с функцией `Matern32` в большинстве случаев. На основе построенных графиков сходимости классической байесовской оптимизации были выделены четыре конфигурации, которые показали наилучшую сходимость. Они являются комбинацией ковариационных функций `Matern52` или `RBF` и функций выбора `PI` или `LogEI`:

– конфигурация `Matern52+LogEI`;
– конфигурация `Matern52+PI`;
– конфигурация `RBF+LogEI`;
– конфигурация `RBF+PI`.

Отметим, что результаты полученные по графикам сходимости не согласуются полностью с результатами кросс-валидации. Заметим, что графики схо-

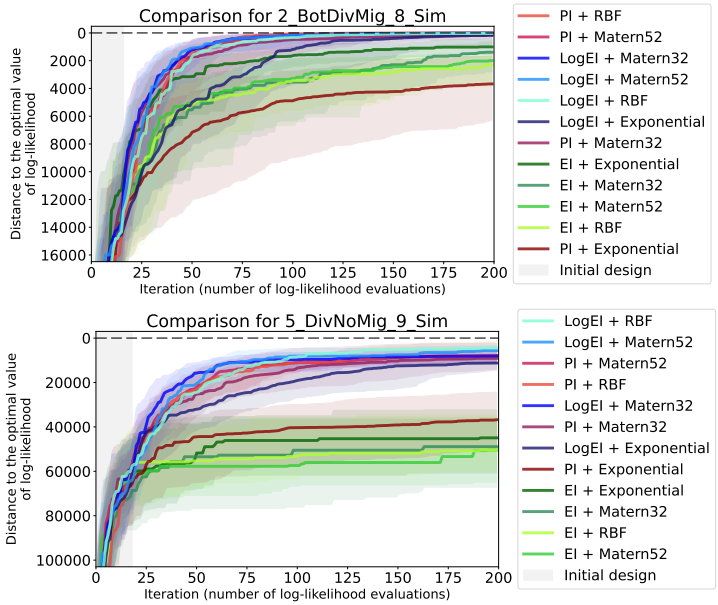


Рисунок 59 – Примеры графиков сходимости двенадцати конфигураций классической байесовской оптимизации для датасетов двух и пяти популяций

димости являются более надежным сравнением, так как они отображают реальную сходимость метода для разных конфигураций на данных. Например, в случаях, когда метрика L_{LOO-CV} указывает на то, что ковариационная функция `Exponential` является лучшим выбором, графики сходимости демонстрируют обратное. Однако оба сравнения сходятся в том, что функция ковариации `Exponential` является наименее эффективной для решения поставленной задачи.

Далее был разработан и протестирован метод байесовской оптимизации с *автоматическим выбором функции ковариации*. Перед запуском первой итерации байесовской оптимизации выполняется процедура начального дизайна — генерации точек случайным образом. Разработанный метод вычисляет метрику кросс-валидации L_{LOO-CV} на этом наборе начальных точек и выбирает функцию ковариации с наибольшим значением метрики. Основываясь на результатах экспериментальных исследований двенадцати конфигураций классической байесовской оптимизации, были отобраны две наиболее эффективные функции выбора: `PI` и `LogEI`. Метод байесовской оптимизации с автоматическим выбором функции ковариации был протестирован для каждой из них. Методы были обозначены, как `PI+Auto` и `LogEI+Auto` соответственно. Было проведено сравнение двух конфигураций между собой, а также с предыдущими лучшими конфигурациями классической байесовской оптимизации.

Для каждого набора данных были построены гистограммы частоты выбора рассматриваемых функций ковариации (рисунок 60). Согласно гистограммам, частота выбора функций различается в зависимости от набора данных, однако `RBF` и `Matern52` выбираются наиболее часто.

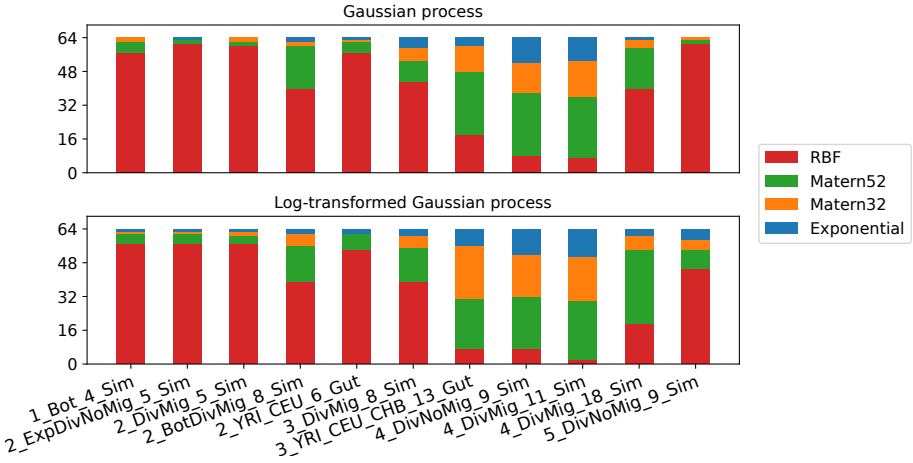


Рисунок 60 – Гистограммы частоты выбора функций ковариации при применении баесовской оптимизации с автоматическим выбором функции ковариации

Были построены графики сходимости для первых 200 итераций для конфигураций `PI+Auto` и `LogEI+Auto` метода с автоматическим выбором функции ковариации. Эти графики представлены на рисунках S4 – S6 в работе [3]. Два примера графиков представлены на рисунке 61. Полученные результаты демонстрируют, что конфигурации имеют равную или лучшую сходимость по сравнению с наилучшими конфигурациями классической байесовской оптимизации. Однако, определение однозначного победителя среди `PI+Auto` и `LogEI+Auto` представляет сложность. Например, для трех наборов данных — `3_DivMig_8_Sim`, `4_DivNoMig_9_Sim` и `4_DivMig_11_Sim` — метод `LogEI+Auto` показывает лучшую производительность по сравнению с методом `PI+Auto`. Для двух наборов данных — `3_YRI_CEU_CHB_13_Gut` и `5_DivNoMig_9_Sim` — результаты оказываются противоположными. Таким образом, невозможно однозначно определить превосходство какой-либо функции выбора для метода байесовской оптимизации с автоматическим выбором функции ковариации.

Комбинирование подходов и разработка ансамблевых методов является перспективным направлением при решении задач оптимизации. Совместное использование нескольких методов может привести к значительно более эффективным результатам. Например, ансамблевый метод байесовской оптимизации

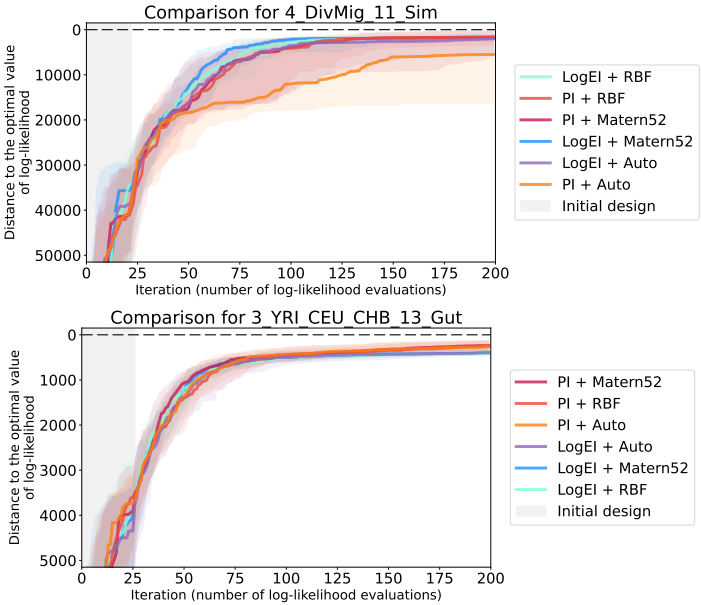


Рисунок 61 – Примеры графиков сходимости двух конфигураций байесовской оптимизации с автоматическим выбором функции ковариации и четырех наилучших конфигураций классической байесовской оптимизации для наборов данных трех и четырех популяций

под названием Squirrel [144] стал одним из победителей конкурса *Black-box Optimization Challenge* в 2020 году.

Основываясь на этих идеях, был разработан ансамблевый метод `Ensemble` байесовской оптимизации для настройки параметров моделей демографической истории по генетическим данным. Он представляет собой байесовскую оптимизацию с автоматическим выбором функции ковариации и с ансамблем из двух функций выбора (`PI` и `LogEI`). На каждой итерации равновероятно выбирается одна из функций `PI` и `LogEI`, и выполняется поиск новой точки для вычисления целевой функции. Набор функций ковариации для автоматического выбора был ограничен двумя наиболее эффективными и часто выбираемыми: `RBF` и `Matern52`. Заметим, что функции ковариации для рассматриваемых функций выбора определяются разными способами: для `PI` регрессия гауссовского процесса строится для целевой функции $f^{moments}$ в то время, как для `LogEI` регрессия использует логарифмированные значения $\log f^{moments}(\theta)$.

Были построены графики сходимости метода `Ensemble` и было выполнено сравнение с двумя наилучшими конфигурациями `PI+Auto` и `LogEI+Auto`. Полученные графики приведены в [3] на рисунках S7 – S9. Два примера графи-

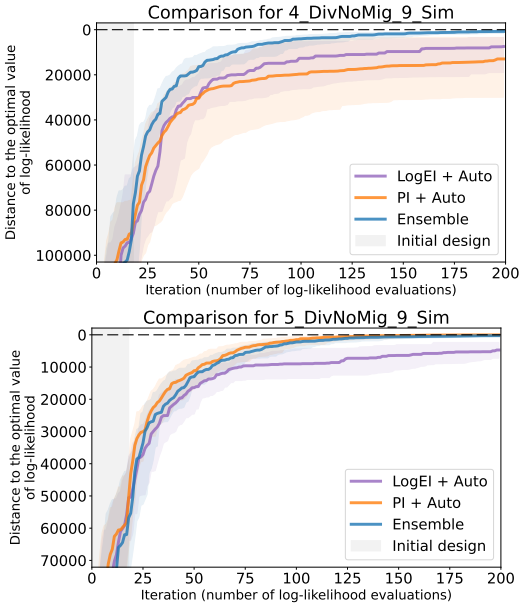


Рисунок 62 – Примеры графиков сходимости ансамблевого метода байесовской оптимизации и двух конфигураций метода с автоматическим выбором функции ковариации для наборов данных четырех и пяти популяций

ков представлены на рисунке 62. Они демонстрируют, что разработанный метод `Ensemble` имеет наилучшую сходимость среди рассмотренных методов на всех наборах данных. Метод `Ensemble` был выбран в качестве финального метода байесовской оптимизации для настройки параметров моделей демографической истории. На некоторых наборах данных он показал значительные улучшения в сходимости по сравнению с другими байесовскими оптимизациями, например, на `4_DivNoMig_9_Sim` (рисунок 62) и на `4_DivMig_11_Sim`.

2.4. Экспериментальные исследования разработанного метода настройки параметров моделей, основанного на комбинации генетического алгоритма и локального поиска для данных одной, двух и трех популяций

Были проведены экспериментальные исследования для определения эффективности разработанного метода для настройки параметров моделей демографической истории по генетическим данным (раздел 2.2), который основан на комбинации генетического алгоритма и локального поиска. Разработанный метод был сравнен с существующими методами настройки параметров моделей первого класса, основанных на методах локальной оптимизации, на симулированных и реальных данных одной, двух и трех популяций. Также метод был при-

менен для настройки параметров моделей расширенного класса, которые включают дискретные параметры динамики изменения численности. Результаты настройки параметров различных моделей, включая расширенные, с использованием набора существующих методов вычисления правдоподобия были получены автором, а их сравнение приведено ниже. Для реальных данных, которые ранее были проанализированы в других исследованиях, были получены демографические истории с лучшим значением правдоподобия и информационного критерия Акаике, чем было получено ранее. В данном разделе приведены результаты этих экспериментальных исследований.

2.4.1. Сравнение с существующими методами настройки параметров на симулированных данных одной, двух и трех популяций

Были проведены экспериментальные исследования на симулированных данных для демонстрации расширенного класса моделей и для выявления эффективности метода настройки параметров, основанного на комбинации генетического алгоритма и локального поиска. Три набора данных были симулированы с использованием следующих демографических историй:

а) История «бутылочного горлышка» для одной популяции,
б) Разделение предковой популяции с асимметричной миграцией между двумя популяциями-потомками,
в) Вторичный контакт с симметричной миграцией для трех популяций после разделения одной из двух популяций-потомков.

Для каждой истории были симулированы данные в виде аллель-частотных спектров с использованием *moments*. Каждый аллель-частотный спектр имел размер в 20 гаплоидных хромосом на популяцию. Длина последовательности равна 10^8 пар оснований, вероятность мутации равна $1{,}25 \cdot 10^{-8}$ на одну позицию генома на одно поколение.

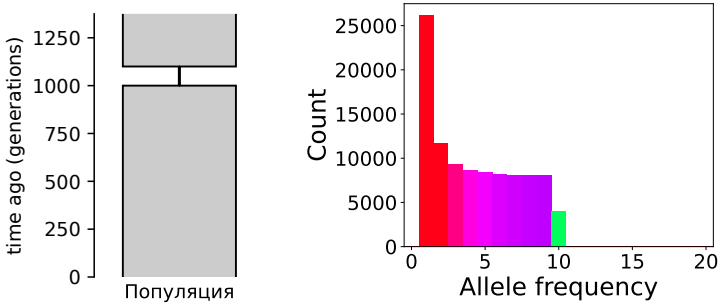
Для каждого набора данных были сравнены три метода настройки параметров: 1) метод Пауэлла с перезапусками, доступный в *moments*, 2) *moments-pipeline*, который реализует набор раундов метода Нелдера-Мида с последовательными запусками, 3) GA+P — разработанный метод на основе комбинации генетического алгоритма и локального поиска — метода Пауэлла. Число перезапусков в первом методе Пауэлла и число раундов для второго метода *dadi pipeline* были выбраны таким образом, чтобы среднее число вычислений целевой функции совпадало со средним значением, полученным для GA+P. Например, в случае одной популяции для метода Пауэлла было использовано 40 перезапусков, а для *moments-pipeline* были использованы пять раундов. Первые четыре раунда включали по 10 запусков метода Нелдера-Мида в каждом, последний пятый раунд включал 20 запусков.

Каждый метод был повторен 50 раз для одной и двух популяций и 10 раз для трех популяций. Для каждого метода были записаны среднее время одного запуска, среднее число вычислений целевой функции, среднее и стандарт-

ное отклонение финального логарифма правдоподобия и наилучшее значение логарифма правдоподобия среди повторов. Было также зарегистрировано среднее время одного запуска каждой оптимизации, однако оно представлено только в информационных целях. Вычислительные ресурсы разных методов следует сравнивать, используя число вычислений целевой функции, так как ни у одного метода нет накладных расходов — вычисления целевой функции занимают почти все время работы метода, а время вычисления правдоподобия может сильно варьироваться в зависимости от значений параметров.

В случае каждого набора симулированных данных вывод демографической истории был проведен для двух моделей. Первая модель — модель первого класса с непрерывными параметрами, которая способна воспроизвести историю, используемую в симуляции. Для этой модели было выполнено сравнение всех трех методов настройки параметров. Вторая использованная модель — новая расширенная модель, которая не только способна восстановить первоначальную историю, но и дополнительно имеет дискретные параметры динамики изменения численности. Для настройки параметров второй модели использовался исключительно метод GA+P.

Для одной популяции была выбрана демографическая история «бутылочное горлышко»: эффективный размер предковой популяции составлял 10 000 особей, затем 1 100 поколений назад произошло «бутылочное горлышко» — размер популяции составлял 100 особей в течение 100 поколений, после чего численность популяции снова увеличилась до 10 000 особей. Рисунок 63 демонстрирует описанную демографическую историю (рисунок 63а), а также симулированные генетические данные в виде аллель-частотного спектра (рисунок 63б). Истинная демографическая история имеет серый цвет, чтобы отобразить тот факт, что она использовалась в симуляции.



(а) (б)

Рисунок 63 – Демографическая история одной популяции, которая была использована для симуляции данных и симулированные генетические данные в виде аллель-частотного спектра

Две модели, представленные на рисунке 64, были использованы для симулированных данных одной популяции: 1) модель 1 из первого класса моделей с пятью непрерывными параметрами (рисунок 64а), 2) модель 2 из расширенного класса моделей с пятью непрерывными параметрами и двумя дискретными параметрами динамики (рисунок 64б).

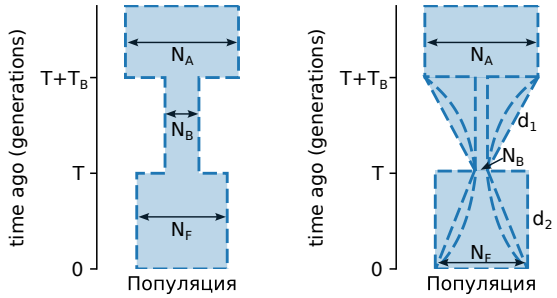


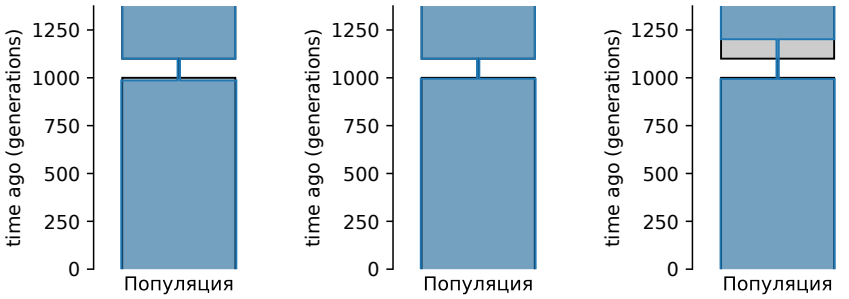
Рисунок 64 – Используемые модели для сравнения методов на симулированных данных одной популяции

Для каждого повтора метод Пауэлла был запущен 40 раз, а *moments-pipeline* состоял из пяти раундов: первые четыре раунда включали по 10 запусков метода Нелдера-Мида в каждом, последний пятый раунд состоял из 20 запусков. Результаты представлены в таблице 6. Изображения полученных демографических историй для модели 1 показаны на рисунке 65: а) метод Пауэлла с перезапусками, б) метод *moments-pipeline*, в) метод GA+P. На рисунке 66 приведены две альтернативные демографические истории для модели 2, полученные методом GA+P. Все истории, изображенные на рисунках 65 и 66, наложены на истинную демографическую историю, использованную в симуляциях, для демонстрации разницы.

Для модели 1 метод GA+P показал лучшее среднее и стандартное отклонение логарифма правдоподобия, однако метод *moments-pipeline* показал лучшее значение максимального правдоподобия среди трех оптимизаций. Для модели 2 метод GA+P показал наилучшие среднее и стандартное отклонение логарифма вероятности по 50 повторам, а также предоставил две альтернативные истории со схожими значениями правдоподобия. Первая из них — вариант 1 включает экспоненциальное уменьшение численности популяции и внезапный рост после, а вторая — вариант 2 имеет постоянные динамики численности популяции и похожа на истинную историю. Данный результат является следствием того, что одному аллель-частотному спектру может соответствовать несколько демографических историй. Похожие демографические истории были представлены в [145], как пример историй, которые имеют одинаковые ожидаемые спектры.

Таблица 6 – Результаты 50 повторов тестируемых методов настройки параметров для моделей демографической истории одной популяции

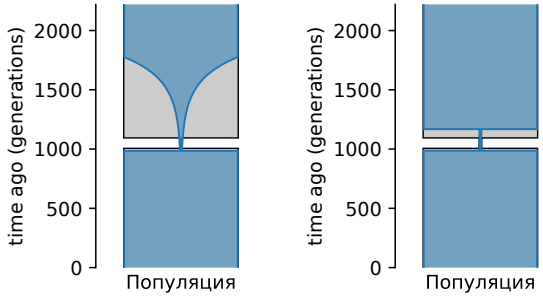
		Модель 1			Модель 2	
	Истинные значения	Метод Пауэлла с перезапусками	*moments-pipeline*	GA+P	GA+P Вариант 1	GA+P Вариант 2
Среднее время	–	$06^{\text{мин}}40^{\text{сек}}$	$05^{\text{мин}}22^{\text{сек}}$	$07^{\text{мин}}25^{\text{сек}}$	$12^{\text{мин}}54^{\text{сек}}$	
Среднее число итераций	–	7 860	7 951	7 515	13 688	
Среднее $f^{moments}$	–	−293,4815	−338,1879	−132,9586	−97,9999	
Ст. откл. $f^{moments}$	–	136,2765	116,9429	99,9209	43,5008	
Лучшее $f^{moments}$	−88,5603	−88,6202	−88,5780	−88,5832	**−88,5616**	−88,5711
N_A	10000	9999	10028	10038	10034	10018
N_B	100	219	203	201	178^e	180
N_F	10000	10180	10159	10094	10050	10011
T_B	100	112	103	205	791	182
T	1000	971	975	978	985	986



(а) (б) (в)

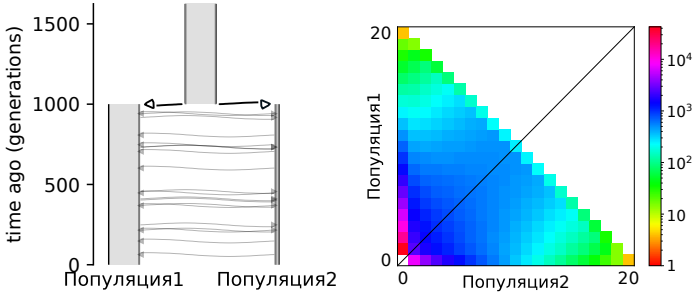
Рисунок 65 – Демографические истории, полученные путем настройки параметров модели 1 разными методами

Для двух популяций была выбрана следующая демографическая история: эффективный размер предковой популяции составлял 10 000, 1 000 поколений назад она разделилась на две популяции с размерами 10 000 особей (Популяция 1) и 1 000 особей (Популяция 2). Непрерывная миграция между популяциями равна: $2{,}5\times10^{-4}$ особей из популяции 2 в популяцию 1 за поколение и $1{,}25\times10^{-4}$ особей из популяции 1 в популяцию 2 за поколение. Рисунок 67 демонстрирует описанную демографическую историю (рисунок 67а), а также симулированные генетические данные в виде аллель-частотного спектра (рисунок 67б). Метод Пауэлла имел шесть перезапусков, а метод *moments-pipeline* был запущен для четырех раундов первые три раунда включали по 10 запусков метода Нелдера-Мида в каждом, последний четвертый раунд состоял из 20 запусков. Две использованные модели изображены на рисунке 68.



(а) (б)

Рисунок 66 – Демографические истории, полученные путем настройки параметров модели 2 методом GA+P



(а) (б)

Рисунок 67 – Демографическая история двух популяции, которая была использована для симуляции данных и симулированные генетические данные в виде аллель-частотного спектра

Результаты представлены в таблице 7. Изображения полученных демографических историй для модели 1 показаны на рисунке 69: а) метод Пауэлла с перезапусками, б) метод *moments-pipeline*, в) метод GA+P. На рисунке 70 приведена демографическая история для модели 2, полученная методом GA+P. Все истории, изображенные на рисунках 69 и 70, наложены на истинную демографическую историю, использованную в симуляциях, для демонстрации разницы.

Все методы для всех моделей предоставили демографические истории, близкие к истинной, которая была использована для симуляции данных. Методы Пауэлла и GA+P для модели 1 показали наиболее точные результаты. Метод Пауэлла в среднем продемонстрировал лучшие результаты, чем метод GA+P. Ме-

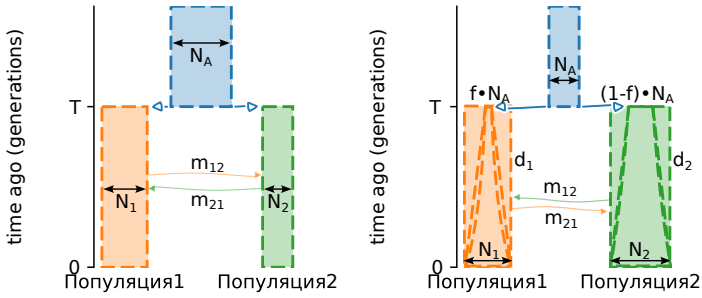


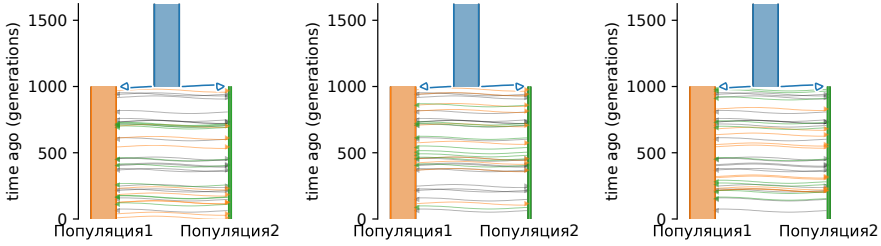
Рисунок 68 – Используемые модели для сравнения методов на симулированных данных двух популяций

Таблица 7 – Результаты 50 повторов тестируемых методов настройки параметров для моделей демографической истории двух популяций

		Модель 1			Модель 2
	Истинные значения	Метод Пауэлла с перезапусками	*moments-pipeline*	GA+P	GA+P
Среднее время	–	$30^{\text{мин}}39^{\text{сек}}$	$55^{\text{мин}}12^{\text{сек}}$	$30^{\text{мин}}34^{\text{сек}}$	$223^{\text{мин}}24^{\text{сек}}$
Среднее число итераций	–	7626	8928	7437	17136
Среднее $f^{moments}$:	–	−1310,948	−1311,016	−1311,269	−1321,195
Ст. откл. $f^{moments}$	–	0,027	0,241	0,451	13,861
Лучшее $f^{moments}$	−1310,931	**1310,931**	−1310,932	**−1310,931**	−1310,983
N_A	10000	10000	10001	10000	10001
N_1	10000	10000	9992	10003	10007
N_2	1000	1000	1000	1000	997
$m_{12}(\times 10^{-4})$	2.50	2.50	2.50	2.50	2.50
$m_{21}(\times 10^{-4})$	1.25	1.25	1.25	1.25	1.26
T	1000	1000	1000	1000	996

тод GA+P корректно восстановил константные динамики изменения численности при настройке параметров модели 2.

Для симуляции данных трех популяций была использована следующая демографическая история. Эффективный размер предковой популяции составлял 10 000, затем 3 000 поколений назад она разделилась на две субпопуляции размером 15 000 (популяция 1) и 5 000 (популяция 2 + популяция 3). Затем вторая субпопуляция разделилась на две новые популяции размером 5 000 (популяция 2) и 10 000 (популяция 3). Непрерывные миграции имели начало только 1 000 поколений назад сразу после второго разделения и являются симметричными: $0{,}25 \times 10^{-4}$ особей на поколение между популяцией 1 и популяцией 2, $0{,}5 \times 10^{-4}$ между популяцией 1 и популяцией 3 и $1{,}5 \times 10^{-4}$ между популяци-



(а) (б) (в)

Рисунок 69 – Демографические истории двух популяций, полученные путем настройки параметров модели 1 разными методами

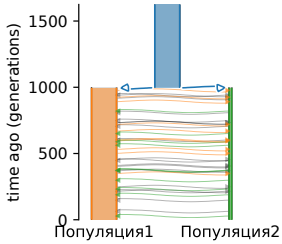
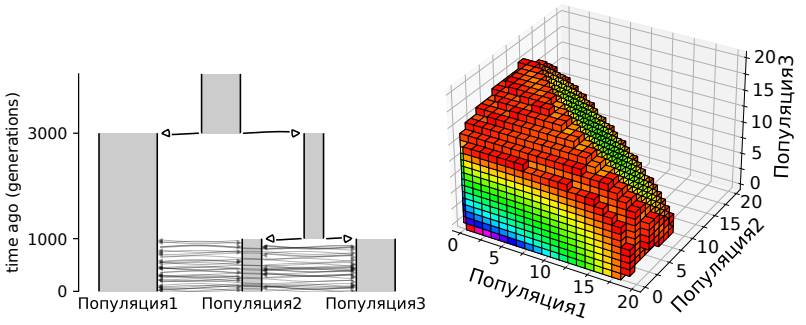


Рисунок 70 – Демографическая история двух популяций, полученная путем настройки параметров модели 2 методом GA+P

ей 2 и популяцией 3. Рисунок 71 демонстрирует описанную демографическую историю (рисунок 71а), а также симулированные генетические данные в виде аллель-частотного спектра (рисунок 71б).

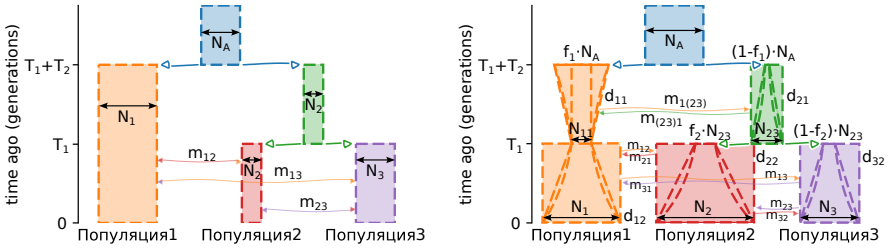
Две модели, представленные на рисунке 72 были использованы для симулированных данных трех популяции: 1) модель 1 из первого класса моделей с десятью параметрами (рисунок 72а), 2) модель 2 из расширенного класса моделей с 18 непрерывными параметрами и пятью дискретными параметрами динамики (рисунок 72б). Метод Пауэлла имел 12 перезапусков, а метод *moments-pipeline* был запущен для шести раундовЖ первые пять раундов включали по 10 запусков метода Нелдера-Мида в каждом, последний цестой раунд состоял из 20 запусков.

Результаты представлены в таблице 8. Изображения полученных демографических историй для модели 1 показаны на рисунке 73: а) метод Пауэлла с перезапусками, б) метод *moments-pipeline*, в) метод GA+P. На рисунке 74 приведена демографическая история для модели 2, полученная методом GA+P. Все истории, изображенные на рисунках 73 и 74 наложены на истинную демографическую историю, использованную в симуляциях, для демонстрации разницы.



(а) (б)

Рисунок 71 – Демографическая история трех популяции, которая была использована для симуляции данных и симулированные генетические данные в виде аллель-частотного спектра



(а) (б)

Рисунок 72 – Используемые модели для сравнения методов на симулированных данных трех популяций

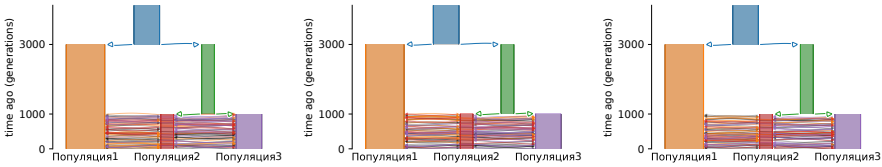
В случае модели 1 для трех популяций метод GA+P продемонстрировал наилучшие значения максимального, среднего и стандартного отклонения значения правдоподобия среди всех наблюдаемых методов. Настройка параметров модели 2 успешно реконструировала симметричные миграции, а дополнительные параметры оказались близки к истинным значениям. Однако эта история имеет ненулевую миграцию после раскола предковой популяции, что может быть связано с малым числом повторов. Все предсказанные динамики, несмотря на то, что не все они являются константными, показывают, что размер популяций оставался почти одинаковым.

Таблица 8 – Результаты 10 повторов тестируемых методов настройки параметров для моделей демографической истории трех популяций

		Модель 1			Модель 2
	Истинные значения	Метод Пауэлла с перезапусками	*moments-pipeline*	GA+P	GA+P
Среднее время	–	$28^{ч}28^{мин}13^{сек}$	$45^{ч}09^{мин}30^{сек}$	$27^{ч}16^{мин}08^{сек}$	$45^{ч}05^{мин}00^{сек}$
Среднее число вычислений	–	22475	19452	21651	71534
Среднее $f^{moments}$	–	−11178,62	−11179,82	−11178,45	−11250,09
Ст. откл. $f^{moments}$	–	0,40	0,72	0,15	123,72
Лучшее $f^{moments}$	−11178,28	−11178,31	−11178,59	−11178,29	−11180,28
N_A	10000	10003	9988	9996	9997
N_{11}	$= N_1$ (15000)	NA	NA	NA	14612
N_{23}	$= N_2$ (5000)	NA	NA	NA	4801
$m_{1(23)}(\times 10^{-4})$	NA (0)	NA	NA	NA	0.14
$m^{(23)1}(\times 10^{-4})$	NA (0)	NA	NA	NA	0.27
N_1	15000	15018	15029	15004	15567^{exp}
N_2	5000	4992	5011	5008	5032
N_3	10000	9984	9850	10026	10050
$m_{12}(\times 10^{-4})$	0,25	0,25	0,24	0,25	0,23
$m_{13}(\times 10^{-4})$	0,50	0,50	0,50	0,50	0,47
$m_{21}(\times 10^{-4})$	$= m_{12}$ (0,25)	NA	NA	NA	0,21
$m_{23}(\times 10^{-4})$	1,50	1,51	1,55	1,50	1,54
$m_{31}(\times 10^{-4})$	$= m_{13}$ (0,50)	NA	NA	NA	0,44
$m_{32}(\times 10^{-4})$	$= m_{23}$ (1,50)	NA	NA	NA	1,60
T_2	2000	1998	1996	1999	2055
T_1	1000	1000	1009	1000	1020

* 9/10 повторов. Наихудший повтор имел максимальное значение логарифма правдоподобия, равное −13801,48.

^{exp} — экспоненциальное изменение численности.



(а) (б) (в)

Рисунок 73 – Демографические истории трех популяций, полученные путем настройки параметров модели 1 разными методами

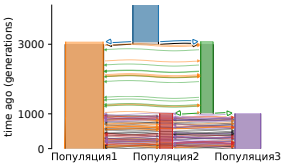


Рисунок 74 – Демографическая история трех популяций, полученная путем настройки параметров модели 2 методом GA+P

2.4.2. Сравнение с существующими методами настройки параметров моделей на данных популяций кошачьей лягушки

Разработанный метод, основанный на комбинации генетического алгоритма и локального поиска, был сравнен с методом множественного запуска метода Нелдера-Мида, реализованным в программном средстве *dadi-pipeline* [49]. В статье [49], в которой был впервые представлен *dadi-pipeline*, его эффективность была продемонстрирована на данных для кошачьей лягушки (*Scotobleps gabonicus*). Разработанный комбинированный метод был запущен для настройки параметров моделей, использованных в статье [49], на тех же данных. Результаты, полученные с помощью разработанного метода, были сравнены с результатами из статьи, полученными с использованием *dadi-pipeline*.

Авторами работы [49] были собраны генетические данные 84 особей кошачьей лягушки из 33 локаций в известной зоне обитания вида на юго-востоке Нигерии, в Камеруне, Экваториальной Гвинее, Габоне и Республике Конго (рисунок 75). Были выделены две основные группы-популяции: 1) северная (Northern)

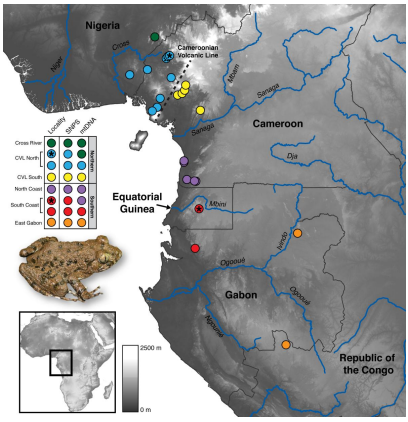


Рисунок 75 – Географическое расположение образцов генетических данных. Источник: [49]

и 2) южная (Southern). Северная группа дополнительно была разделена на три субпопуляции: 1) северная камерунская вулканическая линия CVLN, 2) южная камерунская вулканическая линия CVLS, 3) популяция CrossRiver.

Три аллель-частотных спектра были построены в [49] для разных пар популяций: 1) спектр размера 41×19 для северной (Northern) и южной (Southern) популяций; 2) спектр размера 31×19 для популяций CVLN и CVLS; и 3) спектр размера 15×31 для популяций CrossRiver и CVLN. При построении аллель-частотных спектров были использованы мутации только на независимых позициях, поэтому считаем, что зависимостей в данных нет. На рисунке 76 представлены изображения использованных аллель-частотных спектров.

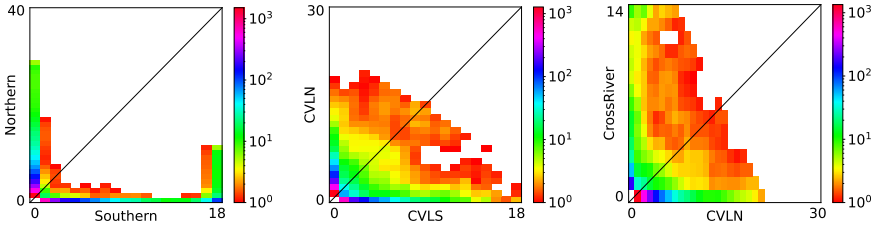


Рисунок 76 – Генетические данные различных пар популяций кошачьей лягушки в виде аллель-частотных спектров

Для каждого из трех наборов генетических данных были выведены демографические истории с использованием двенадцати моделей двух популяций из каталога *dadi-pipeline*, описание которых приведено в работе [1] на странице 4 дополнительных материалов. Для вычисления значения правдоподобия был использован метод ∂a∂i с размером сетки $pts = [50, 60, 70]$. Для каждой модели было найдено максимальное значение правдоподобия с помощью разработанного метода, основанного на генетическом алгоритме. Модели были отсортированы по значению информационного критерия Акаике и результаты были сравнены с результатами, полученными в статье [49] с использованием *dadi-pipeline*. Результаты представлены в работе [1] в таблице S2 для популяций Northern, Southern, в таблице S3 для популяций CVLN, CVLS, в таблице S4 для популяций CrossRiver, CVLN.

В 33 случаях из 36 (92%) правдоподобие, полученное с использованием разработанного метода на основе комбинации генетического алгоритма и локального поиска, оказалось лучше, чем значение, полученное с использованием *dadi-pipeline*. В двух случаях (5%) правдоподобие совпало с результатами из [49]. Только одна модель (no_mig_size) для данных популяций CrossRiver и CVLN получила максимальное значение правдоподобия хуже, чем было получено ранее (таблица S4 в работе [1]). Эффективность разработанного метода позволила построить более корректные сравнения моделей с помощью критерия Акаике и

выбрать лучшие. На рисунке 77 представлены финальные демографические истории популяций кошачьей лягушки. Так как длина генома кошачьей лягушки неизвестна [49], то все значения параметров моделей были получены относительно размера N_A предковой популяции, который остался неизвестным.

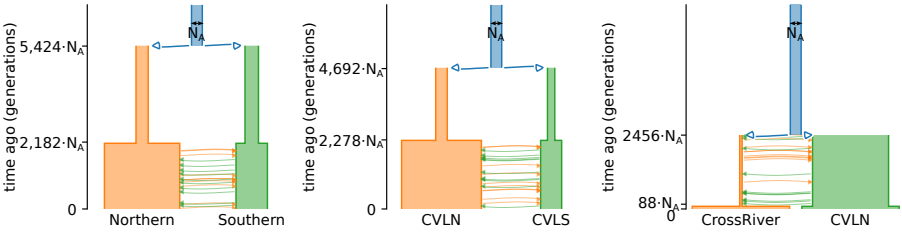


Рисунок 77 – Полученные демографические истории для различных пар популяций кошачьей лягушки

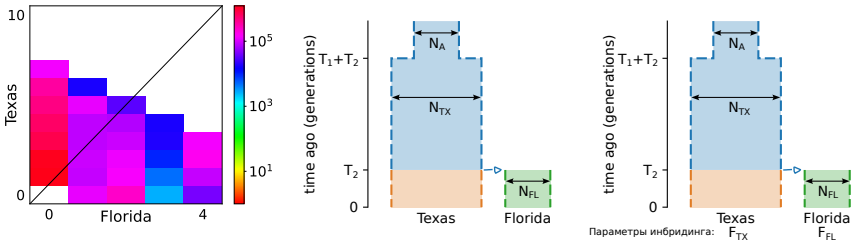
2.4.3. Сравнение с существующими методами настройки параметров моделей на данных двух популяций американской пумы

Разработанный метод настройки параметров моделей, основанный на комбинации генетического алгоритма и локального поиска, был сравнен с методами BFGS и BOBYQA, доступными в программном обеспечении *∂a∂i*. Эффективность методов была сравнена на наборе данных двух популяций американской пумы (*Puma concolor*). Генетические данные пяти особей техасской популяции (Texas) и две особи флоридской популяции (Florida) были представлены в работе [146], демографическая история была выведена в статье [51].

Вероятность мутации американской пумы составляет $\mu = 2.2 \times 10^{-9}$ для одной позиции в геноме за одно поколение. Среднее время на одно поколение равно трем годам, а длина последовательности, представленной в генетических данных, равна $2\,564\,692\,624$ пар оснований [146].

Для вывода демографической истории популяций американской пумы были использованы две модели. Первая модель описывает демографическую историю увеличения численности техасской популяции и отделения от нее флоридской популяции, имеющей постоянный размер. Вторая модель расширяет первую модель и дополнительно включает параметры инбридинга для каждой из двух популяций. На рисунке 78 изображены генетические данные в виде аллель-частотного спектра (рисунок 78а) и две описанные модели. На рисунке 78б изображена модель 1 без инбридинга, на рисунке 78в — модель 2 с инбридингом.

Сравнение эффективности методов было проведено с использованием равных границ на значения параметров модели, которые были применены



(а) (б) (в)

Рисунок 78 – Генетические данные двух популяций американской пумы (а) и рассматриваемые модели демографической истории (б, в)

в исходной работе [51]. Значения размеров популяций лежали в интервале $[10^{-2} \cdot N_A, 10 \cdot N_A]$, параметров времени — в интервале $[0{,}2 \cdot N_A, 20 \cdot N_A]$, а параметров инбридинга — в $[10^{-5}, 1 - 10^{-5}]$, где N_A — размер общей предковой популяции (рисунок 78в). Для вычисления правдоподобия был использован метод $\partial a \partial i$ с размером сетки, равным $pts = \{40, 50, 60\}$.

Всего было сравнено пять методов настройки параметров модели. Разработанный комбинированный метод (GA+BFGS) настройки параметров модели, основанный на комбинации генетического алгоритма и локального поиска — метода BFGS, был сравнен с четырьмя существующими методами: 1) BFGS, 2) BFGS с перезапусками, 3) метод BOBYQA, 4) метод BOBYQA с перезапусками. Число перезапусков у второго и четвертого методов было выбрано таким образом, чтобы среднее число вычислений целевой функции было близко к среднему числу вычислений метода GA+BFGS. Метод BFGS с перезапусками имел 18 и 16 запусков для модели 1 и модели 2 соответственно, а метод BOBYQA с перезапусками потребовал шести и четырех запусков соответственно. Каждый метод был повторен 100 раз и были сравнены наилучший результат, среднее и стандартное отклонения финальных значений правдоподобия.

Результаты для модели 1 без инбридинга представлены в таблице 9. Результаты для модели 2 с инбридингом представлены в таблице 10. Жирным шрифтом выделены наилучшие результаты.

Разработанный метод GA+BFGS показал наилучшее среднее значение правдоподобия среди рассматриваемых методов для обеих моделей. Метод BOBYQA продемонстрировал наилучшее максимальное значение правдоподобия. Однако стоит отметить, что метод GA+BFGS предоставил похожие результаты. Для модели 1 максимальное значение правдоподобия, полученное методом GA+BFGS, составляет $-452\,969{,}48$, а методом BOBYQA — $-452\,969{,}40$. Для модели 3 максимальное значения правдоподобия равны $-317\,239{,}49$ и $-317\,239{,}48$ для методов GA+BFGS и BOBYQA соответственно. Метод BFGS

Таблица 9 – Результаты 100 повторов различных методов для поиска параметров модели 1 без инбридинга демографической истории двух популяций пум

	BFGS		BOBYQA		GA+BFGS
	1 запуск	18 запусков	1 запуск	6 запусков	
Число вычислений правдоподобия	238 ± 53	$4\,307 \pm 200$	755 ± 743	$4\,667 \pm 1\,982$	$4\,103 \pm 1\,930$
Время CPU (мин.)	$0{,}8 \pm 3{,}6$	17 ± 18	$2{,}1 \pm 1{,}9$	13 ± 6	88 ± 43
Лучшее правдоподобие	$-452\,987{,}45$	$-452\,987{,}45$	**$-452\,969{,}40$**	**$-452\,969{,}40$**	$-452\,969{,}48$
Среднее правдоподобие	$-1\,418\,712$	$-455\,783$	$-1\,015\,406$	$-455\,684$	**$-453\,000$**
Ст. откл. правдоподобия	$2\,213\,096$	$22\,275$	$2\,474\,846$	$19\,105$	53

Таблица 10 – Результаты 100 повторов различных методов для поиска параметров модели 2 с инбридингом демографической истории двух популяций пум

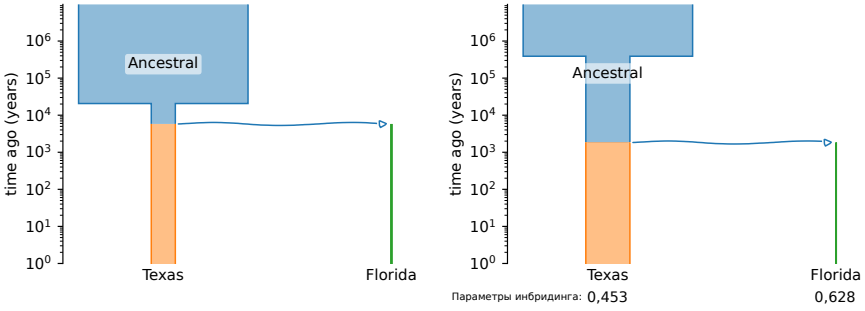
	BFGS		BOBYQA		GA+BFGS
	1 запуск	16 запусков	1 запуск	4 запуска	
Число вычислений правдоподобия	394 ± 82	$6\,245 \pm 324$	$1\,605 \pm 1\,207$	$6\,095 \pm 2\,561$	$6\,193 \pm 2\,680$
Время CPU (мин.)	$1{,}3 \pm 1{,}4$	25 ± 19	12 ± 5	16 ± 7	93 ± 47
Лучшее правдоподобие	$-317\,370{,}88$	$-317\,370{,}88$	**$-317\,239{,}48$**	**$-317\,239{,}48$**	$-317\,239{,}49$
Среднее правдоподобие	$-1\,729\,870$	$-320\,947$	$-381\,979$	$-320\,503$	**$-319\,451$**
Ст. откл. правдоподобия	$4\,339\,276$	$5\,029$	$115\,205$	$8\,753$	$7\,340$

и метод BFGS с перезапусками показали наихудшие результаты среди рассматриваемых методов.

Некоторые настроенные параметры численности популяций для моделей 1 и 2, полученные методами BOBYQA и GA+BFGS, оказались равными верхним граничным условиям. Для получения более достоверных результатов метод GA+BFGS был запущен для граничных условий с расширенными границами для параметров численности популяций. Границы для размеров популяций были выбраны равными $[10^{-4} \cdot N_A, 100 \cdot N_A]$, значения параметров времени были ограничены интервалом $[10^{-3} \cdot N_A, 10 \cdot N_A]$, а параметры инбридинга — интервалом $[10^{-3}, 1 - 10^{-3}]$, где N_A — размер предковой популяции. Границы для параметров времени и инбридинга являются значениями по умолчанию, которые были рекомендованы авторами *дадi* исходя из возможностей метода вычисления правдоподобия.

Финальные настроенные модели демографической истории двух популяций пум изображены на рисунке 79, их параметры представлены в работе [2] таблице S25. Для каждого параметра были получены доверительные интервалы

его значений с помощью численного метода информационной матрицы Годамбе [69]. Для этого метода был использован размер сетки $\epsilon = 10^{-2}$.



(а) (б)

Рисунок 79 – Полученные демографические истории двух популяций американской пумы для (а) модели 1 без инбридинга, (б) модели 2 с инбридингом

Полученные демографические истории имеют лучшие значения правдоподобия ($-452\,475{,}41$ против $-453\,003{,}05$ для модели 1 и $-316\,109{,}44$ против $-318\,058{,}08$ для модели 2), чем те, которые были получены в [51]. Они имеют похожие размеры популяций, за исключением размера флоридской популяции, которая была оценена в 800 особей по сравнению с $1\,200-1\,600$ особей, полученных в работе [51]. Время разделения популяций было оценено, как $4\,000 - 5\,500$ лет назад. Параметры инбридинга, полученные для модели 2, немного выше, чем для той же модели в [51]: $0{,}453$ для популяции Техаса и $0{,}628$ для популяции Флориды.

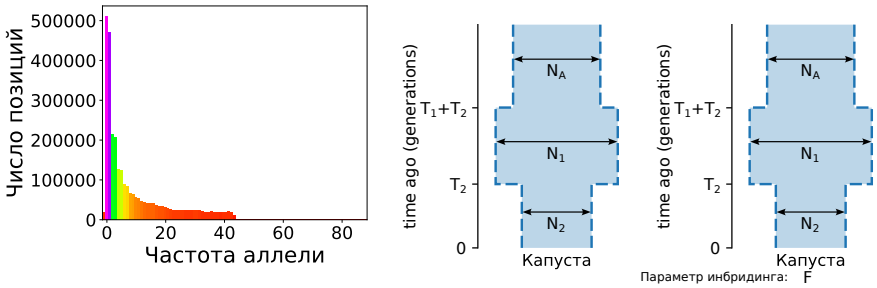
Рассматриваемая модель 1 вложена в модель 2, поэтому они были сравнены с использованием статистического критерия отношения правдоподобия. Данные имеют зависимости, поэтому был применен модифицированный критерий [69]. Статистика отношения правдоподобия составила $\lambda_{adj}^{LRT} = 2\,568{,}59$ (p-value = $\sim$0,0; [69]), что указывает на то, что настроенная модель 2 с инбридингом лучше описывает данные.

2.4.4. Сравнение с существующими методами настройки параметров моделей на данных одной популяции огородной капусты

Разработанный метод, основанный на генетическом алгоритме, был сравнен с методом BOBYQA на наборе данных одной популяции огородной капусты (*Brassica oleracea*). Генетические данные 45 особей были представлены в работах [147, 148], а их демографическая история была выведена в статье [51].

Вероятность мутации для огородной капусты составляет $\mu = 1.5 \times 10^{-8}$ для одной позиции в геноме за одно поколение. Средняя продолжительность одного поколения капусты равна одному году, что было использовано для перевода времени из единиц поколений в года. Длина генетической последовательности, представленной в данных, равна 411 560 319 пар оснований [147, 148].

Как и в случае данных американской пумы, для вывода демографической истории огородной капусты и сравнения методов настройки параметров были применены две модели. Обе модели описывают три интервала времени с различными постоянными размерами популяции. Модель 2 дополнительно содержит параметр инбридинга популяции, в модели 1 он отсутствует — равен нулю. На рисунке 80 изображены генетические данные в виде аллель-частотного спектра (рисунок 80а) и две описанные модели: на рисунке 80б показана модель 1 без инбридинга и на рисунке 80в — модель 2 с инбридингом.



(а) (б) (в)

Рисунок 80 – Генетические данные популяции огородной капусты (а) и рассматриваемые модели демографической истории (б, в)

Ограничения на значения параметров были выбраны как в работе [51] для всех рассматриваемых методов. Размеры популяции лежат в интервале $[10^{-3} \cdot N_A, 50 \cdot N_A]$, время — в интервале $[2 \cdot 10^{-3} \cdot N_A, 100 \cdot N_A]$, где N_A — размер предковой популяции (рисунок 80в). Значения параметров инбридинга обязаны принадлежать промежутку $[10^{-5}, 1 - 10^{-5}]$. Был использован метод вычисления правдоподобия, реализованный в ∂a∂i, с сеткой $pts = \{100, 110, 120\}$.

Для данных огородной капусты были сравнены три метода: 1) метод BOBYQA, 2) метод BOBYQA с перезапусками и 3) разработанный метод GA+BFGS, основанный на комбинации генетического алгоритма и метода локального поиска — метода BFGS. Число перезапусков второго метода было выбрано так, чтобы среднее число вычислений значения правдоподобия совпадало со средним числом вычислений метода GA+BFGS. Оно составило 27 перезапусков для модели 1 и 12 перезапусков для модели 2. Каждый метод был повторен 100 раз, максимально достигнутое значение, среднее и стандартное отклонения правдоподобия были сравнены.

Результаты представлены в таблицах 11 и 12 для моделей 1 и 2 соответственно. Жирным шрифтом выделены наилучшие результаты.

Таблица 11 – Результаты 100 повторов различных методов для поиска параметров модели 1 без инбридинга демографической истории одной популяции огородной капусты

	BOBYQA		GA+BFGS
	1 запуск	27 запусков	
Число вычислений правдоподобия	205 ± 307	$6\,053 \pm 1,733$	$6\,034 \pm 2\,764$
Время CPU (мин.)	6 ± 19	160 ± 131	84 ± 40
Лучшее правдоподобие	$-24\,303\,37$	$\mathbf{-24\,292\,72}$	$-24\,307\,64$
Среднее правдоподобие	$-55\,567$	$\mathbf{-25\,384}$	$-25\,723$
Ст. откл. правдоподобия	19 498	4 371	2,689

Таблица 12 – Результаты 100 повторов различных методов для поиска параметров модели 2 с инбридингом демографической истории одной популяции огородной капусты

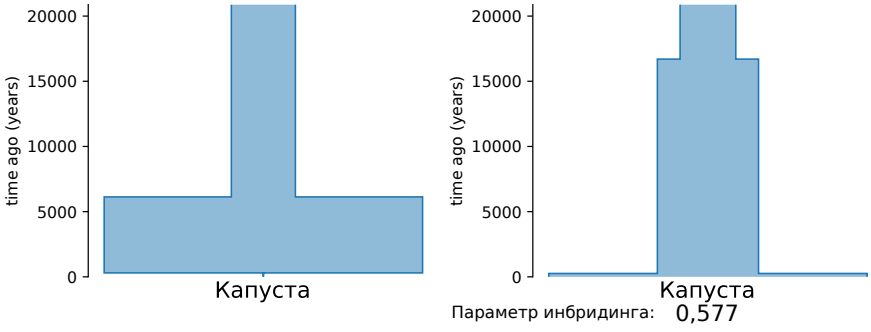
	BOBYQA		GA+BFGS
	1 запуск	12 запусков	
Число вычислений правдоподобия	534 ± 773	$5\,923 \pm 2\,501$	$5\,872 \pm 3\,282$
Время CPU (мин.)	33 ± 46	357 ± 173	418 ± 232
Лучшее правдоподобие	$-4\,271\,25$	$\mathbf{-4\,270\,35}$	$-4\,270\,37$
Среднее правдоподобие	$-26\,023$	$\mathbf{-4\,398}$	$-4\,677$
Ст. откл. правдоподобия	29 721	285	778

Было продемонстрировано, что метод GA+BFGS превосходит метод BOBYQA с единичным запуском. Однако при многократном запуске метода BOBYQA, он оказывается более эффективным, чем метод GA+BFGS. Отметим, что число запусков, необходимых для достижения эффективности метода BOBYQA, варьируется для разных данных и остается неизвестным в общем случае. Метод GA+BFGS не обладает этим недостатком и предоставляет сравнимый результат.

Как и в случае вывода демографической истории американской пумы, некоторые настроенные параметры для моделей огородной капусты были равны верхним граничным условиям. Поэтому были выведены новые значения

параметров при условии расширенных ограничений для параметров численности популяций. Область определения для размеров популяции была выбрана равной промежутку $[10^{-2} \cdot N_A, 10^4 \cdot N_A]$, для параметров времени — $[2 \cdot 10^{-5} \cdot N_A, 10 \cdot N_A]$, где N_A — размер общей предковой популяции. Границы для параметров инбридинга были выбраны равными $[10^{-3}, 1 - 10^{-3}]$. Настройка параметров была выполнена с помощью разработанного метода GA+BFGS.

Финальные демографические истории представлены на рисунке 81, их параметры и доверительные интервалы приведены в работе [2] таблице S30. Доверительные интервалы были оценены с помощью метода информационной матрицы Годамбе, предложенного в работе [69], с размером сетки $\epsilon = 10^{-2}$.



(а) (б)

Рисунок 81 – Полученные демографические истории одной популяции огородной капусты для (а) модели 1 без инбридинга, (б) модели 2 с инбридингом

Полученные демографические истории одомашненной капусты имеют лучшие значения логарифма правдоподобия, чем было получено ранее в работе [51]: $-24\,137{,}13$ против $-24\,330{,}40$ для модели 1 и $-4\,267{,}14$ против $-4\,281{,}14$ для модели 2. Значения для размеров популяции в первую и вторую эпохи выводятся аналогично результатам [51]. Однако величина численности популяции для самого недавнего временного интервала была предсказана меньше, чем было получено ранее для модели 1 без инбридинга (шесть против 592 особей) и больше, чем получено ранее для модели 2 с инбридингом (174 960 000 против 215 000 особей). Продолжительность этого временного интервала оказалась меньше для обеих моделей, чем предполагалось ранее. Отметим, что в случае модели 1 значение этого параметра времени очень близко к нулю.

Рассматриваемые модели были сравнены с использованием теста отношения правдоподобия. Используемые генетические данные имели зависимости, поэтому была использована модифицированная версия теста [69]. Статистика составила $\lambda_{adj}^{LRT} = 127{,}10$ (p-value $= \sim 0{,}0$), что означает, что модель 2 с инбридингом лучше описывает данные, чем модель 1 без инбридинга.

2.4.5. Сравнение методов вычисления правдоподобия на симулированных данных двух популяций орангутанга

Четыре метода вычисления правдоподобия, реализованные в программных средствах *дадi*, *moments*, *momi2* и *momentsLD*, были сравнены с использованием разработанного метода GA+NM, основанного на комбинации генетического алгоритма и локального поиска — метода Нелдера-Мида. Сравнение проводилось на данных, симулированных с использованием библиотеки *stdpopsim* [6, 7]. Эта библиотека имеет каталог из существующих видов, для которых записаны различные характеристики такие, как размер хромосом, генома, скорость мутации и рекомбинации, которые были проанализированы и вычислены в опубликованных ранее исследованиях. Для каждого вида имеется набор ранее полученных демографических историй. Для симуляции данных требуется выбрать вид, его характеристики, демографическую историю и размер данных. Симуляция производится с использованием одного из двух движков: *msprime* [149] или SLiM [150].

Методы вычисления правдоподобия для вывода демографической истории популяций были сравнены на наборе данных, симулярованных для двух популяций орангутангов. Для моделирования был использован сценарий, доступный в библиотеке *stdpopsim* [6]. Набор данных был симулирован с использованием движка *msprime* [149] и включает генетические данные пяти диплоидных особей на каждую популяцию.

Демографическая история борнейского (*Pongo pygmaeus*) и суматранского (*Pongo abelii*) орангутанов была первоначально предложена в работе [151] и показана на рисунке 82. Демографическая история изображена серым цветом, чтобы подчеркнуть, что она была использована для симуляции данных. В частности, эта история является историей изоляции популяций с миграцией, которая включает разделение предковой популяции, последующие экспоненциальные рост и спад численности суматранской и борнейской популяций соответственно. Было симулировано 23 неполовые хромосомы общей длиной $2{,}87 \cdot 10^9$ пар оснований. Скорость мутаций равна $1{,}5 \cdot 10^{-8}$ на позицию на поколение [152]. Усредненные скорости рекомбинации для каждой хромосомы взяты из карты рекомбинации *Pongo abelii*, полученной в [152].

Четыре метода вычисления правдоподобия, реализованных в *дадi*, *moments*, *momi2* и *momentsLD*, были сравнены на полученном наборе данных с использованием семи моделей демографической истории, изображенных на рисунке 83. Первая модель ORAN-NOMIG включает разделение предковой популяции и экспоненциальные изменения численности суматранских и борнейских орангутанов, но не имеет миграций. Модель ORAN-MIG соответствует модели ORAN-NOMIG, но включает асимметричные непрерывные миграции между популяциями после разделения. Две расширенные модели ORAN-STRUCT-NOMIG и ORAN-STRUCT-MIG были также рассмотрены. Они отличаются отсутствием и наличием миграций и имеют по два дискретных параметра динамики численности популяций в каждой. Дополнительно в анализ

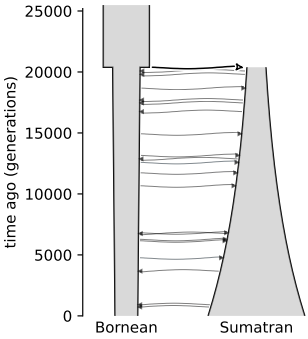


Рисунок 82 – Демографическая история двух популяций орангутангов, использованная для симуляции генетических данных с помощью *stdpopsim* [6]

были включены три модели с единичными миграциями. Они были сравнены только для метода вычисления правдоподобия, реализованного в *momi2*, для того чтобы проверить может ли непрерывная миграция, которая не поддерживается *momi2*, быть заменена на несколько единичных. Модель ORAN-NOMIG без миграций была сравнена с тремя моделями: 1) с одной единичной миграцией (ORAN-PULSE1), 2) с тремя единичными миграциями (ORAN-PULSE3) и 3) с семью единичными миграциями (ORAN-PULSE7). Среднее время одного вычисления правдоподобия и среднее число вычислений в проведенных экспериментах представлены в таблицах 13 и 14 соответственно.

Таблица 13 – Среднее время одного вычисления правдоподобия при использовании метода GA+NM и различных методов вычисления правдоподобия

Model	*∂a∂i*	*moments*	*momi2*	*momentsLD*
ORAN-NOMIG	0.05 ± 0.14	0.03 ± 0.01	0.01 ± 0.006	15.02 ± 25.71
ORAN-MIG	0.24 ± 2.02	0.16 ± 0.11	–	27.98 ± 44.59
ORAN-STRUCT-NOMIG	0.07 ± 0.34	0.06 ± 0.02	0.02 ± 0.01	15.73 ± 27.55
ORAN-STRUCT-MIG	0.17 ± 0.75	0.27 ± 0.16	–	18.09 ± 39.26
ORAN-PULSE1	–	–	0.39 ± 0.16	–
ORAN-PULSE3	–	–	1.18 ± 0.51	–
ORAN-PULSE7	–	–	3.13 ± 1.34	–

Заметим, что модели ORAN-MIG и ORAN-STRUCT-MIG являются корректно заданными — они способны отобразить истинную демографическую историю, которая была использована для симулирования данных. Модели ORAN-NOMIG и ORAN-STRUCT-NOMIG не включают непрерывные миграции, которые присутствуют в истинной истории, а модели ORAN-PULSE-1, ORAN-

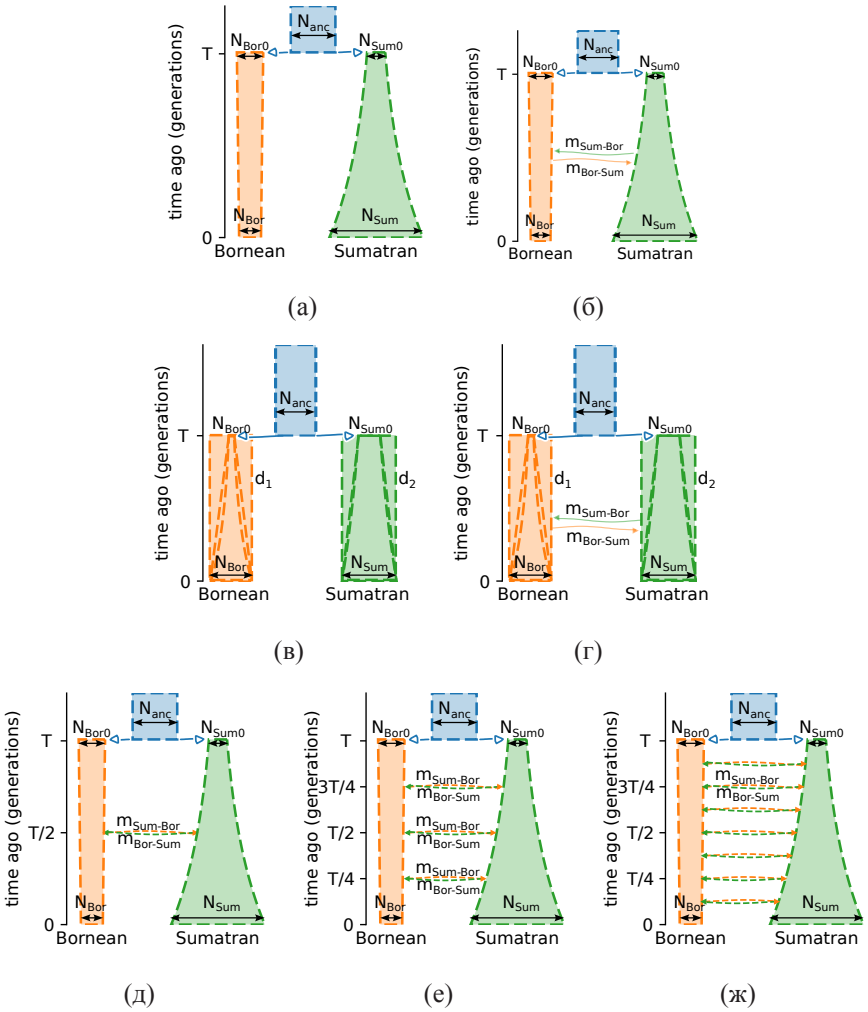


Рисунок 83 – Модели демографической истории двух популяций орангутангов: (а) ORAN-NOMIG, (б) ORAN-MIG, (в) ORAN-STRUCT-NOMIG, (г) ORAN-STRUCT-MIG, (д) ORAN-PULSE-1, (е) ORAN-PULSE-3, (ж) ORAN-PULSE-7

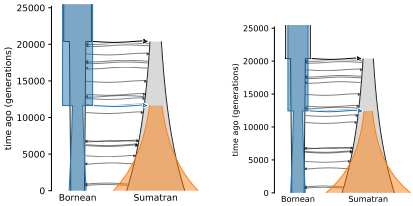
PULSE-3 и ORAN-PULSE-7 заменяют непрерывные миграции на набор единичных.

Полученные значения параметров, настроенные с применением разработанного метода GA+NM на основе комбинации генетического алгоритма и ло-

Таблица 14 – Среднее число вычислений правдоподобия, которое потребовалось для настройки параметров рассматриваемых моделей с использованием метода GA+NM и различных методов вычисления правдоподобия

Model	*∂a∂i*	*moments*	*momi2*	*momentsLD*
ORAN-NOMIG	11,442	8,428	6,993	10,318
ORAN-MIG	11,411	12,320	–	13,570
ORAN-STRUCT-NOMIG	3,977	3,881	7,020	2,570
ORAN-STRUCT-MIG	5,356	7,200	–	10,318
ORAN-PULSE1	–	–	7,667	–
ORAN-PULSE3	–	–	9,160	–
ORAN-PULSE7	–	–	9,178	–

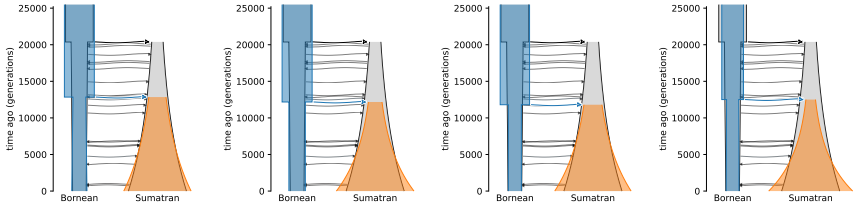
кального поиска для всех рассмотренных моделей приведены в [2] в таблицах S14, S15, S16, 3 и 4. Для моделей ORAN-MIG и ORAN-STRUCT-MIG, которые являются корректно заданными, для всех рассмотренных методов вычисления правдоподобия полученные параметры близки к истинной истории. Для расширенной модели ORAN-STRUCT-MIG все методы предоставили корректные экспоненциальные динамики изменения численности борнейской и суматранской популяций. На рисунках 84, 85 и 86 продемонстрированы результаты для некорректно заданных моделей, которые представляют наибольший интерес. Эти изображения показывают полученные демографические истории для сравнения с наложенными на истинную историю, которая была использована для симуляции данных и изображена серым цветом.



(а) (б)

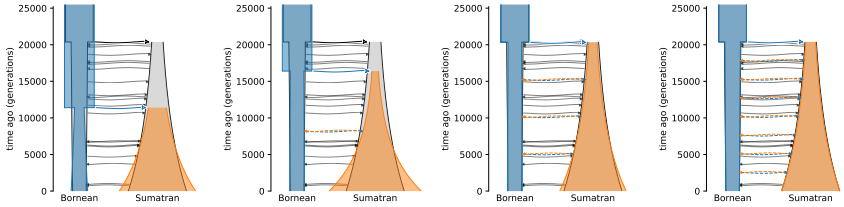
Рисунок 84 – Результаты настроенной модели ORAN-NOMIG для методов вычисления правдоподобия, основанных на (а) аллель-частотном спектре, (б) статистиках неравномерного сцепления генов

В случае модели ORAN-NOMIG без миграций все четыре метода вычисления правдоподобия дают схожие параметры. Полученные значения идентичны для *∂a∂i*, *moments* и *momi2*, которые являются методами вычисления правдоподобия на основе аллель-частотного спектра. Согласно результатам, полученным



(а) (б) (в) (г)

Рисунок 85 – Результаты настроенной модели ORAN-STRUCT-NOMIG для методов вычисления правдоподобия, реализованных в (а) *дадi*, (б) *moments*, (в) *momi2*, (г) *momentsLD*



(а) (б) (в) (г)

Рисунок 86 – Результаты настроенных моделей для метода вычисления правдоподобия, реализованного в *momi2*: (а) модель ORAN-NOMIG, (б) модель ORAN-PULSE-1, (в) модель ORAN-PULSE-3, (г) модель ORAN-PULSE-7

для этих методов, численность предковой популяции составляет $\sim$19 000, для метода *momentsLD*, основанного на статистиках неравномерного сцепления генов, эта численность равна $\sim$14 000. Истинное значение размера предковой популяции равно $\sim$17 934. Демографические истории, полученные настройкой параметров модели ORAN-NOMIG, представлены на рисунке 84. На рисунке 84а представлен общий результат для *дадi*, *moments* и *momi2*, на рисунке 84б — для *momentsLD*. Оценки для современных размеров популяций больше, чем истинные значения, использованные при симуляции. Кроме того, время разделения в полученных результатах меньше: $\sim$12 000 поколений против истинного времени $\sim$20 000 поколений. Эти расхождения между полученными значениями параметров для модели ORAN-NOMIG и истинными значениями можно объяснить тем, что модель чрезмерно упрощена и в ней отсутствует миграция.

Настроенные значения параметров расширенной модели ORAN-STRUCT-NOMIG близки к полученным значениям параметров для модели ORAN-NOMIG и представлены на рисунке 85. Динамики изменения численности популяций правильно определены как экспоненциальные в случае методов *дадi* и

momentsLD (рисунки 85а и 85г). Однако методы вычисления правдоподобия *momi2* и *moments* отдали предпочтение постоянной численности борнейской популяции. Таким образом, некорректная спецификация модели может приводить к ошибочным результатам в настройке параметров динамики изменения численности популяций. Несмотря на эти различия, отметим, что постоянный размер достаточно хорошо аппроксимирует истинную историю борнейской популяции ((рисунки 85в и 85б)).

Наконец, были проанализированы результаты, полученные с использованием метода вычисления правдоподобия, реализованного в *momi2*, для дополнительных моделей ORAN-PULSE-1, ORAN-PULSE-3 и ORAN-PULSE-7 с единичными миграциями. Полученные демографические истории представлены на рисунке 86. Результаты для модели ORAN-NOMIG (рисунок 86а) были сравнены с результатами для трех дополнительных моделей с единичными миграциями (рисунки 86б, 86в и 86г). Полученные интенсивности миграции для единичных миграций некорректно сравнивать с интенсивностями непрерывных миграций, однако отметим, что их значения уменьшаются с увеличением числа единичных миграций в модели. Например, интенсивность единичной миграции из популяции борнейских орангутанов в популяцию суматранских ($m_{Bor-Sum}$) равна $0{,}65$ для модели ORAN-PULSE-1 с одной единичной миграцией, $0{,}057$ для модели ORAN-PULSE-3 с тремя импульсами и $0{,}025$ для модели ORAN-PULSE-7 с семью импульсами. Что является более важным выводом, так это то, что значения других параметров сходятся к истинным. Так, время разделения предковой популяции оценивается в $\sim$11 000 поколений для модели ORAN-NOMIG, $\sim$16 000 поколений для модели ORAN-PULSE-1 и $\sim$20 000 для моделей ORAN-PULSE-3 и ORAN-PULSE-7. Последнее значение близко к истинному значению 20 157, использованному при симулировании генетических данных. Оценки параметров для модели ORAN-PULSE-7 с семью миграциями импульсов являются наиболее точными среди рассмотренных моделей. Увеличение числа импульсных событий приводит к более точным оценкам, но требует больше вычислительных ресурсов, согласно таблицам 13 и 14. Таким образом, непрерывная миграция, которая не поддерживается методом *momi2*, в некоторой степени может быть заменена несколькими единичными миграциями.

2.4.6. Вывод демографической истории трех популяций современного человека

Были проанализированы генетические данные популяций современного человека на территории Российской федерации. Полногеномные данные были получены для 60 особей из трех популяций (рисунок 87):

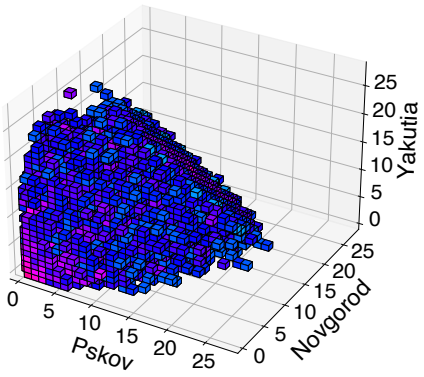
- жители территории Пскова (Pskov);
- жители территории Новгорода (Novgorod);
- жители территории Якутии (Yakutia).

Данные были собраны для семей — существовали родственные связи между некоторыми особями. Для анализа демографической истории было отобрано по 14 представителей, не связанных родственными связями. По этим данным был построен аллель-частотный спектр размера $29 \times 29 \times 29$. Построенный аллель-частотный спектр представлен на рисунке 88. На рисунке 88а приведено изображение спектра в трехмерном пространстве и на рисунках 88б, 88в и 88г изображены проекции этого спектра в двухмерном пространстве.

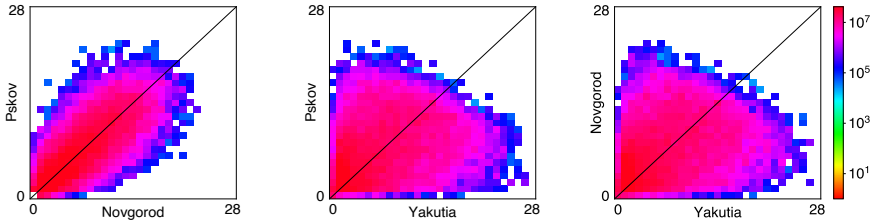
Рисунок 87 – Географическое расположение образцов генетических данных. Источник: [4]

Для вывода демографической истории рассматриваемых популяций была использована модель расширенного класса, представленная на рисунке 89. Модель включает разделение предковой популяции постоянного размера на якутскую популяцию (Yakutia) и общую популяцию жителей Пскова и Новгорода, которая также разделилась и образовала популяции Novgorod и Pskov. Эта модель содержит три элемента временных интервалов, разделенные двумя элементами разделения. Все динамики изменения численности и размеры популяций определены независимыми параметрами и доступны для настройки.

Для настройки параметров модели был применен разработанный метод на основе комбинации генетического алгоритма и метода Пауэлла. В качестве метода вычисления правдоподобия был использован метод, реализованный в *moments*. Настройка параметров была запущена 10 раз и лучший результат получил значение правдоподобия равное $f^{moments}(\theta^\star) = -12\,751$ и представлен на рисунке 90. Размер предковой популяции (параметр N_A) составил 2 250 особей. Популяции Пскова и Новгорода имеют практически идентичные истории и были



(а)



(б) (в) (г)

Рисунок 88 – Генетические данные трех популяций современного человека на территории Российской федерации в виде аллель-частотного спектра

образованы разделением общей популяции всего 1 980 лет назад. Якутская популяция была образована гораздо раньше — 10 380 лет назад, и имела сначала постоянный размер около 900 особей и затем линейное увеличение численности до 7 800 особей. Общая популяция жителей Пскова и Новгорода также имела линейный рост с 380 до 3 700 особей, а популяции Пскова и Новгорода — постоянную численность в 7 000 и 70 000 соответственно.

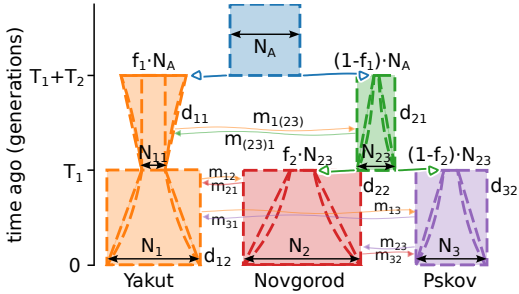


Рисунок 89 – Рассматриваемая модель расширенного класса

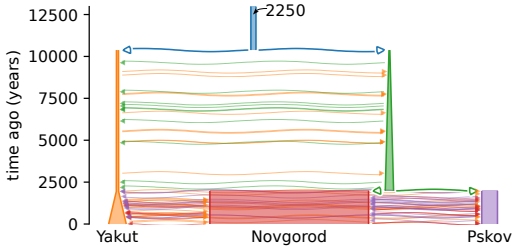


Рисунок 90 – Полученная демографическая история трех популяций современного человека

2.5. Экспериментальные исследования разработанного метода настройки параметров моделей, основанного на комбинации байесовской оптимизации и локального поиска для данных четырех и пяти популяций

Для выявления эффективности разработанного метода, основанного на комбинации байесовской оптимизации и локального поиска, были проведены экспериментальные исследования на симулированных и реальных данных. Сначала комбинированный метод на основе байесовской оптимизации был сравнен с комбинированным методом на основе генетического алгоритма. На множестве наборов данных было продемонстрировано превосходство байесовской оптимизации в случае вывода демографической истории более, чем трех популяций. Затем, разработанный метод был сравнен с существующим методом настройки параметров моделей, основанным на методе локального поиска, на генетических данных четырех и пяти популяций современных людей. Были получены демографические истории с большим значением правдоподобия, чем было получено ранее в других исследованиях для этих данных.

2.5.1. Сравнение с разработанным генетическим алгоритмом на симулированных и реальных данных

В данной работе было предложено два комбинированных метода настройки параметров моделей демографической истории на основе комбинации методов глобального поиска и локального поиска. Первый метод использует генетический алгоритм в качестве метода глобальной оптимизации, второй — ансамблевую байесовскую оптимизацию. Экспериментальные исследования, описанные в разделе 2.4, демонстрируют эффективность генетического алгоритма при сравнении с существующими методами настройки, основанными на локальной оптимизации. Разработанные методы на основе генетического алгоритма и байесовской оптимизации были сравнены на множестве наборов симулированных и реальных данных. Сравнение было продемонстрировано с применением метода вычисления правдоподобия, реализованного в *moments*.

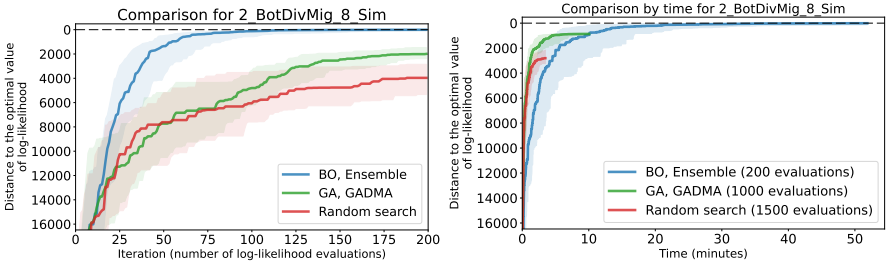
Основное применение байесовской оптимизации — решение задач оптимизации сложновычислимых функций. Она имеет накладные временные расходы, возникающие из-за регрессии на основе гауссовского процесса и оптимизации функции выбора на каждой итерации. Генетический алгоритм и байесовская оптимизация были сравнены с использованием графиков сходимости не только по итерациям, но и по времени.

Были рассмотрены 13 наборов данных с различным числом популяций: от одной популяции до пяти. Рисунок 58 показывает среднее время вычисления правдоподобия с использованием *moments* для каждого из рассматриваемых наборов данных. Все используемые наборы данных являются частью пакета `deminf_data v1.0.0`. Каждый набор включает генетические данные, модель демографической истории и границы значений параметров. Более подробное описание пакета представлено в разделе 2.2.3.

Для каждого набора данных были запущены три метода настройки параметров модели: 1) ансамблевая байесовская оптимизация (BO, Ensemble), 2) генетический алгоритм (GA, GADMA), 3) случайный поиск (Random search). Случайный поиск был применен для дополнительной проверки эффективности первых двух методов. Были построены графики сходимости по итерациям и по времени для тринадцати наборов данных. В данном контексте под итерацией понимается одно вычисление целевой функции. Все полученные графики сходимости представлены в [3] на рисунках S10–S14. Некоторые примеры графиков сходимости представлены на рисунке 91 для двух популяций, на рисунке 92 для трех популяций, на рисунке 93 для четырех популяций и на рисунке 94 для пяти популяций. Сплошные линии отображают медиану сходимости метода, закрашенная область является областью между квартилями. Каждый метод был повторен 64 раза.

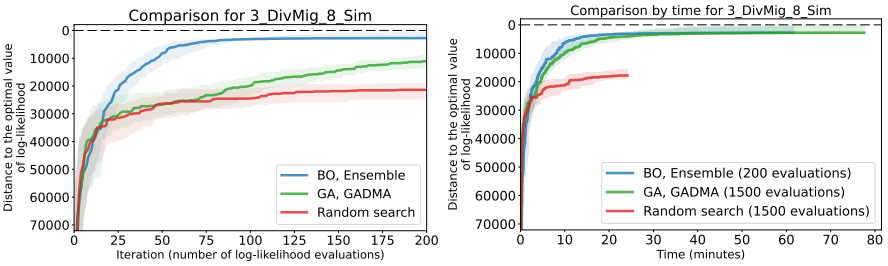
Байесовская оптимизация и генетический алгоритм превосходят случайный поиск на всех наборах данных. Ансамблевая байесовская оптимизация превосходит генетический алгоритм на всех наборах данных согласно полученным

графикам сходимости по итерациям. Однако как упоминалось ранее, байесовская оптимизация имеет накладные временные расходы, что приводит к другим результатам при сравнении сходимости методов по времени. Для всех рассмотренных наборов данных одной и двух популяций генетический алгоритм имел более быструю сходимость в терминах времени, чем байесовская оптимизация. В случае трех популяций, оба метода имели схожую сходимость по времени. В случае четырех и пяти популяций, байесовская оптимизации оказывалась эффективнее, чем генетический алгоритм. В случае пяти популяций баейсовская оптимизация позволяет сократить время настройки параметров на дни и даже недели (рисунок 94б).



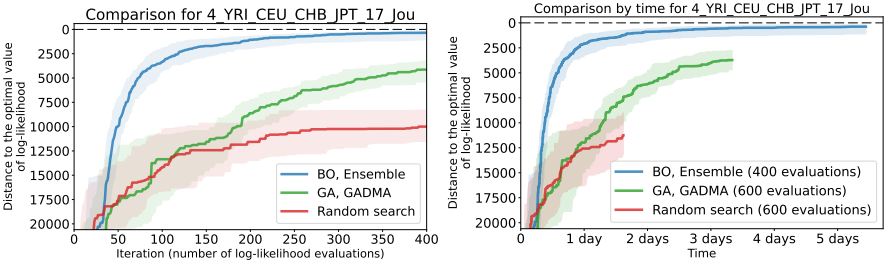
(а) (б)

Рисунок 91 – Сходимость рассматриваемых методов настройки параметров модели демографической истории двух популяций по (а) итерациям, (б) времени



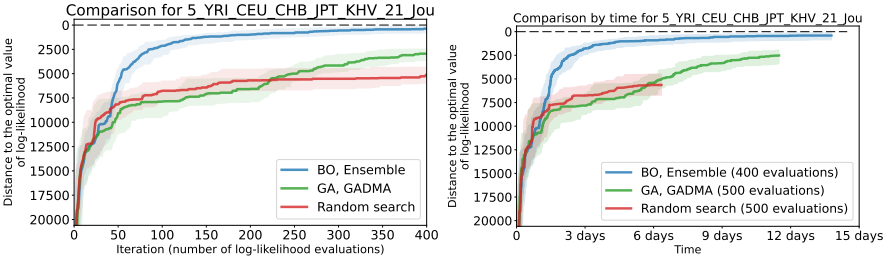
(а) (б)

Рисунок 92 – Сходимость рассматриваемых методов настройки параметров модели демографической истории трех популяций по (а) итерациям, (б) времени



(а) (б)

Рисунок 93 – Сходимость рассматриваемых методов настройки параметров модели демографической истории четырех популяций по (а) итерациям, (б) времени



(а) (б)

Рисунок 94 – Сходимость рассматриваемых методов настройки параметров модели демографической истории пяти популяций по (а) итерациям, (б) времени

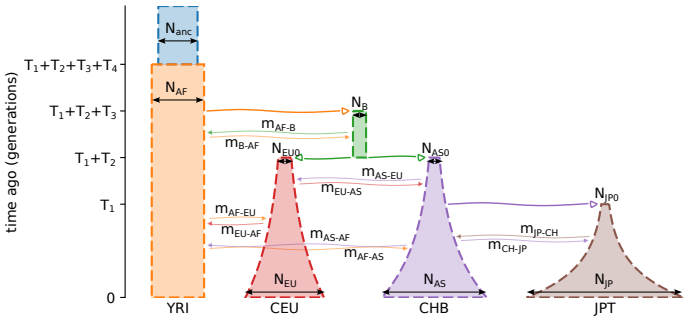
2.5.2. Сравнение с существующим методом настройки параметров моделей на реальных данных четырех и пяти популяций современного человека

Разработанный метод на основе комбинации ансамблевой байесовской оптимизации и локального поиска был сравнен с существующим методом Пауэлла с множественными запусками на реальных данных современного человека. Были использованы два набора генетических данных для четырех и пяти популяций современного человека, построенным в работе [55] по данным аутосомных синонимичных последовательностей из общедоступного проекта 1000 Genomes Project [153, 154]. Данные по четырем популяциям включают: 1) популяцию Йоруба из Ибадана, Нигерия (YRI); 2) европейскую популяцию, представленную жителями штата Юта с североевропейскими и западноевропейскими корнями (CEU); 3) популяцию Ханьцы из Пекина, Китай (CHB); и 4) японскую популяцию из Токио (JPT). Данные по пяти популяциям включают те же четыре популяции и дополнительную пятую популяцию киньских вьетнамцев (KHV). Отметим, что эти два набора данных включены в пакет `deminf_data` и имеют названия — `4_YRI_CEU_CHB_JPT_17_Jou` для четырех популяций и `5_YRI_CEU_CHB_JPT_KHV_21_Jou` для пяти популяций.

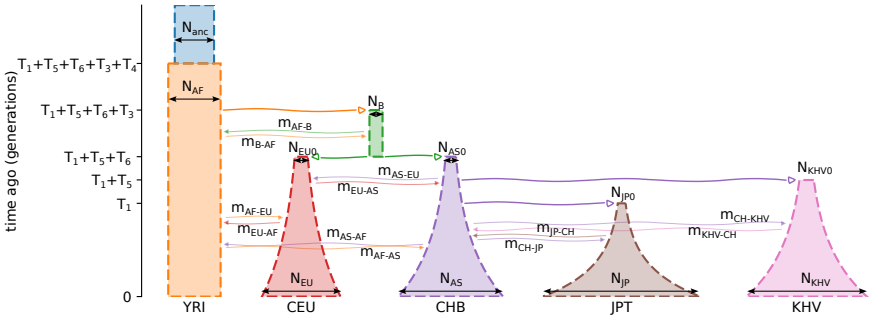
В работе [55] генетические данные четырех и пяти популяций были проанализированы и получены их демографические истории. Настройка параметров моделей была выполнена с применением метода Пауэлла с перезапусками. Были проведены экспериментальные исследования по настройке параметров моделей, использованных в [55], с применением разработанного метода на основе комбинации байесовской оптимизации и локального поиска. Разработанный метод предоставил значения параметров, имеющих лучшее значение правдоподобия, чем те, что были получены ранее в [55].

Использованные модели демографических историй представлены на рисунке 95. Модель 1 для четырех популяций имеет 17 параметров (рисунок 95а), а модель 2 для пяти популяций включает 21 параметр — те же 17 параметров, что в модели 1, и четыре дополнительных (рисунок 95б). Чтобы сократить время вычислений, авторы работы [55] оптимизировали 17 параметров модели 1, затем фиксировали их и настраивали дополнительные четыре параметра второй модели. Обозначим демографические истории, полученные в [55], как *базовые истории*.

Сначала была проведена настройка 17 параметров модели 1 для четырех популяций современного человека. Комбинированный метод включал ансамблевую байесовскую оптимизацию на 400 вычислений целевой функции и последующий метод BFGS локального поиска без ограничений на число вычислений. Настроенные параметры можно найти в работе [3] в таблице S3. На рисунке 96 изображены базовая история и две лучшие демографические истории, полученные в результате 64 повторов разработанного метода. Обе полученные демографические истории имеют значение правдоподобия лучше, чем базовая история (рисунок 96а), полученная в [55]. Лучшая история похожа на базовую, но имеет экспоненциальное снижение популяции JPT с 30 000 до 15 000 особей в отличие



(а)



(б)

Рисунок 95 – Модели демографической истории современного человека
(а) четыре популяции, (б) пять популяций

от экспоненциального роста с 4 000 до 230 000 особей в базовой истории (рисунок 96б). Более того, эта история предполагает гораздо более низкую интенсивность непрерывной миграции между популяциями CHB и JPT. Вторая лучшая история гораздо более похожа на базовую, но имеет более низкие темпы роста популяции JPT и более низкий уровень миграции между популяциями CHB и JPT (рисунок 96в).

После этого 17 параметров, которые являются общими для модели 1 и модели 2, были фиксированы в модели 2 для пяти популяций, а остальные четыре параметра были настроены. При этом использовались те же значения зафиксированных параметров, что и в работе [55]. Запуск комбинированного метода BO Ensemble включал ансамблевую байесовскую оптимизацию, запущенную на 200 вычислений целевой функции и последующим запуском метода BFGS. Потребовалось всего 16 ± 7 вычислений целевой функции, чтобы превысить значение правдоподобия, полученное для базовой истории в [55]. Настроенные парамет-

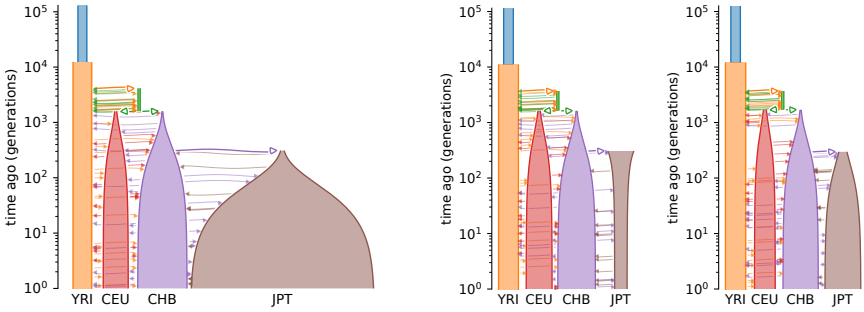


Рисунок 96 – Демографические истории для модели 1 четырех популяций современного человека: (а) базовая история, полученная с помощью метода Пауэлла с перезапусками в работе [55], (б) лучшая история, полученная с помощью разработанного метода BO Ensemble, (в) альтернативная история, полученная с помощью разработанного метода BO Ensemble

ры представлены в работе [3] в таблице S4. На рисунке 97 представлены базовая история и все полученные истории для модели 2. На рисунке 97б показана полученная демографическая история для модели 2 с фиксированными параметрами. Эта история имеет интенсивность миграции между популяциями CHB и KHV в два раза больше, чем в базовой истории. Более того, раскол популяции CHB, который образовал популяцию KHV, в полученной истории произошел раньше: 590 поколений назад, по сравнению с 337 поколениями в базовой истории.

Наконец, разработанный метод был использован для настройки всех параметров модели 2 для пяти популяций. Байесовская оптимизация в комбинированном методе была запущена на 400 вычислений целевой функции, а последующий локальный поиск был запущен без ограничений. Запуск комбинированного метода был повторен 64 раза. Значения настроенных параметров могут быть найдены в работе [3] в таблице S4. Две лучшие альтернативные истории представлены на рисунке 97, как 97в и 97г. Обе истории имеют значения правдоподобия выше, чем базовая история и чем история, полученная при фиксированных 17 параметрах модели 2 (рисунок 97б). История с лучшим значением правдоподобия (рисунок 97в) имеет экспоненциальное сокращение японской популяции JPT, аналогично наилучшей истории для модели 1 четырех популяций. Однако этот результат не подтверждается другими исследованиями [56]. Событие «выхода людей из Африки» в этой истории произошло более миллиона лет назад, что не подтверждается современными археологическими исследованиями. Вторая лучшая история (рисунок 97г) лучше согласуется с современными знаниями [56]. Различия в значениях параметрах между этой историей и базовой историей касаются в основном популяций YRI и CEU. Это можно объяснить малым числом особей этих двух популяций в использованном наборе генетических данных. Аллель-

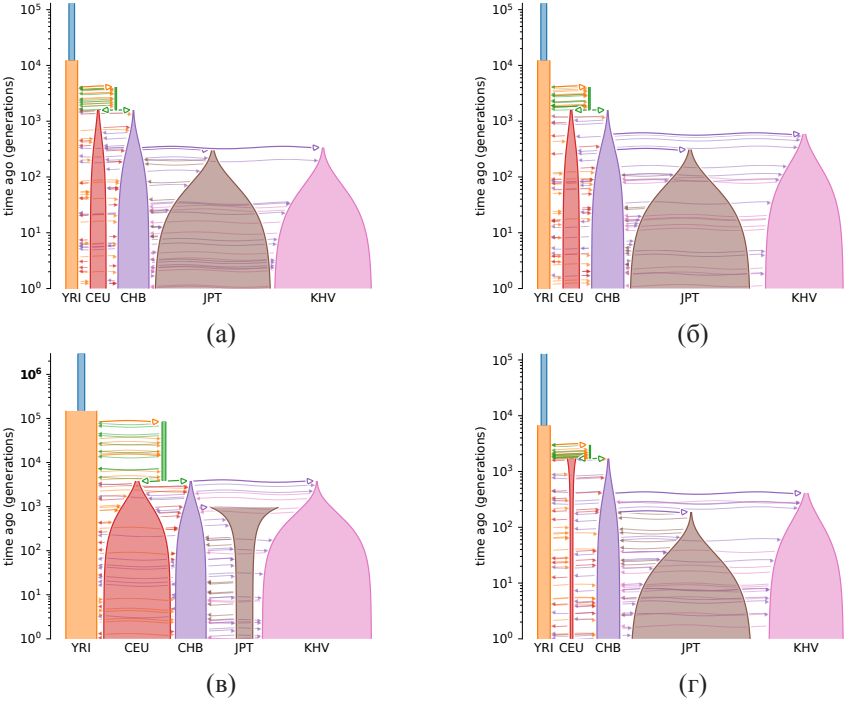


Рисунок 97 – Демографические истории для модели 2 пяти популяций современного человека: (а) базовая история, полученная с помощью метода Пауэлла с перезапусками в работе [55], (б) история, полученная настройкой четырех параметров с помощью разработанного метода BO Ensemble, (в) лучшая история, полученная настройкой 21 параметра с помощью разработанного метода BO Ensemble, (г) альтернативная история, полученная настройкой 21 параметра с помощью разработанного метода BO Ensemble

частотный спектр был построен всего для пяти хромосом для популяций YRI и CEU, в то время как другие популяции были представлены 30 хромосомами.

Выводы по граве 2

1. Разработан расширенный класс моделей демографических историй, который включает модели с дискретными параметрами динамики изменения численности популяции для настройки.
2. Разработан метод настройки параметров моделей на основе комбинации генетического алгоритма и локального поиска. Гиперпараметры разработанного метода были настроены автоматически для более эффективного решения поставленной задачи.
3. Разработан метод настройки параметров моделей на основе комбинации байесовской оптимизации и локального поиска. Гиперпараметры разработанного метода были настроены с использованием экспериментальных исследований на множестве наборов данных.
4. Разработанные методы настройки параметров могут быть использованы как для настройки существующих моделей первого и второго класса, так и моделей разработанного расширенного класса.
5. На основе экспериментальных исследований на симулированных и реальных данных было продемонстрировано, что разработанные комбинированные методы настройки параметров являются более эффективными, чем существующие методы, основанные на методах локальной оптимизации. Они предоставляют параметры, имеющие лучшее значение правдоподобия, чем методы локальной оптимизации.
6. Проведены экспериментальные исследования сравнения разработанного генетического алгоритма и байесовской оптимизации. Генетический алгоритм имеет более быструю сходимость при настройке параметров моделей одной, двух и трех популяций. Байесовская оптимизация оказывается более эффективной, чем генетический алгоритм, в случае более трех популяций. В случае четырех и пяти популяций баейсовская оптимизация позволяет найти решение, близкое к оптимуму, на 50-80% быстрее, чем генетический алгоритм.
7. Для реальных данных получены и проанализированы демографические истории, имеющие лучшее значение правдоподобия или информационного критерия Акаике, чем полученные ранее по тем же данным в других работах.

Глава 3. Метод автоматического перебора расширенных моделей с разным числом параметров и настройки параметров по генетическим данным одной, двух и трех популяций

Классические методы вывода демографической истории популяций предполагают использование параметризованной модели демографической истории. Для более надежного результата требуется вручную перебирать множество моделей и выбирать ту, которая наилучшим образом описывает генетические данные. В данной работе был разработан класс расширенных моделей, которые включают параметры динамики изменения численности для настройки и уже элиминируют часть этого перебора. Следующим шагом в автоматизации всего процесса вывода демографической истории популяции является метод автоматического перебора моделей расширенного класса.

В данной главе представлено описание разработанного метода автоматического перебора моделей, а также результаты экспериментальных исследований на реальных данных.

3.1. Метод автоматического перебора моделей расширенного класса

Был разработан метод автоматического перебора моделей расширенного класса. Пользователю необходимо лишь задать минимальные и максимальные ограничения на модель и метод самостоятельно выполнит перебор моделей расширенного класса в предоставленных границах.

3.1.1. Разработка метода автоматического перебора моделей расширенного класса

На каждой итерации разработанный метод использует метод настройки параметров на основе комбинации глобальной и локальной оптимизации — разработанный комбинированный метод на основе генетического алгоритма и метода BFGS. Метод начинает с создания и настройки расширенной модели, которая удовлетворяет входным минимальным ограничениям. Затем на каждой итерации текущая модель изменяется, увеличивается число ее параметров, и снова запускается процесс настройки ее параметров по генетическим данным. При достижении максимальных ограничений на модели, метод останавливается и происходит сравнение всех перебранных моделей с помощью информационного критерия Акаике. В результате работы, выбирается настроенная модель, которая наилучшим образом описывает генетические данные. Таким образом, можно описать следующие шаги разработанного метода:

а) Создать текущую модель, как модель с минимальными ограничениями.
б) Настроить параметры для текущей модели по генетическим данным.
в) Создать следующую модель, подходящую под ограничения.

г) Повторить пункты б), в) пока не будут достигнуты максимальные ограничения.
д) Сравнить множество перебранных моделей и выбрать лучшую.

Блок-схема разработанного метода представлена на рисунке 98. Псевдокод метода представлен в листинге 7. На вход метод получает генетические данные $\mathfrak{D}$, метод вычисления правдоподобия $f_{\mathcal{M}}$, минимальные, максимальные ограничения моделей S_{min}, S_{max} соответственно и набор констант B, который задает какие параметры включены в модели и будет определен далее.

Листинг 7 – Процедура автоматического перебора моделей демографической истории популяций по генетическим данным

```
1:  function GetBestDemographicHistory(𝔇, f_M, S_min, S_max, B)
2:      results ← [ ]
3:      S_cur ← S_min
4:      M ← GetModelFromS(S_min, B)
5:      θ* ← GetBestParams(f_M, 𝔇)
6:      results[S_min] ← θ*
7:      while S_cur ≠ S^F do
8:          M, θ_cur, S_cur ← ChangeModel(M, θ*, S_cur, S_max, B)
9:          results[S_cur] ← GetBestParams(f_M, 𝔇, θ_cur)
10:     return BestByAIC(results, f_M, 𝔇)
```

На вход:

– генетические данные и информация о наличии зависимости в данных;
– метод вычисления правдоподобия;
– условия-ограничения для расширенных моделей.

На выход:

– множество моделей, подходящих под условия-ограничения, и их настроенные значения параметров;
– выбор лучшей модели с использованием критерия Акаике [43][69].

Задание ограничений на модели. Для использования метода автоматического перебора требуется определить ограничения на модели. Для задания ограничений модели предлагается использовать число временных интервалов:

– будем рассматривать максимум три популяции. В случае одной популяции в модели не будет разделения, в случае двух популяций будет одно разделение популяций и в случае трех популяций — два разделения;
– будем задавать три числа $S = \{s_1, s_2, s_3\}$. Первое число s_1 определяет число временных интервалов до первого разделения (включая самый

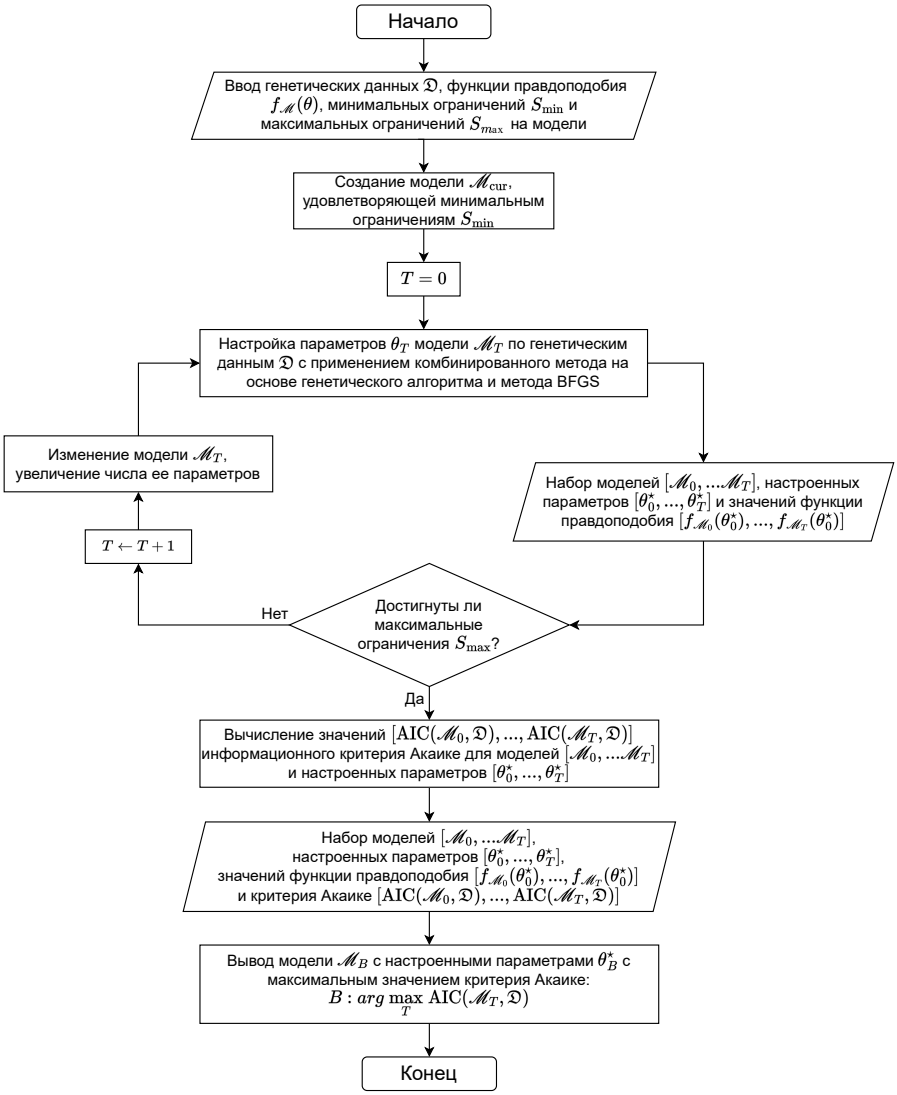


Рисунок 98 – Блок-схема разработанного метода автоматического перебора моделей расширенного класса с разным числом параметров

первый бесконечный интервал), второе число s_2 задает число временных интервалов между первым и вторым разделением, последнее третье число s_3 равно числу временных интервалов после второго разделения.

Дополнительно определим набор булевых констант B:

$$B = \{b_{\text{mig}}, b_{\text{sym\_mig}}, b_{\text{inbr}}, b_{\text{const}}, b_{\text{lin}}, b_{\text{exp}}\},$$

которые задают какие параметры включены в модели. Например, если b_{mig} истинна, то модель содержит параметры миграции для каждого временного интервала. В противном случае в модели отсутствуют параметры миграции. Обозначенные константы соответствуют следующим параметрам и применимы для всех временных интервалов модели одновременно:

- b_{mig} обозначает наличие или отсутствие параметров миграции;
- $b_{\text{sym\_mig}}$ определяет являются ли миграции симметричными или асимметричными — задаются разными параметрами;
- b_{inbr} определяет наличие или отсутствие параметров инбридинга;
- b_{const} определяет включена ли постоянная динамика численности в область определения параметров динамики;
- b_{lin} определяет включена ли линейная динамика численности в область определения параметров динамики;
- b_{exp} определяет включена ли экспоненциальная динамика численности в область определения параметров динамики.

Определение 23. Ограничение модели — три числа $S = \{s_1, s_2, s_3\}$ временных интервалов.

Для заданного ограничения $S = \{s_1, s_2, s_3\}$ и набора булевых констант B модель расширенного класса создается автоматически со всеми возможными параметрами, включая динамики изменения численности. Минимальное ограничение моделей — это минимальное $S_{\min} = \{s_1, s_2, s_3\}$ число временных интервалов в модели, максимальное ограничение — это максимально возможное $S_{\max} = \{s_1, s_2, s_3\}$ число временных интервалов в модели. Набор булевых констант B, которые задают параметры моделей. Например, если b_{mig}, остается зафиксированным во время метода автоматического перебора моделей.

Ограничение включает три числа, так как разработанный метод применим только для одной, двух или трех популяций. Пример модели, соответствующей ограничению (2,1,1), представлен на рисунке 99.

Пример модели двух популяций для структуры (2,1,0) представлен на рисунке 100. На рисунке 101 приведены демографические истории, которые соответствуют модели со следующими значениями параметров:

а) `Nanc`: 7200, `T1`: 30000, `N11`: 40000, `D11`: Lin, `f`: 0.8, `T2`: 50000, `N21`: 500, `N22`: 500, `D21`: Exp, `D22`: Sud;

б) `Nanc`: 40000, `T1`: 50000, `N11`: 7000, `D11`: Sud, `f`: 0.1, `T2`: 50000, `N21`: 5000, `N22`: 5000, `D21`: Lin, `D22`: Exp;

в) `Nanc`: 7000, `T1`: 40000, `N11`: 20000, `D11`: Sud, `f`: 0.2, `T2`: 80000, `N21`: 20000, `N22`: 500, `D21`: Exp, `D22`: Lin.

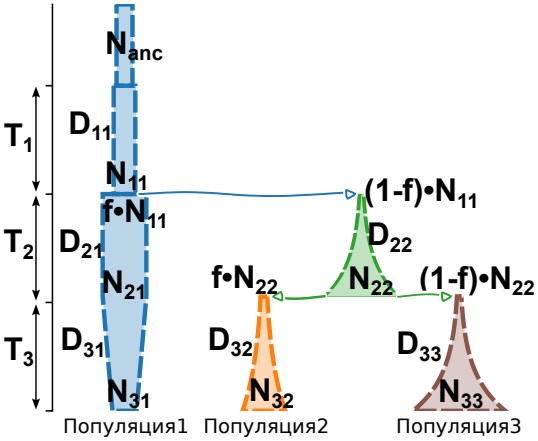


Рисунок 99 – Пример модели трех популяций, которая соответствует ограничению (2,1,1)

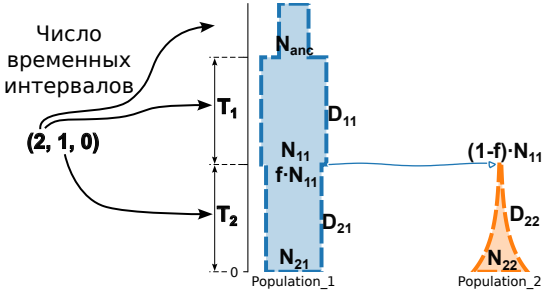
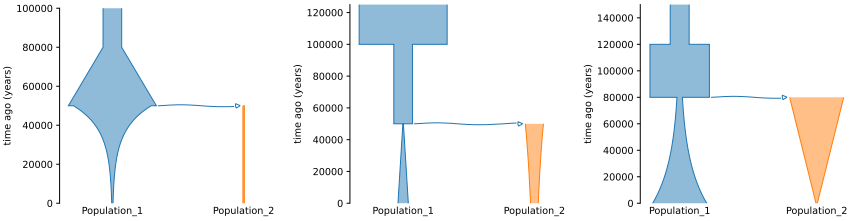


Рисунок 100 – Пример модели двух популяций, соответствующей ограничению (2,1,0)



(а) (б) (в)

Рисунок 101 – Демографические истории для модели, представленной на рисунке 100, при разных значениях ее параметров

Изменение модели демографической истории. Для изменения модели в разработанном методе автоматического перебора, был предложен метод увеличения числа временных интервалов в ограничении модели:

- из трех частей временной оси (до первого разделения, между первым и вторым разделениями, после второго разделения) случайным образом выбрать часть, где еще не достигнуто финальное число временных интервалов;
- случайным образом выбрать временной интервал в части;
- разделить временной интервал на две части, создать новые параметры, вычислить их значения в соответствии со значениями старых параметров.

Опишем разработанный метод автоматического перебора моделей с использованием введенных понятий ограничений. На первом шаге создается модель с числом временных интервалов, определенным минимальным ограничением $S_{\min}$, и осуществляется настройка ее параметров. Затем на каждом последующем шаге увеличивается число временных интервалов в модели, и производится настройка параметров для модели с большим числом параметров. Процесс завершается, когда найдены оптимальные параметры для модели с числом временных интервалов, которое соответствует максимальному ограничению $S_{\max}$. На последнем шаге происходит сравнение моделей с разными структурами с использованием статистического критерия Акаике.

Вход:

- генетические данные и информация о наличии зависимости в данных;
- набор булевых констант B, которые определяют какие параметры будут включены в модели;
- минимальное ограничение $S_{\min}$;
- максимальное ограничение $S_{\max}$.

Выход:

- оптимальные значения параметров для моделей, соответствующих разным ограничениям;
- информация о лучшей модели при сравнении с использованием информационного критерия Акаике.

Листинг 8 – Процедура изменения модели для метода автоматического перебора

function ChangeModel($\mathcal{M}, \theta^\star, S_{\text{cur}}, S_{\max}, B$)
    $D \leftarrow \{S^F[i] - S^*[i]\}_{i=1}^{P}$ ▷ Ищем где можно увеличить число временных интервалов
    $Q \leftarrow$ DiscreteRandom($\{i\}_{i=1}^{P}, D$)
    $d \leftarrow \sum_{i=1}^{Q} S^*[i] +$ UniformRandom($\{i\}_{i=1}^{S^*[Q]}$) ▷ Выбираем индекс интервала
    $S_{\text{cur}}[Q] \leftarrow S_{\text{cur}}[Q] + 1$ ▷ Новое число временных интервалов
    $\mathcal{M}_{\text{new}} \leftarrow$ InsertTimeInterval($\mathcal{M}, d$) ▷ Изменяем модель
        ▷ Вычисляем значения параметров новой модели так, чтобы она соответствовала той же истории, что и предыдущая настроенная модель
    $\theta_{\text{new}} \leftarrow$ GetParamsForSameHistory($\mathcal{M}, \theta^\star, \mathcal{M}_{\text{new}}$)
    return $\mathcal{M}_{\text{new}}, \theta_{\text{new}}, S_{\text{cur}}$

3.1.2. Реализация разработанного метода автоматического перебора моделей расширенного класса

Для реализации разработанного метода автоматического перебора моделей был разработан модуль `core`, который включает класс `CoreRun`, и был расширен модуль `models`. Структура классов показана на рисунке 102.

В модуль `models` был добавлен новый класс `StructureDemographicModel`, который наследует класс `EpochDemographicModel` расширенных моделей и описывает модели расширенного класса, определенные ограничением S и набором булевых констант B. Объект этого класса имеет следующие атрибуты:

- `initial_s` — минимальное ограничение на число временных интервалов модели;
- `initial_s` — максимальное ограничение на число временных интервалов модели;
- `has_dyns` — набор булевых констант $\{b_{\text{const}}, b_{\text{lin}}, b_{\text{exp}}\}$, определяющих область определения параметров динамики;
- `has_inbr` — булева константа b_{inr}, определяющая наличие параметров инбридинга;
- `has_migs` — булева константа b_{mig}, определяющая наличие параметров миграции;
- `sim_migs` — булева константа $b_{\text{sym\_mig}}$, определяющая являются ли миграции симметричными.

У класса `StrcutureDemographicModel` реализованы четыре процедуры. Процедура `from_s(s)` строит модель с числом интервалов, равным ограничению `s`. Процедура `get_s(s)` возвращает число временных интервалов между разделениями популяций в модели. Остальные две процедуры реализуют важные части разработанного метода автоматического перебора. Процедура

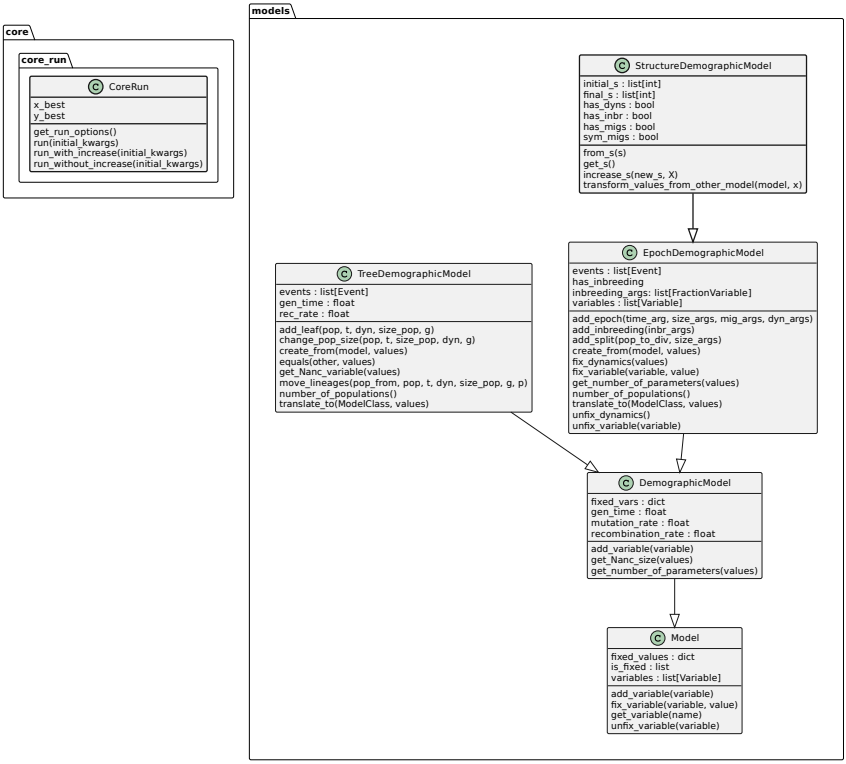


Рисунок 102 – Два класса, реализованных для разработанного метода автоматического перебора моделей

`increase_s(s)` производит изменение модели и увеличение числа временных интервалов — процедура ChangeModel в псевдокоде. Последняя процедура `transform_values_from_other_model` позволяет вычислить параметры модели по значениям вложенной модели таким образом, чтобы обе соответствовали одной демографической истории — процедура GetParamsForSameHistory в псевдокоде.

Новый модуль `core` содержит единственный класс `CoreRun`, который реализует разработанный метод автоматического перебора моделей. Метод реализован в процедуре `run_with_increase`, которая принимает на вход аргументы запуска, включающие генетические данные, метод вычисления правдоподобия и ограничения моделей, и возвращает настроенную модель, которая имеет максимальное значение критерия Акаике.

3.2. Экспериментальные исследования разработанного метода автоматического перебора моделей расширенного класса

В данном разделе представлены проведенные экспериментальные исследования по выводу демографической истории популяций с помощью разработанного метода автоматического перебора расширенных моделей.

Используемые методы вычисления правдоподобия используют статистики генетических данных, которые могут являться довольно ограниченным источником информации [92]. Например, предыдущие исследования показали, что аллель-частотный одной популяции может соответствовать различным демографическим сценариям [145, 155]. Проведенные экспериментальные исследования в разделе 2.4.1 также подтверждают эти результаты. Чем больше параметров содержат модели, тем больше шансы найти ложную демографическую историю при использовании аллель-частотного спектра. Поэтому в проведенных экспериментальных исследованиях рассмотрены довольно ограниченные модели. Однако разработанный метод может быть использован в будущем с более широкими ограничениями на модели при анализе более надежных данных.

3.2.1. Вывод демографической истории трех популяций современного человека

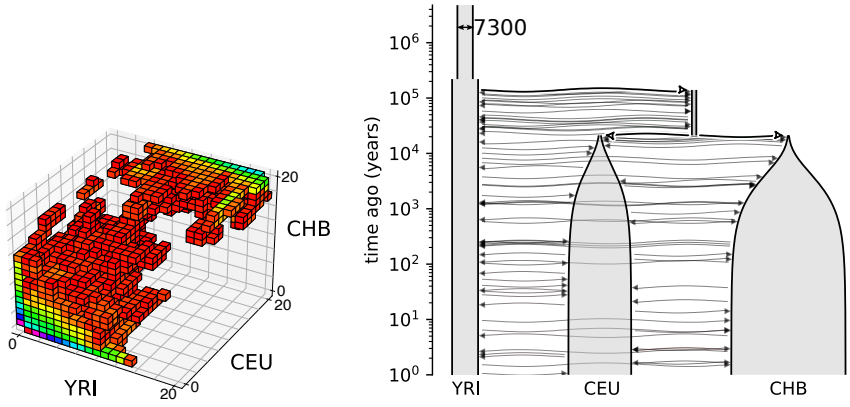
Одной из наиболее популярных демографических историй человеческих популяций является так называемая история «выхода из Африки» для трех популяций [45, 55, 156]:

- YRI — представители народа Йоруба из Ибадана, Нигерия;
- CEU — европейская популяция, представленная жителями штата Юта с предками из Северной и Западной Европы;
- CHB — представители народа Хань из Пекина, Китай.

Разработанный метод автоматического перебора моделей демографической истории был применен на данных этих трех популяций, ранее проанализированных в [45]. В качестве метода вычисления правдоподобия был использован метод, реализованный в ∂а∂i с размером сетки $pts = \{40, 50, 60\}$. На рисунке 103 представлены данные и демографическая история, полученная по этим данным в работе [45]. В работе [45] была настроена модель демографической истории с 13 параметрами с помощью множественных запусков локальной оптимизации. Размер популяции YRI в использованной модели был представлен как два интервала константной численности, миграции были симметричными, а динамики изменения численности были равны константным, за исключением последнего временного интервала для популяций CEU и CHB, где был выбран экспоненциальный рост.

Аллель-частотный спектр размера $21 \times 21 \times 21$ был построен в работе [45] на основе данных из [157] (рисунок 103а). Данные включали мутации всех биаллельных позиций из некодирующих областей 219 генов. Общая длина последовательности составила $4{,}04 \cdot 10^6$ пар оснований. Скорость мутации была выбрана,

равной $2{,}35 \cdot 10^{-8}$ на позицию на поколение, как было использовано в [45]. Для перевода времени в года было использовано среднее время одного поколения, равное 25 годам.



(а) (б)

Рисунок 103 – Генетические данные в виде аллель-частотного спектра и демографическая история, полученная ранее в работе [45]

Для применения разработанного метода автоматического перебора моделей требуется определить минимальные и максимальные ограничения. Так как был использован метод на основе аллель-частотного спектра, ограничения на число временных интервалов в моделях были выбраны небольшими. Минимальное ограничение было равно (1,1,1), что означает один временной интервал до первого разделения популяции, один интервал между первым и вторым разделениями популяций и один — после второго. Максимальное ограничение было выбрано равным (2,1,1), что описывает модель с двумя интервалами до первого разделения, с одним интервалом между первым и вторым разделениями и одним интервалом после второго разделения. Таким образом, метод перебирал только две возможные модели — модель 1 и модель 2, соответствующие минимальному и максимальному ограничению.

Настроенные параметры моделей, наилучшие значения правдоподобия, а также информационного критерия Акаике (CLAIC) могут быть найдены в работе [1] в таблице 3. Согласно этим результатам [1], наилучшая демографическая история была получена для модели, соответствующей ограничению (2,1,1) на число интервалов. На рисунке 104 представлено сравнение настроенной модели с результатом, полученным ранее в [45]. Полученная демографическая история имеет больше параметров, чем модель, использованная в [45]: 20 непрерывных параметров против 13. Несмотря на это информационный критерий Акаике (CLAIC) выделяет модель, полученную разработанным методом, как наилуч-

шую. Демографические истории, полученные разработанным методом и в работе [45], весьма схожи. Значения современных размеров европейской (CEU) и азиатской (CHB) популяций были получены немного меньше, чем ранее. Основными отличиями, однако, являются значения интенсивностей миграций и численность общей евразийской популяции, которая экспоненциально растет с 200 до 1 500 особей. Для сравнения в истории из [45] эта численность является константной, равной 2 000 особей, что является чуть менее реалистичным, чем экспоненциальный рост. Интенсивности миграций в полученном результате являются асимметричными и имеют значения выше, чем в истории, найденной в [45]. Наиболее интенсивная миграция произошла между популяциями YRI и CEU, а после второго разделения — между популяциями CEU и CHB. Отметим, что интенсивности миграций соотносятся с известным географическим положением: чем более географически удалены друг от друга популяции, тем меньше полученная интенсивность миграции.

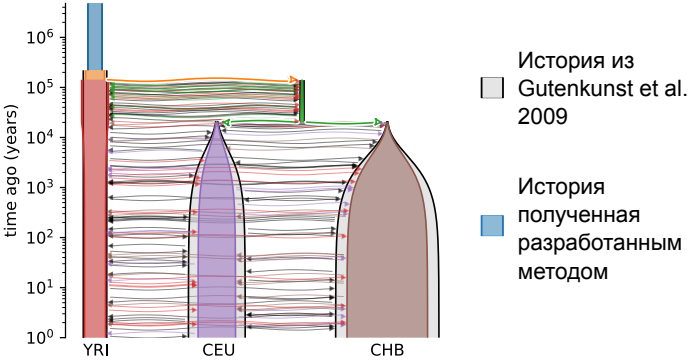


Рисунок 104 – Демографическая история, полученная разработанным методом

3.2.2. Вывод демографической истории популяций кошачьей лягушки

Разработанный метод автоматического перебора моделей демографической истории был использован для вывода демографических историй популяций кошачьей лягушки. В работе [49] были представлены генетические данные в виде аллель-частотного спектра для трех пар популяций, а также был выполнен ручной перебор моделей. При построении данных генетические данные были отобраны и только независимые позиции были использованы. Как следствие модели с разным числом параметров можно корректно сравнивать с помощью информационного критерия Акаике [43].

Рассматриваемые данные кошачьей лягушки были ранее проанализированы во второй главе в разделе 2.4.2 в экспериментальных исследованиях разработанного метода настройки параметров моделей на основе генетического алгоритма. Для всех моделей, ранее настроенных в статье [49], экспериментальные исследования позволили получить параметры, обеспечивающие лучшее значение правдоподобия.

В этом разделе представлены результаты экспериментальных исследований применения метода автоматического перебора моделей. Метод был применен для вывода демографической истории трех пар популяций: 1) северная (Northern) и южная (Southern) популяции; 2) популяции CVLN и CVLS; и 3) популяции CrossRiver и CVLN. Минимальное ограничение на модели было выбрано равным одному временному интервалу до разделения и одному временному интервалу после разделения популяций. Максимальное ограничение составило два временных интервала до разделения и три временных интервала после разделения популяций. Модели были сравнены с использованием информационного критерия Акаике (AIC).

Результаты представлены в работе [1] в таблице S2 для популяций Northern, Southern, в таблице S3 для популяций CVLN, CVLS, в таблице S4 для популяций CrossRiver, CVLN. полученные демографические истории изображены на рисунке 105.

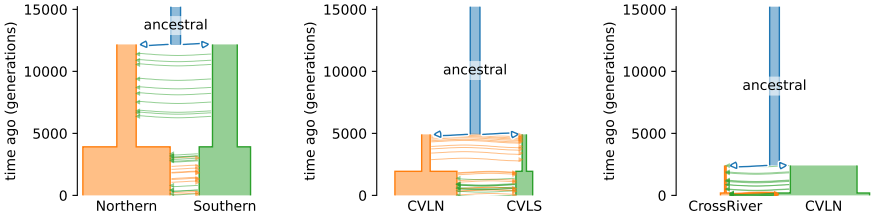


Рисунок 105 – Полученные демографические истории для различных пар популяций кошачьей лягушки

Все демографические истории популяций кошачьей лягушки, полученные разработанным методом автоматического перебора, имеют один временной интервал до разделения и два интервала после разделения предковой популяции. Две из трех полученных демографических историй имеют наилучшие значения информационного критерия Акаике при сравнении с моделями, перебранными вручную в работе [49] (таблицы S2 и S4 в работе [1]). Для данных популяций CVLN и CVLS демографическая история, полученная методом автоматического перебора, имеет значение AIC хуже, чем наилучшая модель, полученная ручным перебором в [49]. Однако заметим, что на основании этой модели, построенной

автоматически, можно рассмотреть аналогичную модель с исключенным параметром миграции из популяции CVLS в популяцию CVLN. Для такой вложенной модели значение информационного критерия Акаике будет равно 926.2, что является наилучшим значением среди всех ранее полученных моделей.

Настроенные модели демонстрируют односторонние миграции во время первого временного интервала после разделения. Наличие одного интервала до разделения предковой популяции и двух интервалов после разделения согласуется с результатами, полученными в [49]. Таким образом, разработанный метод автоматического перебора моделей позволил построить и настроить модели демографических историй, обеспечивающих наилучшее значение правдоподобия, чем было получено ранее с использованием ручного перебора.

3.2.3. Вывод демографической истории двух и трех популяций голубой акулы

Разработанный метод автоматического перебора моделей был использован для вывода демографической истории популяций голубых акул. Генетические данные включали полногеномные последовательности 376 особей (рисунок 106). Было выделено три популяции, представленные в этих данных:

- северная популяция (Northern) — север Атлантического океана и Средиземное море:
- южная популяция (Southern) — Индийский океан и юго-западная часть Тихого океана;
- южноафриканская популяция (SAF).

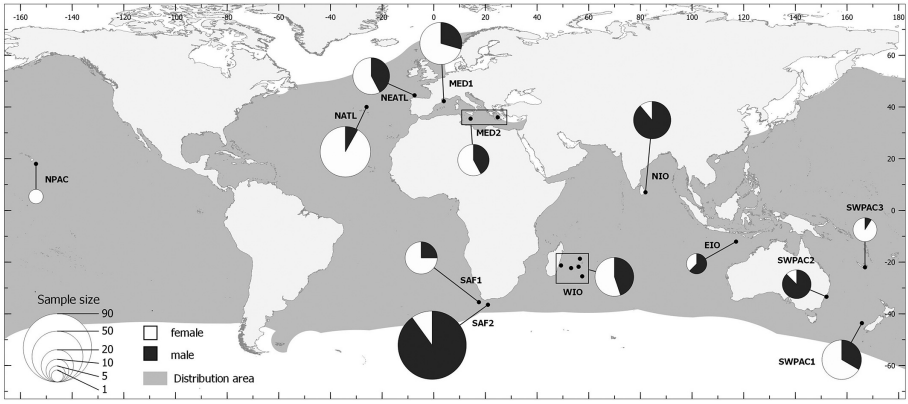


Рисунок 106 – Географическое расположение образцов генетических данных. Источник: [5]

На основе генетических данных были построены несколько аллель-частотных спектров, представленных на рисунке 107. Для вывода демографической истории были построены три спектра. Первый спектр для двух популяций

(рисунок 107а) размера 51×51 был использован для вывода демографической истории северной и южной популяций. Для вывода демографической истории трех популяций было использовано два спектра разного размера: $21 \times 21 \times 21$ (рисунок 107б) и $51 \times 51 \times 51$ (рисунок 107в).

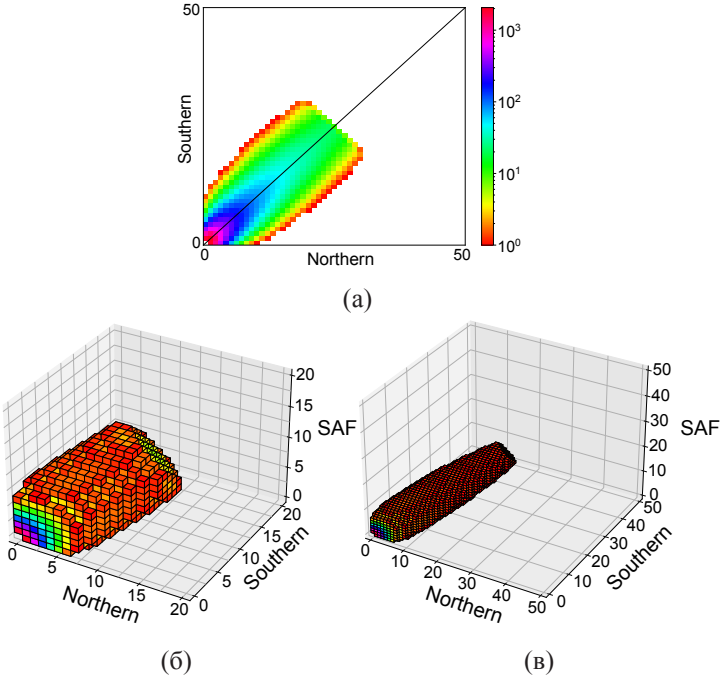


Рисунок 107 – Генетические данные в виде аллель-частотных спектров для (а) двух популяций, (б) и (в) трех популяций

Для вывода демографической истории популяций был использован метод вычисления правдоподобия, реализованный в *moments*. Каждая настройка параметров была повторена 50 раз и выбраны лучшие результаты. Скорость мутаций была выбрана, равной 10^{-8} на позицию на поколение [158, 159], длина генома составила 2 598 195 пар оснований [160]. Для перевода значений параметров времени из поколений в года было использовано среднее время одного поколения, равное девяти годам, поскольку ранее опубликованные оценки составляли 8,1 лет [161] и 8,2, 9,8 лет [162] для южноафриканской и северной популяций.

Для вывода демографической истории популяций голубой акулы был разработан многоступенчатый подход, схема которого изображена на рисунке 108. Сначала, был применен разработанный метод автоматического перебора расширенных моделей для вывода демографической истории северной (Northern) и южной (Southern) популяций по аллель-частотному спектру, представленному

на рисунке 107а. Этот шаг позволяет автоматически перебрать модели, а также настроить динамики изменения численности популяций.

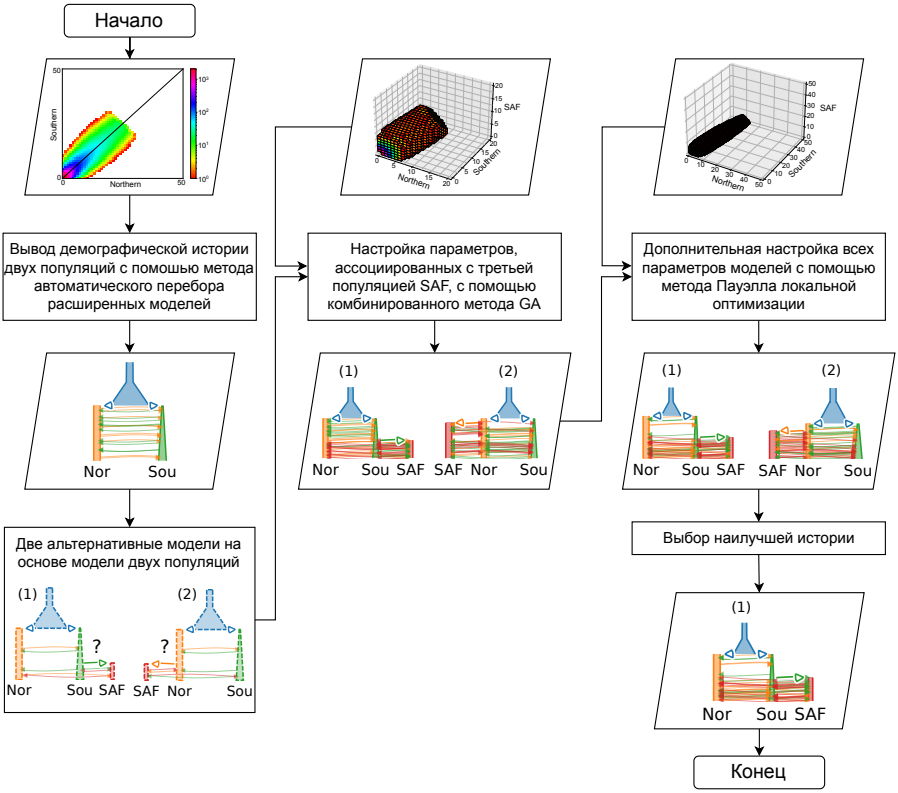


Рисунок 108 – Схема вывода демографической истории популяций голубой акулы

Затем, наилучшая модель двух популяций была модифицирована и третья популяция была добавлена. Было рассмотрено две модифицированные модели: 1) модель 1, в которой южноафриканская популяция SAF отделилась от южной популяции Southern, 2) модель 2, в которой южноафриканская популяция SAF отделилась от северной популяции Northern. Численность южноафриканской популяции была выбрана постоянной, было добавлено семь новых параметров. Для увеличения точности финальных результатов настройка параметров двух модифицированных моделей трех популяций была произведена в два этапа. На первом этапе все параметры, ассоциированные с моделью двух популяций, были фиксированы. С помощью разработанного метода на основе комбинации генетического алгоритма и метода Пауэлла локальной оптимизации была выполнена

настройка семи новых параметров, ассоциированных с историей южноафриканской популяции SAF. Настройка была проведена по аллель-частотному спектру размера $21 \times 21 \times 21$, представленному на рисунке 107б. На втором этапе настройка была проведена для всех параметров моделей с использованием метода Пауэлла локальной оптимизации по аллель-частотному спектру размера $51 \times 51 \times 51$, представленному на рисунке 107в. В конце две модели были сравнены по значению правдоподобия — они имеют равное число параметров.

Для вывода демографической истории двух популяций был использован разработанный метод автоматического перебора моделей. Были использованы следующие ограничения на модели: минимальное число временных интервалов — (1,1,0), максимальное число временных интервалов — (2,1,0). Таким образом, метод перебирал только две модели. Первая модель имела один временной интервал до разделения и один после. Вторая модель включала два временных интервала до разделения и один после. Такое малое число временных интервалов обусловлено ограниченными возможностями аллель-частотного спектра, указанными ранее. Модели были сравнены с использованием модифицированного критерия Акаике (CLAIC), так как данные имели зависимости.

Результаты показали, что модель с двумя временными интервалами до разделения имеет лучшее значение CLAIC и, следовательно, лучше описывает генетические данные. Финальная настроенная модель двух популяций представлена на рисунке 109. Значения настроенных параметров могут быть найдены в работе [5] в таблице 4. Численность предковой популяции составила около $30\,000$ особей, и она линейно росла до $170\,000$ особей, после чего разделилась на северную и южную популяции. Этот линейный рост начался около $1\,170\,000$ лет назад, а разделение около $5\,000$ лет назад. Северная популяция имела постоянную численность в $5\,000$ особей, а южная популяция имела слабый линейный рост численности от $4\,000$ особей в момент образования до $6\,000$ особей в настоящий момент. Миграции между популяциями были асимметричные и миграция из южной популяции в северную гораздо интенсивнее миграции в обратном направлении.

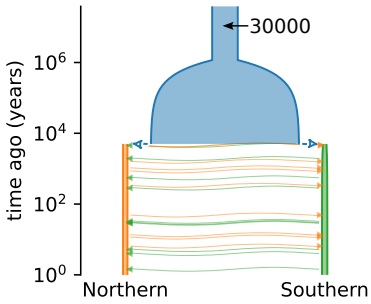


Рисунок 109 – Демографическая история двух популяций голубой акулы

Вывод демографической истории трех популяций с использованием двух модифицированных моделей и многоступенчатой настройки параметров показал, что модель 1 лучше описывает генетические данные, чем модель 2. В частности, это означает, что южноафриканская популяция отделилась от южной популяции, а не от северной. Рисунок лучшей демографической истории представлен на рисунке 109. Настроенные параметры представлены в [5] в таблице S7.

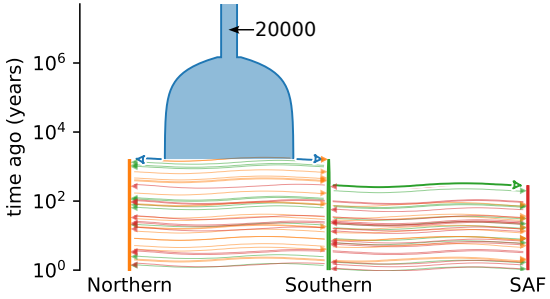


Рисунок 110 – Демографическая история трех популяций голубой акулы

История трех популяций немного отличается от истории двух популяций. Отметим, что она является более надежной, так как включает больше популяций и учитывает больше генетических данных. Размер предковой популяции составил 20 000 особей. Эта численность начала линейно увеличиваться около 1,4 миллиона лет назад и выросла до 165 000 особей. Предковая популяция разделилась 1 600 лет назад на северную и южную популяции. От южной популяции 300 лет назад отделилась южноафриканская популяция. Численность северной популяции составила 2 000 особей, численность южной популяции выросла с 1 500 до 3 000 особей, а размер южноафриканской популяции оставался примерно постоянным около 1 500 особей. Миграции между северной и южноафриканской популяциями отсутствовали. Миграция из южноафриканской популяции в южную была самая интенсивная.

Согласно полученным результатам, линейный рост предковой популяции начался в раннем плейстоцене, а раскол на северную и южную популяции произошел во время эпохи голоцена. Апробация результатов коллегами из области зоологии [5] позволила предположить, что палеоклиматические события спровоцировали расхождение северной и южной популяций. В эпоху голоцена температура морской поверхности в тропиках имела тенденции к потеплению и так продолжалось до момента времени 5 000 лет назад. После этого момента и до настоящего времени температура морской поверхности хоть и колебалась, но имела тенденцию к глобальной стабилизации. Например, в работе [71] было выявлено потепление примерно на $2°C$ в западной тропической части Атлантического океана и восточной тропической части Тихого океана с раннего голоцена до настоя-

щего времени. Было выявлено глобальное похолодание в Северном полушарии около 5 000 лет назад [72].

Записи озерных отложений из Гренландии также позволяют предположить, что около 4 500 и 650 лет назад температура поверхности перестала нагреваться и начала колебаться, в том числе и в отрицательную сторону [72, 73]. Более того, это колебание температуры морской поверхности происходило и в Южном полушарии: было показано, что температура в австралийско-новозеландском регионе во время голоцена имела тенденцию к снижению [163], как и в Южном океане [164, 165]. Согласно полученным результатам разделение предковой популяции как раз произошло около 5 000 лет назад, когда произошло изменение тенденции изменения температуры. Эти колебания температуры морской поверхности в прошлом могли способствовать разделению северной и южной популяций. Сдвиги в сезонах размножения голубых акул между двумя полушариями могли способствовать сохранению этого разделения: в Северном полушарии размножение происходит летом (июль, август) [166], как и в юго-западной экваториальной части Атлантического океана [167], а в Индийском океане — с октября по декабрь [168].

Полученные демографические истории трех популяций показала, что численность голубых акул сильно сократилась при разделении предковой популяции: численность северной и южной популяций при разделении предковой популяции составили только 2–3% от размера предковой популяции. Кроме того, результаты показали довольно низкие современные размеры популяции. Полученные размеры 4 000 – 6 000 особей согласуются с оценками, полученными ранее в [57, 58].

Выводы по главе 3

1. Разработан метод автоматического перебора расширенных моделей демографической истории одной, двух и трех популяций по генетическим данным.
2. Метод был применен для вывода демографической истории «выхода из Африки» для трех популяций современных людей. Полученная история имеет не только лучшее значение правдоподобия, чем ранее полученная история по тем же данным, но и лучшее значение информационного критерия Акаике. Результаты согласуется с другими исследованиями.
3. Метод был применен для вывода демографической истории трех пар популяций кошачьей лягушки. Полученные демографические истории имеет не только лучшее значение правдоподобия, чем истории, ранее полученные по тем же данным ручным перебором моделей, но и лучшие значения информационного критерия Акаике.
4. Выведена демографическая история трех популяций голубой акулы по данным, которые ранее не были проанализированы. Полученная демо-

графическая история согласуется с другими исследованиями этого биологического вида.

Глава 4. Программный комплекс GADMA для вывода демографической истории популяций по генетическим данным и расширение библиотек *stdpopsim* и *demes*

В данной главе описаны программный комплекс GADMA, который реализует все разработанные модели и методы, а также расширение библиотек *stdpopsim* и *demes*, использованные для проведения экспериментальных исследований и представления результатов.

В разделе 4.1 приведено описание программного комплекса GADMA (Global search Algorithm for Demographic Model Analysis). Все экспериментальные исследования методов в данной работе были проведены с использованием этого программного комплекса. Приведена структура комплекса и ее основные компоненты, часть которых была описана ранее.

Раздел 4.1 включает описание существующих библиотек *stdpopsim* и *demes*, которые были расширены для использования в данной работе. Библиотека *stdpopsim* позволяет проводить симуляции генетических данных с использованием каталога видов и их демографических историй. Основное назначение библиотеки *demes* в текстовом и визуальном представлении демографических историй. Все изображения историй в данной работе были получены с использованием *demes*.

4.1. Программный комплекс GADMA для вывода демографической истории популяций по генетическим данным

Все разработанные модели и методы были реализованы в программном комплексе GADMA. Основным назначением GADMA является повышение эффективности, а также снижение уровня сложности процесса вывода демографической истории популяций по генетическим данным. Программный комплекс GADMA с открытым исходным кодом доступен в репозитории `https://github.com/ctlab/GADMA`. Его документация доступна `https://gadma.readthedocs.io`.

Программный комплекс включает в себя набор методов настройки параметров моделей и «движков» — методов вычисления правдоподобия. Всего реализованы четыре «движка», которые соответствуют существующим методам: *∂a∂i*, *moments*, *momi2*, *momentsLD*. Программный комплекс имеет два режима:

а) Режим заданной модели. В этом режиме пользователь самостоятельно задает интересующую его модель демографической истории популяций. Комплекс GADMA настраивает параметры заданной модели.

б) Режим автоматического перебора моделей. Пользователь задает минимальные и максимальные ограничения на модели, а GADMA производит автоматический перебор моделей в пределах заданных ограничений, настраивает их параметры и выбирает наилучшую модель.

Для запуска GADMA пользователю достаточно выбрать режим, движок и метод настройки параметров моделей.

4.1.1. Структура программного комплекса GADMA

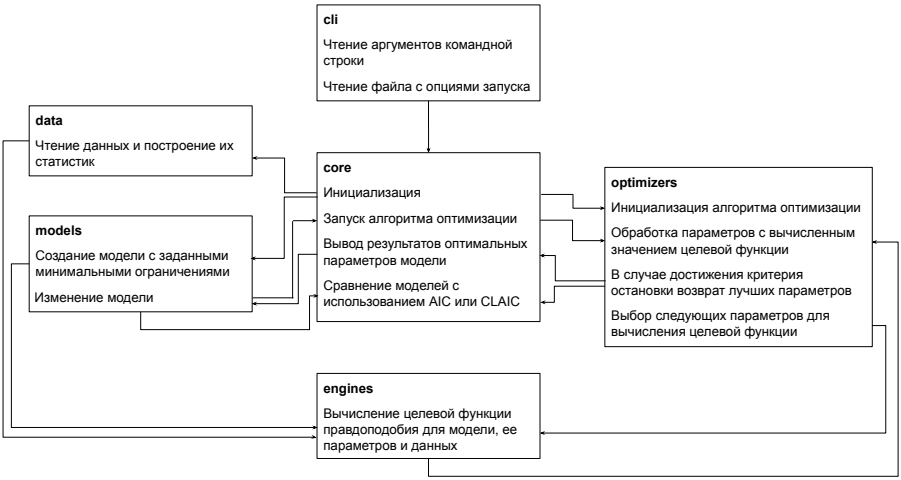


Рисунок 111 – Структура программного комплекса GADMA

На рисунке 111 приведена общая структура программного комплекса. Она включает в себя шесть основных модулей: `core`, `cli`, `data`, `models`, `engines` и `optimizers`.

Модуль `core` является входной точкой программного комплекса и управляет остальными модулями. Основными входными данными является аргументы командной строки и файл с опциями для запуска. Чтение входных данных происходит в **модуле `cli`** и возвращается в корневой модуль. Затем происходит инициализация согласно полученным опциям. Происходит чтение генетических данных с помощью модуля `data` и создание модели демографической истории согласно опциям, выбранным пользователем. Если выбран режим заданной модели, то именно она и анализируется. В противном случае при выборе режима автоматического перебора создается модель, соответствующая минимальным ограничениям, которая в дальнейшем будет изменяться. Затем происходит запуск метода оптимизации из модуля `optimizers`, который возвращает настроенные параметры модели. Модуль выводит найденные параметры. Если был выбран режим автоматического перебора, то следует обращение к модулю `models` для изменения модели и затем снова происходит запуск метода оптимизации. Так повторяется до тех пор, пока не будет достигнуты максимальные ограничения на модели. В конце работы в режиме автоматического перебора модуль `core` сравнивает все полученные модели с использованием метрик AIC или

CLAIC. Выбор метрики зависит от наличия зависимостей в данных — информация, которая указывается пользователем.

Модуль `data` содержит инструменты для работы и хранения генетических данных. Он позволяет прочитать генетические данные, которые могут быть представлены в разных форматах, и строит по ним статистики для дальнейшего использования такие, как аллель-частотный спектр или статистики неравновесного сцепления генов. Подробное описание этих статистик представлено в разделе 1.4.2. Рисунок 112 показывает структуру классов модуля. Абстрактный класс `DataHolder` хранит указатель на генетические данные. Класс `VCFDataHolder` хранит генетические данные в формате VCF [169]. Класс `SFSDataHolder` позволяет хранить указатель на файл с аллель-частотным спектром, который может быть в нескольких форматах [45, 46, 170].

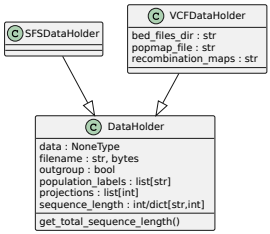


Рисунок 112 – Структура классов модуля `data` программного комплекса GADMA

Модуль `models` предназначен для создания и хранения моделей. Он уже был описан ранее в разделах 2.1 и 3.1.2. Структура классов была представлена на рисунках 38 и 102.

Модуль `optimizers` содержит методы настройки параметров моделей по генетическим данным. На рисунке 55 ранее была представлена структура классов этого модуля. Модуль реализует метод на основе комбинации методов глобальной и локальной оптимизации. На выбор пользователя представлены следующие методы глобальной оптимизации:

– генетический алгоритм, описанный в разделе 2.2 и реализованный классом `GeneticAlgorithm`;
– байесовская оптимизация, описанная в разделе 2.3 и реализованная классами `SMACBayesianOptimizer` и `SMACBOEnsemble`.

GADMA предоставляет выбор из следующих методов локальной оптимизации:

– метод BFGS;
– метод L-BFGS-B;
– метод Пауэлла;
– метод Нелдера-Мида.

Модуль `engines` включает реализацию движков GADMA — методов вычисления правдоподобия, которые используются методами оптимизации для настройки параметров моделей по генетическим данным. Структура классов представлена на рисунке 113.

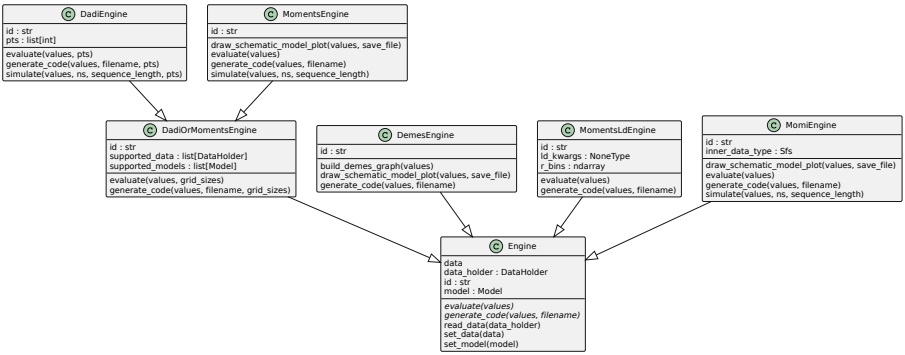


Рисунок 113 – Структура классов модуля `engines` программного комплекса GADMA

Модуль включает абстрактный класс `Engine`, от которого наследуются все движки. Объекты этого класса имеют атрибут `id` для идентификации движка, генетические данные `data_holder`, представленные объектом класса `DataHolder` и атрибут `model` заданной модели демографической истории популяций, являющейся объектом класса `Model`. Основной процедурой класса `Engine` является абстрактная процедура `evaluate`, которая вычисляет значение правдоподобия генетических данных при условии модели с заданными параметрами `values`. Процедура `generate_code` генерирует спецификацию модели с заданными параметрами с использованием интерфейса движка. Например, для движка `dadi` — это будет процедура для языка программирования Python, пример которой показан на рисунке 15.

Программный комплекс GADMA реализует четыре движка для вычисления правдоподобия:

- движок `dadi`, реализующий метод аппроксимации диффузией библиотеки *∂a∂i* — класс `DadiEngine`;
- движок `moments`, реализующий метод моментов для аллель-частотного спектра библиотеки *moments* — класс `momentsEngine`;
- движок `momi2`, реализующий метод непрерывной модели Морана библиотеки *momi2* — класс `MomiEngine`;
- движок `momentsLD`, реализующий метод моментов для статистик неравновесного сцепления генов библиотеки *momentsLD* — класс `MomentsLdEngine`.

Один дополнительный движок `demes`, реализованный классом `DemesEngine`, включен в GADMA для визуального представления демографических историй. Он использует библиотеку *demes*, которая будет описана далее. Для рисования демографических историй используется процедура `draw_schematic_model_plot`. GADMA предоставляет выбор из трех движков для визуального представления демографических историй:

– движок `moments`;
– движок `momi2`;
– движок `demes`.

4.1.2. Входные данные и интерфейс запуска

На вход программный комплекс GADMA принимает файл с опциями запуска, список которых с описанием приведен в таблице 15. Запуск GADMA для заданного файла `params_file` с опциями выполняется из командной строки следующим образом:

```
$ gadma -p params_file
```

Пример входного файла с опциями изображен на рисунке 114.

Таблица 15 – Список опций входного файла программного комплекса GADMA

`Output directory`	Директория для записи результатов
Информация о генетических данных	
`Input data`	Путь к файлу с генетическими данными
`Population labels`	Названия рассматриваемых популяций
`Projections`	Число образцов для каждой популяции
`Sequence length`	Длина представленной последовательности
`Linked SNP's`	Информация о наличии или отсутствии зависимостей в генетических данных, которая используется для выбора AIC или CLAIC
`Directory with bootstrap`	Директория с множеством сгенерированных данных для вычисления CLAIC
Информация о популяциях	
`Mutation rate`	Скорость мутации одной позиции генома на одно поколение
`Recombination rate`	Вероятность рекомбинации между позициями генома, расположенными на расстоянии миллиона пар оснований
`Time for generation`	Среднее время одного поколения
Выбор движка	
`Engine`	Идентификатор движка — метода вычисления правдоподобия
Режим настройки параметров заданной модели	
`Custom model`	Путь к файлу со спецификацией модели
`Lower bound`	Нижние границы значений параметров
`Upper bound`	Верхние границы значений параметров
Режим автоматического перебора моделей	
`Initial structure`	Минимальное ограничение моделей, задающее минимальное число временных интервалов
`Final structure`	Максимальное ограничение моделей, задающее максимальное число временных интервалов
`Dynamics`	Множество значений параметров динамики изменения численности популяций
`No migrations`	Наличие или отсутствие параметров непрерывной миграции в моделях
`Symmetric migrations`	Определяет являются ли миграции симметричными
`Inbreeding`	Наличие или отсутствие параметров инбридинга в моделях
Выбор компонент метода настройки параметров моделей	
`Global optimizer`	Метод глобальной оптимизации
`Local optimizer`	Метод локальной оптимизации

(Продолжение таблицы 15)

Число повторов вывода демографической истории	
`Number of repeats`	Число повторов вывода демографической истории популяций для выбора наилучшего результата
`Number of processes`	Число доступных ядер для параллельного запуска повторов
Опции выходных данных	
`Model plot engine`	Выбор движка для визуального представления демографических историй
`Draw models every N iteration`	Частота генерации визуального представления демографических историй с использованием выбранного движка
`Print models' code every N iteration`	Частота генерации текстового представления демографических историй для всех движков GADMA
`Verbose`	Частота вывода промежуточных результатов

```
%%bash
cat params_file
```

```
# Set data first
Input file: dadi_2pops_CVLN_CVLS_snps.txt
# As we have SNP's file format we need to set the following settings:
Population labels: CVLN, CVLS
Projections: 10, 10  # we downsample AFS for fast example 30, 18 original sizes
Outgroup: False


# Output folder. It should be empty.
Output directory: gadma_result


# Set engine for simulations. We use default moments
Engine: moments
# But we specify grid size for dadi for its usage in generated code
Pts: 30, 40, 50


# Now set structures
Initial structure: 1,1
Final structure: 2,1

# We could specify some additional properties of our model
# We want asymmetric migrations
Symmetric migrations: False
# If True then any population splits into two new in some fraction.
# If False then two new populations after split have its own initial
# sizes. We choose the last option.
Split fractions: False


# No output in stdout
Silence: True

# How many repeats to run and how many processes to use.
Number of repeats: 2
Number of processes: 2
```

Рисунок 114 – Пример входного файла с опциями запуска для программного комплекса GADMA

4.1.3. Выходные данные

Все промежуточные и конечные результаты работы программного комплекса GADMA записываются и сохраняются в указанную пользователем директорию (`Output directory`). Пример структуры этой директории показан на рисунке 115. GADMA позволяет вывести демографическую историю популяций, используя несколько запусков-повторов и выбор наилучшего результата. В основной директории создаются пронумерованные папки, которые содержат результаты каждого повтора. Например, запуск GADMA, выходные данные которого показаны на рисунке 115, содержал два повтора вывода демографической истории. В основную часть сохраняется наилучший результат среди повторов.

```
gadma_result
├── 1
│   ├── current_best_logLL_model_dadi_code.py
│   ├── current_best_logLL_model_moments_code.py
│   ├── eval_file
│   ├── final_best_logLL_model_dadi_code.py
│   ├── final_best_logLL_model_moments_code.py
│   ├── final_best_logLL_model.png
│   ├── GADMA_GA.log
│   ├── save_file
│   ├── save_file_1_1
│   └── save_file_2_1
├── 2
│   ├── current_best_logLL_model_dadi_code.py
│   ├── current_best_logLL_model_moments_code.py
│   ├── eval_file
│   ├── final_best_logLL_model_dadi_code.py
│   ├── final_best_logLL_model_moments_code.py
│   ├── final_best_logLL_model.png
│   ├── GADMA_GA.log
│   ├── save_file
│   ├── save_file_1_1
│   └── save_file_2_1
├── best_logLL_model_dadi_code.py
├── best_logLL_model_moments_code.py
├── best_logLL_model.png
├── extra_params_file
├── GADMA.log
└── params_file
```

Рисунок 115 – Пример структуры директории с результатами запуска

Для полученной демографической истории в директорию записываются текстовое и визуальное представление. Текстовое представление генерируется для всех доступных движков, например, на рисунке 115 файлы `best_logLL_model_dadi_code.py` и `best_logLL_model_dadi_code.py` являются текстовым представлением полученной демографической истории для *dadi* и *moments* соответственно. Визуальное представление демографической истории представлено в файле `best_logLL_model.png`, оно включает изображение демографической истории, сгенерированное одним из движков, а также представление использованных статистик генетических данных. На рисунке 116 приведен пример выходного визуального изображения.

Запуск GADMA происходит из командной строки, где также выводится информация о результатах запуска. Пример вывода GADMA в командной строке показан на рисунке 117.

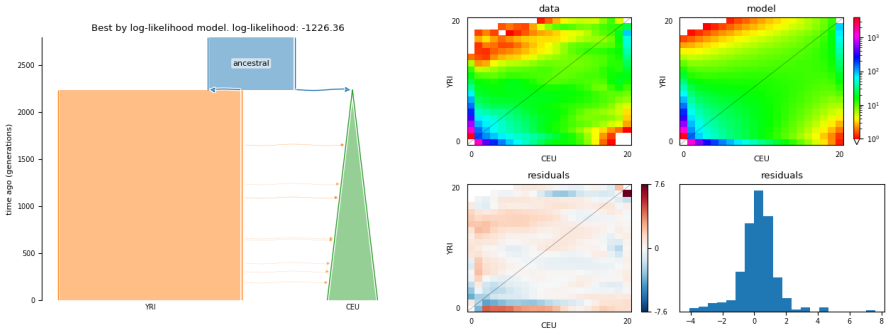


Рисунок 116 – Пример визуального представления демографической истории популяций и использованных статистик генетических данных, созданное GADMA

```
Data reading
Read preprocessed data
Number of populations: 2
Projections: [20, 20]
Population labels: ['YRI', 'CEU']
Outgroup: None
--Successful data reading--

--Successful arguments parsing--

Parameters of launch are saved in output directory: output/params_file
All output is saved in output directory: output/GADMA.log

--Start pipeline--
Run launch number 4
Run launch number 3
Run launch number 1
Run launch number 2

[026:11:23]
All best by log-likelihood models
Number  log-likelihood  Model
Run 1   -95.95  [ [Nanc = 7729],        [ 2806.556(t1), [2745.65(nu11)], [Exp(dyn11)] ],        [ 1 pop split   75.45% (s1)
[2071.542(s1*nu11), 674.108((1-s1)*nu11)] ],    [ 2169.764(t2), [15377.474(nu21), 5530.813(nu22)], [[0, 6.70e-05(m2_12)],
[6.70e-05(m2_12), 0]], [Sud(dyn21), Lin(dyn22)] ] ]     f
Run 2   -141.34 [ [Nanc = 7985],        [ 2844.141(t1), [1535.839(nu11)], [Exp(dyn11)] ],       [ 1 pop split   13.19% (s1)
[202.501(s1*nu11), 1333.338((1-s1)*nu11)] ],    [ 3666.04(t2), [15732.236(nu21), 3193.998(nu22)], [[0, 9.21e-05(m2_12)],
[9.21e-05(m2_12), 0]], [Sud(dyn21), Sud(dyn22)] ] ]     f
Run 4   -191.74 [ [Nanc = 6750],        [ 2101.89(t1), [589.234(nu11)], [Lin(dyn11)] ], [ 1 pop split   63.02% (s1)
[371.318(s1*nu11), 217.916((1-s1)*nu11)] ],     [ 1221.031(t2), [11515.474(nu21), 181624.759(nu22)], [[0, 5.70e-05(m2_12)],
[5.70e-05(m2_12), 0]], [Sud(dyn21), Exp(dyn22)] ] ]     f
Run 3   -244.30 [ [Nanc = 9752],        [ 3755.298(t1), [7740.985(nu11)], [Lin(dyn11)] ],       [ 1 pop split   0.10% (s1)
[7.741(s1*nu11), 7733.244((1-s1)*nu11)] ],      [ 22987.06(t2), [15177.316(nu21), 3105.092(nu22)], [[0, 1.16e-04(m2_12)],
[1.16e-04(m2_12), 0]], [Lin(dyn21), Lin(dyn22)] ] ]     f

You can find code and the picture of the best model in the output directory.


--Finish pipeline--
```

Рисунок 117 – Пример вывода GADMA в командной строке

Функциональные ограничения на применение. При использовании метода автоматического перебора моделей вывод демографической истории в GADMA ограничен тремя популяциями в силу ограничений самого метода. В случае режима настройки параметров заданной модели демографической истории популяций программный комплекс GADMA ограничен применимостью методов вычисления правдоподобия включенных движков. Так, например, движки *∂a∂i* и *moments* могут анализировать до трех и пяти популяций соответственно, а метод вычисления правдоподобия, реализованный в *momi2*, не поддерживает непрерывные миграции. Движок *∂a∂i* является единственным движком, поддерживающим вывод коэффициентов инбридинга. Полный список ограничений движков представлен в таблице 16.

Таблица 16 – Ограничения программного комплекса GADMA при использовании разных движков

	∂a∂i	*moments*	*momi2*	*momentsLD*
Максимальное число популяций в режиме заданной модели	Три	Пять	Произвольное	Произвольное
Максимальное число популяций в режиме автоматического перебора	Три	Три	Три	Три
Учитывает степень рекомбинации	Нет	Нет	Нет	Да
Поддерживает линейное изменение численности	Да	Да	Нет	Да
Поддерживает вывод непрерывной миграции	Да	Да	Нет	Да
Поддерживает вывод коэффициентов инбридинга	Да	Нет	Нет	Нет

4.1.4. Разработка и сопровождение программного комплекса

Исходный код программного комплекса GADMA находится в **открытом доступе** на GitHub под лицензией GPLv3: `https://github.com/ctlab/GADMA`. При разработке была использована распределённая система управления версиями (git), что позволило привлечь группу специалистов к совместной работе над проектом. Всего в разработке программного комплекса приняли участие семь человек. Разработчиком, внесшим наибольший вклад (более 85 %), является диссертант, остальные участники — студенты, которые выполняли работу под руководством диссертанта.

Веб-сервис GitHub позволил осуществлять сопровождение программного комплекса за счет использования **системы отслеживания ошибок** (issue). Это позволило обнаружить и исправить ряд дефектов программного комплекса, а также получить отзывы и пожелания внешних участников.

Публичные версии программного комплекса доступны в каталоге *PyPI* (Python Package Index) программного обеспечения, написанного на языке программирования Python, и в дистрибутиве *Anaconda*. Это означает, что GADMA

может быть легко установлена вместе с зависимостями с помощью команд *pip* и *conda* в терминале.

Была использована **система автоматизации *GitHub Actions*** для программного комплекса GADMA. *GitHub Actions* — система непрерывной интеграции и непрерывного развертывания, которая позволяет выполнить сборку, тестирование и публикацию кода программного обеспечения.

Общедоступная документация была создана с использованием генератора документации *Sphinx*, который позволяет на основе файлов, представленных в формате reStructuredText построить документацию в формате HTML для дальнейшего размещения в сети интернет. Документация включает в себя подробное описание установки и использования программного комплекса, набор примеров использования и полученных результатов, список частозадаваемых вопросов с ответами и список ссылок на исследовательские работы. Кроме того, GADMA является библиотекой и может быть использована для решения задач в других областях, где возникает задача оптимизации, поэтому документация включает автоматически созданную документацию интерфейса прикладного программирования (API) GADMA, в которой описаны основные классы. При каждом обновлении кодовой базы проекта система *GitHub Actions* автоматически создает документацию в формате HTML и размещает новую версию в сети интернет по ссылке: `https://gadma.readthedocs.io`.

Для программного комплекса GADMA была обеспечена возможность проведения **модульного тестирования** (unit testing). Тесты могут быть запущены локально с использованием исходного кода, однако их основное назначение — автоматическое тестирование в системе *GitHub Actions* на различных платформах при обновлении кодовой базы. Система *GitHub Actions* автоматически собирает комплекс и запускает тесты для следующих платформ: Linux, Windows, MacOS. По результатам автоматического тестирования создается отчет о покрытии кода тестами. Этот отчет загружается на сервис *CodeCov*, где он является общедоступным по ссылке: `https://app.codecov.io/gh/ctlab/GADMA`. Покрытие кода последней версии GADMA составило 96,65% и пример отчета, доступного на сервисе *CodeCov*, показан на рисунке 118.

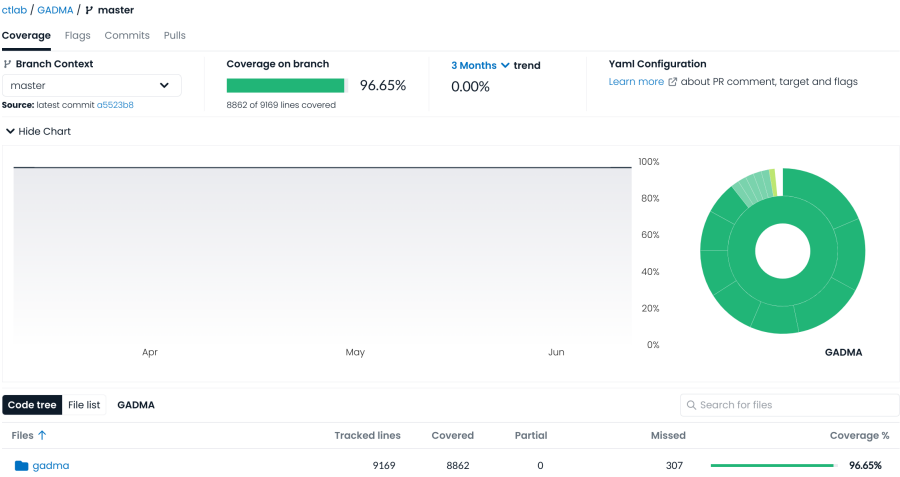


Рисунок 118 – Пример отчета о покрытии программного кода GADMA тестами на сервисе *CodeCov*

4.2. Расширение библиотек *stdpopsim* и *demes* для проведения экспериментальных исследований и представления результатов

В данном разделе описаны основные изменения, сделанные для расширения библиотек *stdpopsim* и *demes*. Эти библиотеки были использованы при проведении экспериментальных исследований в данной работе, библиотека *demes* была также использована для визуального представления демографических историй.

4.2.1. Расширение библиотеки *stdpopsim* для симулирования генетических данных

Библиотека *stdpopsim* — поддерживаемая сообществом PopSim библиотека стандартных моделей популяционной генетики для симулирования генетических данных [6, 7]. Библиотека предоставляет каталог существующих биологических видов (рисунок 119). Для каждого биологического вида представлена информация о геноме — число хромосом, длина хромосом, и другая информация, которая используется в популяционной генетике — скорость мутации, вероятности рекомбинации, карты рекомбинации. Для многих видов представлены демографические истории, ранее полученные в опубликованных исследованиях.

Библиотека позволяет легко проводить симуляции для целого ряда организмов. *Stdpopsim* имеет интерфейс прикладного программирования (API) на языке Python и удобный интерфейс командной строки, что позволяет пользователям с минимальным опытом программирования использовать эту библиотеку.

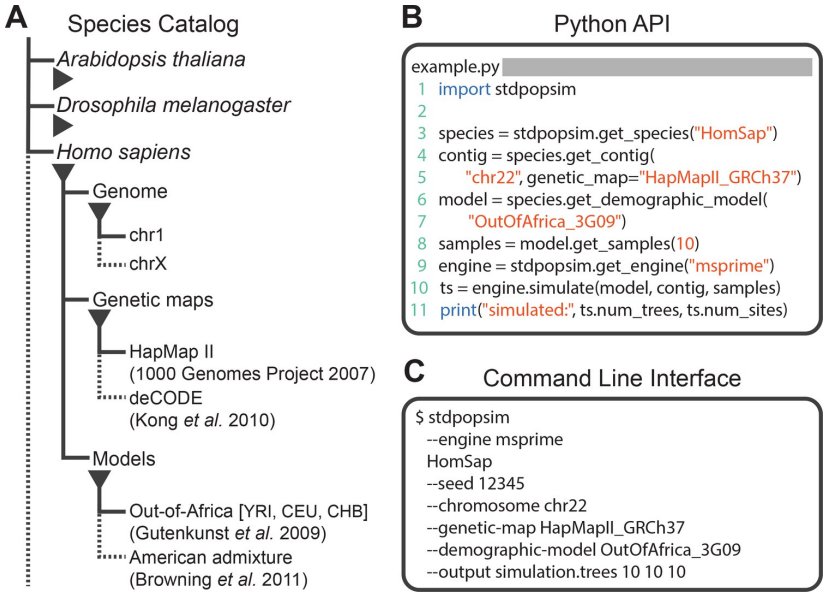


Рисунок 119 – Иерархическая структура каталога библиотеки *stdpopsim*, интерфейс прикладного программирования (API) и интерфейс командной строки. Источник: [6]

Симуляции выполняются с применением одного из двух методов: *msprime* [149], SLiM [150]. Пользователю достаточно выбрать метод симуляции, вид организмов, демографическую историю и число образцов и получить симулированные генетические данные.

Библиотека имеет открытый исходный код, доступный по адресу `https://github.com/popsim-consortium/stdpopsim` и общедоступную документацию: `https://popsim-consortium.github.io/stdpopsim-docs`. Разработка ведется широкой группой разработчиков-исследователей с использованием веб-сервиса GitHub с непрерывной интеграцией. На момент 2023 года число участников проекта насчитывает больше 50. При расширении каталога библиотеки используется система двойной проверки или контроля качества: сначала один участник проекта добавляет объект — биологический вид или демографическую историю, затем другой участник выполняет добавление того же объекта независимо (quality control). Автоматическая система сравнивает оба объекта и, в случае их совпадения, они добавляются в кодовую базу каталога библиотеки. Такой подход позволяет выполнять контроль качества и избегать ошибок разработки.

Автором диссертации был внесен следующий вклад в разработку и расширение библиотеки *stdpopsim*:

– добавление биологического вида *Heliconius melpomene* в каталог (контроль качества): https://github.com/popsim-consortium/stdpopsim/pull/1165;
– добавление демографической истории PapuansOutOfAfrica_10J19 десяти популяций современного человека для биологического вида *Homo Sapiens* в каталог (контроль качества): https://github.com/popsim-consortium/stdpopsim/pull/387;
– добавление демографической истории African3Epoch_1H18 для биологического вида *Arabidopsis thaliana* в каталог: https://github.com/popsim-consortium/stdpopsim/pull/270;
– тестирование библиотеки и выявление дефектов, публикация описания дефектов в системе отслеживания ошибок: https://github.com/popsim-consortium/stdpopsim/issues/701;
– добавление документации: https://github.com/popsim-consortium/stdpopsim/pull/333.

Библиотека *stdpopsim* была применена для симулирования данных при проведении экспериментальных исследований разработанного метода настройки параметров моделей на основе комбинации генетического алгоритма и локального поиска, которые представлены в разделе 2.4.5.

4.2.2. Расширение библиотеки *demes* для текстового и визуального представления демографических историй

Библиотека *demes* позволяет построить и использовать текстовое и визуальное представление демографических историй. Библиотека также была разработана сообществом PopSim, как и библиотека *stdpopsim*. В проекте по разработке принимали участие семь участников. Тестовое представление реализовано в широко используемом формате YAML [171], который является языком сериализации данных, обеспечивающим хороший баланс между человеческой и машинной читабельностью. Спецификация гарантирует отсутствие двусмысленности интерпретации. Общедоступная документация включает в себя обширный набор тестовых примеров и их ожидаемый результат. Пример текстового и соответствующего визуального представления для демографической истории представлено на рисунке 120.

Библиотека имеет открытый исходный код, доступный по адресу https://github.com/popsim-consortium/demes-python и общедоступную документацию: https://popsim-consortium.github.io/demes-docs.

Автором диссертации был внесен следующий вклад в разработку и расширение библиотеки *demes*:

– добавление линейной функции изменения численности популяций;
– разработка части программного кода библиотеки (5 %);
– интеграция библиотеки *demes* в программный комплекс GADMA.

```yaml
# Comments start with a hash.
description:
  Two-deme isolation-with-migration model.
time_units: generations
defaults:
  epoch:
    start_size: 1000
demes:
  - name: A
    description: The ancestral deme
    epochs:
      - end_time: 100
  - name: X
    description: First descendant deme.
    ancestors: [A]
  - name: Y
    description: Second descendant deme.
    ancestors: [A]
    epochs:
      - end_time: 50
      - end_size: 3000
migrations:
  - demes: [X, Y]
    rate: 1e-4
```

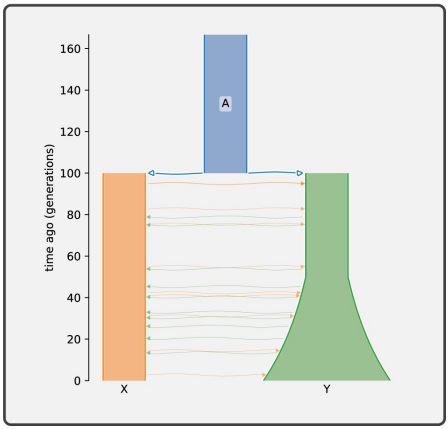


Рисунок 120 – Пример тестового и визуального представления демографической истории, полученных с применением *demes*. Источник: [8]

Выводы по граве 4

1. Описан разработанный программный комплекс GADMA для вывода демографической истории популяций, реализующий разработанные модели и методы.
2. Программный комплекс имеет репозиторий с открытым исходным кодом, доступным по адресу `https://github.com/ctlab/GADMA`, общедоступную документацию и систему автоматического тестирования программного кода.
3. Каталог доступных биологических видов для симуляции данных в библиотеке *stdpopsim* был расширен. Библиотека была протестирована и использована при проведении экспериментальных исследований в данной работе.
4. Библиотека *demes* позволяет построить текстовое и визуальное представление демографических историй. Библиотека была расширена добавлением линейной динамики изменения численности популяций и была интегрирована в программный комплекс GADMA.
5. Все визуальные представления демографических историй, представленные в данной работе, получены с применением библиотеки *demes*.

Заключение

Основные результаты работы состоят в следующем:

- проведено исследование текущего состояния предметной области, уточнение задачи и способов оценки результатов;
- формализована постановка задачи построения и настройки моделей метрических деревьев с функциями на ребрах на примере задачи вывода демографической истории популяций по генетическим данным;
- разработан метод автоматической настройки параметров моделей метрических деревьев с функциями на ребрах на основе комбинации методов глобальной и локальной оптимизации на примере задачи вывода демографической истории популяций по генетическим данным;
- разработан метод автоматического перебора моделей метрических деревьев с функциями на ребрах на примере задачи вывода демографической истории популяций по генетическим данным;
- спроектирован и реализован программный комплекс, включающий разработанные модели и методы для вывода демографической истории популяций по генетическим данным;
- проведены экспериментальные исследования, подтверждающие эффективность разработанных моделей и методов, а также их применимость для вывода демографической истории популяций по генетическим данным, проведен анализ результатов экспериментов.

Для оценки качества настройки моделей демографических историй в данной работе было использовано значение функции правдоподобия. Результаты экспериментов показывают, что метод настройки параметров моделей на основе комбинации генетического алгоритма и локального поиска позволил в 88% случаев (37 моделей из 42 протестированных) найти параметры модели, обеспечивающие лучшее значение правдоподобия, чем параметры, найденные существующими ранее методами. На симулированных данных разработанный метод позволил найти решения, которые на 97% ближе к оптимуму в случае одной популяции и на 66% ближе к оптимуму в случае трех популяций, чем решения, полученные существующими методами. Настройка гиперпараметров генетического алгоритма позволила ускорить реализацию в среднем на 10% с сохранением эффективности метода.

Была подтверждена эффективность метода настройки параметров моделей на основе байесовской оптимизации и локальной оптимизации в условиях сложновычислимой целевой функции. Разработанный метод позволил найти значения параметров, обеспечивающих лучшее значение правдоподобия, чем существующие методы, для двух ранее проанализированных данных четырех и пяти популяций. Было показано, что байесовская оптимизация достигает решения, близкого к оптимуму, на 50-80% быстрее, чем генетический алгоритм, в случае вывода демографической истории четырех и пяти популяций.

Метод автоматического перебора моделей позволяет автоматически строить и настраивать модели в заданных ограничениях на конфигурацию. Сравнение моделей демографических историй с разным числом параметров было осуществлено с использованием информационного критерия Акаике (AIC). Экспериментальные исследования показали, что в трех из четырех случаях метод позволил найти модель, обеспечивающую лучшее значение AIC, чем было получено ранее ручным перебором. В четвертом случае, полученная модель позволила установить излишние параметры в конфигурации и построить вложенную модель, которая в итоге обеспечила наилучшее значение AIC для данных.

В качестве перспективных направлений исследования можно выделить совершенствование метода автоматического перебора моделей с целью поиска оптимального набора параметров конфигурации, а также разработку методов настройки моделей метрического дерева с функциями на ребрах, которые позволяют осуществлять настройку не только функциональных параметров, но и поиск оптимальной структуры дерева.

Список литературы

1. ***Noskova E.***, *Ulyantsev V.*, *Koepfli K.-P.*, *O'Brien S. J.*, *Dobrynin P.* GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data // GigaScience. — 2020. — Vol. 9, no. 3. — giaa005. — DOI: 10.1093/gigascience/giaa005.
2. ***Noskova E.***, *Abramov N.*, *Iliutkin S.*, *Sidorin A.*, *Dobrynin P.*, *Ulyantsev V.* GADMA2: more efficient and flexible demographic inference from genetic data // GigaScience. — 2023. — Vol. 12. — giad059. — DOI: 10.1093/gigascience/giad059.
3. ***Noskova E.***, *Borovitskiy V.* Bayesian optimization for demographic inference // G3, Genes | Genomes | Genetics. — 2023. — Vol. 13, no. 7. — DOI: 10.1093/g3journal/jkad080. — jkad080.
4. *Zhernakova D. V.*, ..., *Ulyantsev V.*, ***Noskova E.***, ..., *O'Brien S. J.* Genome-wide sequence analyses of ethnic populations across Russia // Genomics. — 2020. — Vol. 112, no. 1. — Pp. 442–458. — DOI: 10.1016/j.ygeno.2019.03.007.
5. *Nikolic N.*, *Devloo-Delva F.*, *Bailleul D.*, ***Noskova E.***, ..., *Arnaud-Haond S.* Stepping up to genome scan allows stock differentiation in the worldwide distributed blue shark Prionace glauca // Molecular Ecology. — 2023. — Vol. 32, no. 5. — Pp. 1000–1019. — DOI: 10.1111/mec.16822.
6. *Adrion J. R.*, ..., ***Noskova E.***, ..., *Kern A. D.* A community-maintained standard library of population genetic models // eLife. — 2020. — Vol. 9. — e54967. — DOI: 10.7554/eLife.54967.
7. *Lauterbur M. E.*, ..., ***Noskova E.***, ..., *Gronau I.* Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations // eLife / ed. by Z. Gao, M. Przeworski. — 2023. — June. — Vol. 12. — DOI: 10.7554/eLife.84874.
8. *Gower G.*, *Ragsdale A. P.*, *Bisschop G.*, *Gutenkunst R. N.*, *Hartfield M.*, ***Noskova E.***, *Schiffels S.*, *Struck T. J.*, *Kelleher J.*, *Thornton K. R.* Demes: a standard format for demographic models // Genetics. — 2022. — Vol. 222, no. 3. — DOI: 10.1093/genetics/iyac131. — iyac131.
9. *Кириллов А. Н.* Динамические системы с переменной структурой и размерностью // Известия высших учебных заведений. Приборостроение. — 2009. — Т. 52, № 3. — С. 23–28.
10. *Aldous D.* The continuum random tree III // The annals of probability. — 1993. — Pp. 248–289.
11. *Berkolaiko G.*, *Kuchment P.* Introduction to quantum graphs. — American Mathematical Soc., 2013.
12. *Kottos T.*, *Smilansky U.* Quantum chaos on graphs // Physical review letters. — 1997. — Vol. 79, no. 24. — P. 4794.

13. *Exner P.*, *Kovařík H.* Quantum waveguides. — Springer, 2015.

14. *Kuchment P.*, *Kunyansky L.* Differential operators on graphs and photonic crystals // Advances in Computational Mathematics. — 2002. — Vol. 16. — Pp. 263–290.

15. *Goebel T.*, *Waters M. R.*, *O'Rourke D. H.* The late Pleistocene dispersal of modern humans in the Americas // Science. — 2008. — Vol. 319, no. 5869. — Pp. 1497–1502.

16. *Mellars P.* Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia // Science. — 2006. — Vol. 313, no. 5788. — Pp. 796–800.

17. *Nielsen R.*, *Hellmann I.*, *Hubisz M.*, *Bustamante C.*, *Clark A. G.* Recent and ongoing selection in the human genome // Nature Reviews Genetics. — 2007. — Vol. 8, no. 11. — Pp. 857–868.

18. *Райгородский А.* Модели случайных графов и их применения // Труды Московского физико-технического института. — 2010. — Т. 2, № 4. — С. 130–140.

19. *Райгородский А.* Модели случайных графов. — Litres, 2022.

20. *Ben-Gal I.* Bayesian networks // Encyclopedia of statistics in quality and reliability. — 2008.

21. *Gruber A.*, *Ben-Gal I.* Efficient Bayesian network learning for system optimization in reliability engineering // Quality Technology & Quantitative Management. — 2012. — Vol. 9, no. 1. — Pp. 97–114.

22. *Gruber A.*, *Ben-Gal I.* A targeted Bayesian network learning for classification // Quality Technology & Quantitative Management. — 2019. — Vol. 16, no. 3. — Pp. 243–261.

23. *Ben-Gal I.*, *Shani A.*, *Gohr A.*, *Grau J.*, *Arviv S.*, *Shmilovici A.*, *Posch S.*, *Grosse I.* Identification of transcription factor binding sites with variable-order Bayesian networks // Bioinformatics. — 2005. — Vol. 21, no. 11. — Pp. 2657–2666.

24. *Clark L. A.*, *Pregibon D.* Tree-based models // Statistical models in S. — Routledge, 2017. — Pp. 377–419.

25. *Kotsiantis S. B.* Decision trees: a recent overview // Artificial Intelligence Review. — 2013. — Vol. 39. — Pp. 261–283.

26. *Болтянский В. Г.*, *Солтан П. С.* Комбинаторная геометрия и классы выпуклости // Успехи математических наук. — 1978. — Т. 33, 1 (199. — С. 3–42.

27. *П. С. Солтан Д. К. Замбицкий К. Ф. П.* Экстремальные задачи на графах и алгоритмы их решения // Акад. наук Молд. ССР, Ин-т математики, Вычисл. центр. — 1973.

28. *Dress A. W.* Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: a note on combinatorial properties of metric spaces // Advances in Mathematics. — 1984. — Vol. 53, no. 3. — Pp. 321–402.

29. *Buneman P.* A note on the metric properties of trees // J. Combin. Theory Ser. B. — 1974. — Vol. 17, no. 1. — Pp. 48–50.

30. *Aldous D.* The Continuum Random Tree. I // The Annals of Probability. — 1991. — Vol. 19, no. 1. — Pp. 1–28. — DOI: 10.1214/aop/1176990534. — URL: https://doi.org/10.1214/aop/1176990534.

31. *Aldous D.* The continuum random tree. II. An overview // Stochastic analysis. — 1991. — Vol. 167. — Pp. 23–70.

32. *Матвеев С.*, *Матвеев А. С.*, *Розенберг И. Н.*, *Уманский В.* [и др.]. Создание координатных моделей железнодорожного пути в виде взвешенных метрических графов // Известия высших учебных заведений. Северо-Кавказский регион. Технические науки. — 2010. — № 5. — С. 7–11.

33. *Лёвин Б.*, *Матвеев С.*, *Матвеев А.*, *Розенберг И.*, *Уманский В.* Системы интеллектуальной навигации и графы // Открытое образование. — 2011. — № 2–2. — С. 67–69.

34. *Матвеев С.* Интеллектуальная навигация: ГЛОНАСС и координатные модели // Мир транспорта. — 2013. — № 4. — С. 20–27.

35. *Fahrmeir L.*, *Kneib T.*, *Lang S.*, *Marx B.*, *Fahrmeir L.*, *Kneib T.*, *Lang S.*, *Marx B.* Regression models. — Springer, 2013.

36. *Snee R. D.* Validation of regression models: methods and examples // Technometrics. — 1977. — Vol. 19, no. 4. — Pp. 415–428.

37. *Schiffels S.*, *Wang K.* MSMC and MSMC2: the multiple sequentially markovian coalescent // Statistical population genomics. — Humana, 2020. — Pp. 147–165.

38. *Dai L.* Nonlinear dynamics of piecewise constant systems and implementation of piecewise constant arguments. — World Scientific, 2008.

39. *Leenaerts D.*, *Van Bokhoven W. M.* Piecewise linear modeling and analysis. — Springer Science & Business Media, 2013.

40. *Friedman M.* Piecewise exponential models for survival data with covariates // The Annals of Statistics. — 1982. — Vol. 10, no. 1. — Pp. 101–113.

41. *Muggeo V. M.* Selecting number of breakpoints in segmented regression: implementation in the R package segmented // Technical report. — 2020.

42. *Malash G. F.*, *El-Khaiary M. I.* Piecewise linear regression: A statistical method for the analysis of experimental adsorption data by the intraparticle-diffusion models // Chemical Engineering Journal. — 2010. — Vol. 163, no. 3. — Pp. 256–263.

43. *Akaike H.* A new look at the statistical model identification // IEEE Transactions on Automatic Control. — 1974. — Vol. 19, no. 6. — Pp. 716–723.

44. *Berkolaiko G.* Quantum Graphs and Their Applications: Proceedings of an AMS-IMS-SIAM Joint Summer Research Conference on Quantum Graphs and Their Applications, June 19-23, 2005, Snowbird, Utah. Vol. 415. — American Mathematical Soc., 2006.

45. *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. — 2009. — Vol. 5, no. 10. — e1000695.

46. *Kamm J.*, *Terhorst J.*, *Durbin R.*, *Song Y. S.* Efficiently inferring the demographic history of many populations with allele count data // Journal of the American Statistical Association. — 2020. — Vol. 115, no. 531. — Pp. 1472–1487.

47. *Ragsdale A. P.*, *Gravel S.* Models of archaic admixture and recent history from two-locus statistics // PLoS genetics. — 2019. — Vol. 15, no. 6. — e1008204.

48. *Ragsdale A. P.*, *Gravel S.* Unbiased estimation of linkage disequilibrium from unphased data // Molecular Biology and Evolution. — 2020. — Vol. 37, no. 3. — Pp. 923–932.

49. *Portik D. M.*, *Leaché A. D.*, *Rivera D.*, *Barej M. F.*, *Burger M.*, *Hirschfeld M.*, *Rödel M.*, *Blackburn D. C.*, *Fujita M. K.* Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection // Molecular ecology. — 2017. — Vol. 26, no. 19. — Pp. 5245–5263.

50. *Leaché A. D.* [et al.]. Exploring rain forest diversification using demographic model testing in the African foam-nest treefrog Chiromantis rufescens // Journal of Biogeography. — 2019. — Vol. 46, no. 12. — Pp. 2706–2721.

51. *Blischak P. D.*, *Barker M. S.*, *Gutenkunst R. N.* Inferring the demographic history of inbred species from genome-wide SNP frequency data // Molecular biology and evolution. — 2020. — Vol. 37, no. 7. — Pp. 2124–2136.

52. *Поляк Б. Т.* Введение в оптимизацию. — 1983.

53. *Пантелеев А. В.*, *Метлицкая Д. В.*, *Алешина Е. А.* Методы глобальной оптимизации: метаэвристические стратегии и алгоритмы. — Andrey Panteleev, 2013.

54. *Rippe J. P.*, *Dixon G.*, *Fuller Z. L.*, *Liao Y.*, *Matz M.* Environmental specialization and cryptic genetic divergence in two massive coral species from the Florida Keys Reef Tract // Molecular Ecology. — 2021. — Vol. 30, no. 14. — Pp. 3468–3484.

55. *Jouganous J.*, *Long W.*, *Ragsdale A. P.*, *Gravel S.* Inferring the joint demographic history of multiple populations: beyond the diffusion approximation // Genetics. — 2017. — Vol. 206, no. 3. — Pp. 1549–1567.

56. *Nielsen R.*, *Akey J. M.*, *Jakobsson M.*, *Pritchard J. K.*, *Tishkoff S.*, *Willerslev E.* Tracing the peopling of the world through genomics // Nature. — 2017. — Vol. 541, no. 7637. — Pp. 302–310.

57. *Veríssimo A.*, *Sampaio Í.*, *McDowell J. R.*, *Alexandrino P.*, *Mucientes G.*, *Queiroz N.*, *Silva C. da*, *Jones C. S.*, *Noble L. R.* World without borders—genetic population structure of a highly migratory marine predator, the blue shark (Prionace glauca) // Ecology and Evolution. — 2017. — Vol. 7, no. 13. — Pp. 4768–4781.

58. *King J.*, *Wetklo M.*, *Supernault J.*, *Taguchi M.*, *Yokawa K.*, *Sosa-Nishizaki O.*, *Withler R.* Genetic analysis of stock structure of blue shark (Prionace glauca) in the north Pacific ocean // Fisheries Research. — 2015. — Vol. 172. — Pp. 181–189.

59. *Сивцева Т. М.*, *Осаковский В. Л.* ГЕНОМ ЯКУТСКОГО ЭТНОСА // Наука и техника в Якутии. — 2020. — Т. 1, № 38. — С. 7–11.

60. *Schwarz G.* Estimating the dimension of a model // The Annals of Statistics. — 1978. — Vol. 6, no. 2. — Pp. 461–464.

61. *Vuong Q. H.* Likelihood ratio tests for model selection and non-nested hypotheses // Econometrica: journal of the Econometric Society. — 1989. — Pp. 307–333.

62. *Broyden C. G.* The convergence of a class of double-rank minimization algorithms: 2. The new algorithm // IMA Journal of Applied Mathematics. — 1970. — Vol. 6, no. 3. — Pp. 222–231.

63. *Fletcher R.* A new approach to variable metric algorithms // The Computer Journal. — 1970. — Vol. 13, no. 3. — Pp. 317–322.

64. *Goldfarb D.* A family of variable-metric methods derived by variational means // Mathematics of Computation. — 1970. — Vol. 24, no. 109. — Pp. 23–26.

65. *Shanno D. F.* Conditioning of quasi-Newton methods for function minimization // Mathematics of Computation. — 1970. — Vol. 24, no. 111. — Pp. 647–656.

66. *Nelder J. A.*, *Mead R.* A simplex method for function minimization // The Computer Journal. — 1965. — Vol. 7, no. 4. — Pp. 308–313.

67. *Powell M. J.* An efficient method for finding the minimum of a function of several variables without calculating derivatives // The Computer Journal. — 1964. — Vol. 7, no. 2. — Pp. 155–162.

68. *Gao X.*, *Song P. X.-K.* Composite likelihood Bayesian information criteria for model selection in high-dimensional data // Journal of the American Statistical Association. — 2010. — Vol. 105, no. 492. — Pp. 1531–1540.

69. *Coffman A. J.*, *Hsieh P. H.*, *Gravel S.*, *Gutenkunst R. N.* Computationally efficient composite likelihood statistics for demographic inference // Molecular biology and evolution. — 2016. — Vol. 33, no. 2. — Pp. 591–593.

70. *Schraiber J. G.*, *Akey J. M.* Methods and models for unravelling human evolutionary history // Nature Reviews Genetics. — 2015. — Vol. 16, no. 12. — Pp. 727–740.

71. *Leduc G.*, *Schneider R.*, *Kim J.-H.*, *Lohmann G.* Holocene and Eemian sea surface temperature trends as revealed by alkenone and Mg/Ca paleothermometry // Quaternary Science Reviews. — 2010. — Vol. 29, no. 7/8. — Pp. 989–1004.

72. *Masson-Delmotte V.*, *Schulz M.*, *Abe-Ouchi A.*, *Beer J.*, *Ganopolski A.*, *Gonzáles Rouco J.*, *Jansen E.*, *Lambeck K.*, *Luterbacher J.*, *Naish T.*, [et al.]. Information from paleoclimate archives. — 2013.

73. *Olsen J.*, *Anderson N. J.*, *Knudsen M. F.* Variability of the North Atlantic Oscillation over the past 5,200 years // Nature Geoscience. — 2012. — Vol. 5, no. 11. — Pp. 808–812.

74. *Nabulsi M. M.*, *Tamim H.*, *Sabbagh M.*, *Obeid M. Y.*, *Yunis K. A.*, *Bitar F. F.* Parental consanguinity and congenital heart malformations in a developing country // American journal of medical genetics Part A. — 2003. — Vol. 116, no. 4. — Pp. 342–347.

75. *Wright S.* Coefficients of inbreeding and relationship // The American Naturalist. — 1922. — Vol. 56, no. 645. — Pp. 330–338.

76. *Kimura M.* On the probability of fixation of mutant genes in a population // Genetics. — 1962. — Vol. 47, no. 6. — P. 713.

77. *Kimura M.* Diffusion models in population genetics // Journal of Applied Probability. — 1964. — Vol. 1, no. 2. — Pp. 177–232.

78. *Ohta T.*, *Kimura M.* Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation // Genetics. — 1969. — Vol. 63, no. 1. — P. 229.

79. *Hill W. G.*, *Robertson A.* The effect of linkage on limits to artificial selection // Genetics Research. — 1966. — Vol. 8, no. 3. — Pp. 269–294.

80. *Hill W. G.*, *Robertson A.* Linkage disequilibrium in finite populations // Theoretical and applied genetics. — 1968. — Vol. 38, no. 6. — Pp. 226–231.

81. *Kuhner M. K.*, *Yamato J.*, *Felsenstein J.* Maximum likelihood estimation of population growth rates based on the coalescent // Genetics. — 1998. — Vol. 149, no. 1. — Pp. 429–434.

82. *Kamm J. A.*, *Terhorst J.*, *Song Y. S.* Efficient computation of the joint sample frequency spectra for multiple populations // Journal of Computational and Graphical Statistics. — 2017. — Vol. 26, no. 1. — Pp. 182–194.

83. *Steinrücken M.*, *Kamm J.*, *Spence J. P.*, *Song Y. S.* Inference of complex population histories using whole-genome sequences from multiple populations // Proceedings of the National Academy of Sciences. — 2019. — Vol. 116, no. 34. — Pp. 17115–17120.

84. *Excoffier L.*, *Marchi N.*, *Marques D. A.*, *Matthey-Doret R.*, *Gouy A.*, *Sousa V. C.* fastsimcoal2: demographic inference under complex evolutionary scenarios // Bioinformatics. — 2021.

85. *DeWitt W. S.*, *Harris K. D.*, *Ragsdale A. P.*, *Harris K.* Nonparametric coalescent inference of mutation spectrum history and demography // Proceedings of the National Academy of Sciences. — 2021. — Vol. 118, no. 21.

86. *Chang J.*, *Cooper G.* A practical difference scheme for Fokker-Planck equations // Journal of Computational Physics. — 1970. — Vol. 6, no. 1. — Pp. 1–16.

87. *Byrd R. H.*, *Lu P.*, *Nocedal J.*, *Zhu C.* A limited memory algorithm for bound constrained optimization // SIAM Journal on scientific computing. — 1995. — Vol. 16, no. 5. — Pp. 1190–1208.

88. *Powell M. J.* The BOBYQA algorithm for bound constrained optimization without derivatives // Cambridge NA Report NA2009/06, University of Cambridge, Cambridge. — 2009. — Vol. 26.

89. *Gravel S.*, *National Heart L.*, *Project B. I. (G. E. S.* Predicting discovery rates of genomic features // Genetics. — 2014. — Vol. 197, no. 2. — Pp. 601–610.

90. *Baolin Z.*, *Wenzhi L.* On alternating segment Crank-Nicolson scheme // Parallel Computing. — 1994. — Vol. 20, no. 6. — Pp. 897–902.

91. *Virtanen P.* [et al.]. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python // Nature Methods. — 2020. — Vol. 17. — Pp. 261–272. — DOI: 10.1038/s41592-019-0686-2.

92. *Beichman A. C.*, *Huerta-Sanchez E.*, *Lohmueller K. E.* Using genomic data to infer historic population dynamics of nonmodel organisms // Annual Review of Ecology, Evolution, and Systematics. — 2018. — Vol. 49. — Pp. 433–456.

93. *Fisher R. A.* XVII.—The distribution of gene ratios for rare mutations // Proceedings of the Royal Society of Edinburgh. — 1931. — Vol. 50. — Pp. 204–219.

94. *Watterson G.* The effect of linkage in a finite random-mating population // Theoretical Population Biology. — 1970. — Vol. 1, no. 1. — Pp. 72–87.

95. *Ragsdale A. P.*, *Gutenkunst R. N.* Inferring demographic history using two-locus statistics // Genetics. — 2017. — genetics–117.

96. *Fisher R. A.* XXI.—On the dominance ratio // Proceedings of the royal society of Edinburgh. — 1923. — Vol. 42. — Pp. 321–341.

97. *Wright S.* Evolution in Mendelian populations // Genetics. — 1931. — Vol. 16, no. 2. — P. 97.

98. *Kimura M.* Stochastic processes and distribution of gene frequencies under natural selection. — Citeseer, 1954.

99. *Wakeley J. H.* Coalescent theory: an introduction // Science. — 2009.

100. *Durrett R.*, *Durrett R.* Probability models for DNA sequence evolution. Vol. 2. — Springer, 2008.

101. *Moran P. A. P.* Random processes in genetics // Mathematical Proceedings of the Cambridge Philosophical Society. Vol. 54. — Cambridge University Press. 1958. — Pp. 60–71.

102. *Kingman J. F. C.* The coalescent // Stochastic processes and their applications. — 1982. — Vol. 13, no. 3. — Pp. 235–248.

103. *Watterson G.* On the number of segregating sites in genetical models without recombination // Theoretical population biology. — 1975. — Vol. 7, no. 2. — Pp. 256–276.

104. *Peaceman D. W.*, *Rachford Jr H. H.* The numerical solution of parabolic and elliptic differential equations // Journal of the Society for industrial and Applied Mathematics. — 1955. — Vol. 3, no. 1. — Pp. 28–41.

105. *Press W. H.* Numerical recipes 3rd edition: The art of scientific computing. — Cambridge university press, 2007.

106. *Березин И. С.*, *Жидков Н. П.* Методы вычислений. — 1962.

107. *Демидович Б. П.*, *Марон И. А.* Основы вычислительной математики. — Государственное издательство физико-математической литературы, 1963.

108. *Ziegel E.* Numerical recipes: The art of scientific computing. — 1987.

109. *Gutenkunst R. N.* dadi.CUDA: Accelerating population genetics inference with graphics processing units // Molecular biology and evolution. — 2021. — Vol. 38, no. 5. — Pp. 2177–2178.

110. *Crank J.*, *Nicolson P.* A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type // Mathematical proceedings of the Cambridge philosophical society. Vol. 43. — Cambridge University Press. 1947. — Pp. 50–67.

111. *Rosenbrock H.* An automatic method for finding the greatest or least value of a function // The computer journal. — 1960. — Vol. 3, no. 3. — Pp. 175–184.

112. *Johnson S. G.* The NLopt nonlinear-optimization package. —. — URL: http://github.com/stevengj/nlopt.

113. *Horowitz J. L.* The bootstrap // Handbook of econometrics. Vol. 5. — Elsevier, 2001. — Pp. 3159–3228.

114. *Efron B.*, *Hinkley D. V.* Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information // Biometrika. — 1978. — Vol. 65, no. 3. — Pp. 457–483.

115. *Cattelan M.*, *Sartori N.* Empirical and simulated adjustments of composite likelihood ratio statistics // Journal of Statistical Computation and Simulation. — 2016. — Vol. 86, no. 5. — Pp. 1056–1067.

116. *Varin C.*, *Vidoni P.* A note on composite likelihood inference and model selection // Biometrika. — 2005. — Vol. 92, no. 3. — Pp. 519–528.

117. *Rotnitzky A.*, *Jewell N. P.* Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data // Biometrika. — 1990. — Vol. 77, no. 3. — Pp. 485–497.

118. *Holland J.* Adaptation in natural and artificial systems: an introductory analysis with application to biology // Control and Artificial Intelligence. — 1975.

119. *Chowdhury B.*, *Garai A.*, *Garai G.* An optimized approach for annotation of large eukaryotic genomic sequences using genetic algorithm // BMC bioinformatics. — 2017. — Vol. 18, no. 1. — Pp. 1–13.

120. *Chowdhury B.*, *Garai G.* A review on multiple sequence alignment from the perspective of genetic algorithm // Genomics. — 2017. — Vol. 109, no. 5/6. — Pp. 419–431.

121. *Unger R.* The genetic algorithm approach to protein structure prediction // Applications of Evolutionary Computation in Chemistry. — 2004. — Pp. 153–175.

122. *Spiegel J. O.*, *Durrant J. D.* AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization // Journal of cheminformatics. — 2020. — Vol. 12, no. 1. — Pp. 1–16.

123. *Yang C.-H.*, *Moi S.-H.*, *Lin Y.-D.*, *Chuang L.-Y.* Genetic algorithm combined with a local search method for identifying susceptibility genes // Journal of Artificial Intelligence and Soft Computing Research. — 2016. — Vol. 6, no. 3. — Pp. 203–212.

124. *Schumer M.*, *Steiglitz K.* Adaptive step size random search // IEEE Transactions on Automatic Control. — 1968. — Vol. 13, no. 3. — Pp. 270–276.

125. *Pukelsheim F.* The three sigma rule // The American Statistician. — 1994. — Vol. 48, no. 2. — Pp. 88–91.

126. *Hutter F.*, *Hoos H. H.*, *Leyton-Brown K.* Sequential model-based optimization for general algorithm configuration // International conference on learning and intelligent optimization. — Springer. 2011. — Pp. 507–523.

127. *Lindauer M.*, *Eggensperger K.*, *Feurer M.*, *Biedenkapp A.*, *Deng D.*, *Benjamins C.*, *Ruhkopf T.*, *Sass R.*, *Hutter F.* SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization // Journal of Machine Learning Research. — 2022. — Vol. 23, no. 54. — Pp. 1–9. — URL: http://jmlr.org/papers/v23/21-0888.html.

128. *Huber C. D.*, *Durvasula A.*, *Hancock A. M.*, *Lohmueller K. E.* Gene expression drives the evolution of dominance // Nature communications. — 2018. — Vol. 9, no. 1. — Pp. 1–11.

129. *McCoy R. C.*, *Garud N. R.*, *Kelley J. L.*, *Boggs C. L.*, *Petrov D. A.* Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population // Molecular Ecology. — 2014. — Vol. 23, no. 1. — Pp. 136–150.

130. *Kushner H. J.* A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. — 1964.

131. *Močkus J.* On Bayesian methods for seeking the extremum // Optimization techniques IFIP technical conference. — Springer. 1975. — Pp. 400–404.

132. *Mockus J.* The Bayesian approach to local optimization // Bayesian Approach to Global Optimization. — Springer, 1989. — Pp. 125–156.

133. *Rasmussen C. E.*, *Williams C. K.* Gaussian Processes for Machine Learning. — MIT Press, 2006.

134. *Hutter F.*, *Hoos H. H.*, *Leyton-Brown K.*, *Murphy K. P.* An experimental investigation of model-based parameter optimisation: SPO and beyond // Proceedings of the 11th Annual conference on Genetic and evolutionary computation. — 2009. — Pp. 271–278.

135. *Stein M.* Interpolation of spatial data: some theory for kriging. — New York : Springer Science & Business Media, 2012.

136. *Gradshteyn I. S.*, *Ryzhik I. M.* Table of Integrals, Series, and Products. — 7th ed. — Academic Press, 2014.

137. *Benoit C.* Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéaires en nombre inférieur à celui des inconnues (Procédé du Commandant Cholesky) // Bulletin géodésique. — 1924. — T. 2, n° 1. — P. 67-77.

138. *The GPyOpt authors*. GPyOpt: A Bayesian Optimization framework in python. — 2016. — http://github.com/SheffieldML/GPyOpt.

139. *Balandat M.*, *Karrer B.*, *Jiang D. R.*, *Daulton S.*, *Letham B.*, *Wilson A. G.*, *Bakshy E.* BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization // Advances in Neural Information Processing Systems 33. — 2020. — URL: http://arxiv.org/abs/1910.06403.

140. *Lindauer M.*, *Eggensperger K.*, *Feurer M.*, *Biedenkapp A.*, *Deng D.*, *Benjamins C.*, *Ruhkopf T.*, *Sass R.*, *Hutter F.* SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization // Journal of Machine Learning Research. — 2022. — Vol. 23, no. 54. — Pp. 1–9. — URL: http://jmlr.org/papers/v23/21-0888.html.

141. *Lago J.*, *De Ridder F.*, *Vrancx P.*, *De Schutter B.* Forecasting day-ahead electricity prices in Europe: The importance of considering market integration // Applied energy. — 2018. — Vol. 211. — Pp. 890–903.

142. *Hewamalage H.*, *Bergmeir C.*, *Bandara K.* Recurrent neural networks for time series forecasting: Current status and future directions // International Journal of Forecasting. — 2021. — Vol. 37, no. 1. — Pp. 388–427.

143. *Wu S.*, *Song X.*, *Feng Z.*, *Wu X.* NFLAT: Non-Flat-Lattice Transformer for Chinese Named Entity Recognition // arXiv preprint arXiv:2205.05832. — 2022.

144. *Awad N.*, *Shala G.*, *Deng D.*, *Mallik N.*, *Feurer M.*, *Eggensperger K.*, *Biedenkapp A.*, *Vermetten D.*, *Wang H.*, *Doerr C.*, [et al.]. Squirrel: A switching hyperparameter optimizer // arXiv preprint arXiv:2012.08180. — 2020.

145. *Myers S.*, *Fefferman C.*, *Patterson N.* Can one learn history from the allelic spectrum? // Theoretical Population Biology. — 2008. — Vol. 73, no. 3. — Pp. 342–348.

146. *Ochoa A.*, *Onorato D. P.*, *Fitak R. R.*, *Roelke-Parker M. E.*, *Culver M.* De novo assembly and annotation from parental and F1 puma genomes of the Florida panther genetic restoration program // G3: Genes, Genomes, Genetics. — 2019. — Vol. 9, no. 11. — Pp. 3531–3536.

147. *Cheng F.* [et al.]. Genome resequencing and comparative variome analysis in a Brassica rapa and Brassica oleracea collection. Sci Data 3: 160119. — 2016.

148. *Cheng F.*, *Sun R.*, *Hou X.*, *Zheng H.*, *Zhang F.*, *Zhang Y.*, *Liu B.*, *Liang J.*, *Zhuang M.*, *Liu Y.*, [et al.]. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea // Nature genetics. — 2016. — Vol. 48, no. 10. — Pp. 1218–1224.

149. *Kelleher J.*, *Etheridge A. M.*, *McVean G.* Efficient coalescent simulation and genealogical analysis for large sample sizes // PLoS computational biology. — 2016. — Vol. 12, no. 5. — e1004842.

150. *Haller B. C.*, *Messer P. W.* SLiM 3: forward genetic simulations beyond the Wright–Fisher model // Molecular biology and evolution. — 2019. — Vol. 36, no. 3. — Pp. 632–637.

151. *Locke D. P.*, *Hillier L. W.*, *Warren W. C.*, *Worley K. C.*, *Nazareth L. V.*, *Muzny D. M.*, *Yang S.-P.*, *Wang Z.*, *Chinwalla A. T.*, *Minx P.*, [et al.]. Comparative and demographic analysis of orang-utan genomes // Nature. — 2011. — Vol. 469, no. 7331. — Pp. 529–533.

152. *Nater A.*, *Mattle-Greminger M. P.*, *Nurcahyo A.*, *Nowak M. G.*, *De Manuel M.*, *Desai T.*, *Groves C.*, *Pybus M.*, *Sonay T. B.*, *Roos C.*, [et al.]. Morphometric, behavioral, and genomic evidence for a new orangutan species // Current Biology. — 2017. — Vol. 27, no. 22. — Pp. 3487–3498.

153. *Sudmant P. H.*, *Rausch T.*, *Gardner E. J.*, *Handsaker R. E.*, *Abyzov A.*, *Huddleston J.*, *Zhang Y.*, *Ye K.*, *Jun G.*, *Hsi-Yang Fritz M.*, [et al.]. An integrated map of structural variation in 2,504 human genomes // Nature. — 2015. — Vol. 526, no. 7571. — Pp. 75–81.

154. *The 1000 Genomes Project Consortium*. A global reference for human genetic variation // Nature. — 2015. — Vol. 526, no. 7571. — P. 68.

155. *Rosen Z.*, *Bhaskar A.*, *Roch S.*, *Song Y. S.* Geometry of the sample frequency spectrum and the perils of demographic inference // Genetics. — 2018. — Vol. 210, no. 2. — Pp. 665–682.

156. *Harris K.*, *Nielsen R.* Inferring demographic history from a spectrum of shared haplotype lengths // PLoS Genetics. — 2013. — Vol. 9, no. 6. — e1003521.

157. *Sharp R. R.*, *Barrett C. J.* The environmental genome project: ethical, legal, and social implications. // Environmental Health Perspectives. — 2000. — Vol. 108, no. 4. — P. 279.

158. *Pirog A.* Structure génétique des populations et biologie de la reproduction chez le requin bouledogue Carcharhinus leucas et le requin tigre Galeocerdo cuvier : thèse de doctorat / Pirog Agathe. — La Réunion, 2018.

159. *Duncan K.*, *Martin A.*, *Bowen B.*, *De Couet H.* Global phylogeography of the scalloped hammerhead shark (Sphyrna lewini) // Molecular ecology. — 2006. — Vol. 15, no. 8. — Pp. 2239–2251.

160. *Rougeux C.*, *Bernatchez L.*, *Gagnaire P.-A.* Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (Coregonus clupeaformis) // Genome biology and evolution. — 2017. — Vol. 9, no. 8. — Pp. 2057–2074.

161. *Poisson F.* Compilation of information on blue shark (Prionace glauca), silky shark (Carcharhinus falciformis), oceanic whitetip shark (Carcharhinus longimanus), scalloped hammerhead (Sphyrna lewini) and shortfin mako (Isurus oxyrinchus) in the Indian Ocean // 3rd Session of the IOTC Working Party on Ecosystems and Bycatch, July 11-13 2007, Victoria, Seychelles. — 2007.

162. *Cortés E.*, *Arocha F.*, *Beerkircher L.*, *Carvalho F.*, *Domingo A.*, *Heupel M.*, *Holtzhausen H.*, *Santos M. N.*, *Ribera M.*, *Simpfendorfer C.* Ecological risk assessment of pelagic sharks caught in Atlantic pelagic longline fisheries // Aquatic Living Resources. — 2010. — Vol. 23, no. 1. — Pp. 25–34.

163. *Bostock H. C.*, *Barrows T. T.*, *Carter L.*, *Chase Z.*, *Cortese G.*, *Dunbar G. B.*, *Ellwood M.*, *Hayward B.*, *Howard W.*, *Neil H.*, [et al.]. A review of the Australian–New Zealand sector of the Southern Ocean over the last 30 ka (Aus-INTIMATE project) // Quaternary Science Reviews. — 2013. — Vol. 74. — Pp. 35–57.

164. *Kaiser J.*, *Schefuß E.*, *Lamy F.*, *Mohtadi M.*, *Hebbeln D.* Glacial to Holocene changes in sea surface temperature and coastal vegetation in north central Chile: high versus low latitude forcing // Quaternary Science Reviews. — 2008. — Vol. 27, no. 21/22. — Pp. 2064–2075.

165. *Shevenell A. E.*, *Ingalls A.*, *Domack E.*, *Kelly C.* Holocene Southern Ocean surface temperature variability west of the Antarctic Peninsula // Nature. — 2011. — Vol. 470, no. 7333. — Pp. 250–254.

166. *Fujinami Y.*, *Semba Y.*, *Okamoto H.*, *Ohshimo S.*, *Tanaka S.* Reproductive biology of the blue shark (Prionace glauca) in the western North Pacific Ocean // Marine and Freshwater Research. — 2017. — Vol. 68, no. 11. — Pp. 2018–2027.

167. *Coelho R.*, *Mejuto J.*, *Domingo A.*, *Yokawa K.*, *Liu K.-M.*, *Cortés E.*, *Romanov E. V.*, *Da Silva C.*, *Hazin F.*, *Arocha F.*, [et al.]. Distribution patterns and population structure of the blue shark (Prionace glauca) in the Atlantic and Indian Oceans // Fish and Fisheries. — 2018. — Vol. 19, no. 1. — Pp. 90–106.

168. *Druon J.-N.*, *Campana S.*, *Vandeperre F.*, *Hazin F. H.*, *Bowlby H.*, *Coelho R.*, *Queiroz N.*, *Serena F.*, *Abascal F.*, *Damalas D.*, [et al.]. Global-scale environmental niche and habitat of blue shark (Prionace glauca) by size and sex: a pivotal step to improving stock management // Frontiers in Marine Science. — 2022. — Vol. 9. — P. 451.

169. *Danecek P.*, *Auton A.*, *Abecasis G.*, *Albers C. A.*, *Banks E.*, *DePristo M. A.*, *Handsaker R. E.*, *Lunter G.*, *Marth G. T.*, *Sherry S. T.*, [et al.]. The variant call format and VCFtools // Bioinformatics. — 2011. — Vol. 27, no. 15. — Pp. 2156–2158.

170. *Excoffier L.*, *Dupanloup I.*, *Huerta-Sánchez E.*, *Sousa V. C.*, *Foll M.* Robust demographic inference from genomic and SNP data // PLoS genetics. — 2013. — Vol. 9, no. 10. — e1003905.

171. *Ben-Kiki O.*, *Evans C.*, *Ingerson B.* Yaml ain't markup language (yaml™) version 1.1 // Working Draft 2008. — 2009. — Vol. 5. — P. 11.

Список иллюстраций

P.1 Примеры визуального представления демографических историй одной и двух популяций . . . 17
P.2 Пример модели демографической истории двух популяции в виде метрического дерева с функциями на ребрах . . . 18
P.3 Пример входа и выхода существующих программных решений для вывода демографической истории популяций по генетическим данным . . . 18
P.4 Пример задания модели первого класса с использованием интерфейса библиотеки *∂a∂*i . . . 19
P.5 Примеры работы методов локальной оптимизации при поиске максимума функции, изображенной на рисунке (а): (б) метод BFGS, (в) метод Нелдера-Мида, (г) метод Пауэлла. . . . 20
P.6 Пример расширенной модели демографической истории двух популяций и демографические истории при разных значениях параметра `Dyn` . . . 21
P.7 Результаты настроенных моделей для метода вычисления правдоподобия, реализованного в *momi2* (а) модель без миграций, и модели с (б) одной, (в) тремя, (г) семью единичными миграциями 24
P.8 Полученная демографическая история трех популяций современного человека . . . 24
P.9 Сходимость по времени методов настройки параметров моделей для (а) двух популяций, (б) пяти популяций . . . 25
P.10 Сравнение демографической истории, полученной разработанным методом (ВО), и демографической истории из [55] . . . 26
P.11 Демографическая история, полученная разработанным методом . 26
P.12 Демографическая история трех популяций голубой акулы . . . 27
P.13 Структура программного комплекса GADMA . . . 28
S.1 Examples of visual representations of the demographic histories of one and two populations . . . 38
S.2 Example of the model of demographic history of two populations in the form of a metric tree with functions on edges . . . 39
S.3 Example input and output of existing software solutions for the demographic history inference from genetic data . . . 39
S.4 Example specification of the first class model using the *∂a∂*i library interface . . . 40
S.5 Examples of local optimization methods applied to maximize the function shown in panel (a): (b) BFGS method, (c) Nelder-Mead method, (d) Powell method. . . . 41
S.6 Example of an extended demographic model for two populations and the corresponding demographic histories for different values of the parameter `Dyn` . . . 42

S.7 Results of tuned models for the likelihood computation method implemented in *momi2* (a) model without migrations, and models with (b) one, (c) three, (d) seven pulse migrations 44
S.8 Obtained demographic history of three populations of modern humans 45
S.9 Convergence over time of parameter tuning methods for (a) two populations, (b) five populations. 45
S.10 Comparison of the demographic history obtained using the developed method (BO) and the demographic history from [55]. 46
S.11 Demographic history obtained by the developed method 47
S.12 Demographic history of three blue shark populations 47
S.13 Structure of the GADMA software framework. 48
1 Пример двух популяций газели вида *Dama gazelle*; 1) популяция mhorr, 2) популяция addra 60
2 Примеры визуального представления демографических историй одной и двух популяций 62
3 Пример демографической истории трех популяций современного человека и карта перемещения этих популяций, построенная по археологическим данным. 63
4 Пример дерева разделений популяций 64
5 Пример демографической истории популяций как дерева разделений с заданными временами и функциями численности, набора миграций и дополнительных констант 65
6 Пример модели M_1 демографической истории одной популяции с одним параметром 67
7 Пример модели M_2 демографической истории одной популяции с четырьмя параметрами 68
8 Пример вложенных моделей: модель M_1 вложена в модель M_2 . . 68
9 Метрические деревья с функциями на ребрах как модели демографических историй а) одной популяции, б) двух популяций . . 69
10 Общая схема поиска демографической истории популяций с использованием параметрической модели 70
11 Пример входа и выхода существующих программных решений для вывода демографической истории популяций по генетическим данным 71
12 Общая схема существующих методов вычисления значения правдоподобия 73
13 Пример модели $M = \langle \Theta, \mathcal{E}, \mathfrak{F} \rangle$ первого класса 77
14 Модель демографической истории с параметрами и демографические истории при разных значениях параметров 78
15 Пример задания модели демографической истории с использованием интерфейса библиотеки *∂a∂i* 79
16 Пример задания модели демографической истории с использованием интерфейса библиотеки *moments* 79

17 Пример задания модели демографической истории с использованием интерфейса библиотеки *momentsLD* 80
18 Пример модели $M = \langle \Theta, \mathcal{E}, \mathfrak{F} >$ второго класса 82
19 Модель демографической истории с параметрами и демографические истории при разных значениях параметров 83
20 Пример задания модели демографической истории с использованием интерфейса библиотеки *momi2* 84
21 Главные компоненты генетического материала клетки. 87
22 Понятие аллели как варианта участка ДНК и генотипа как совокупность аллелей организма. 88
23 Пример построения аллель-частотного спектра для двух популяций 90
24 Примеры аллель-частотного спектра для: а) одной популяции; б) двух популяций; в) трех популяций 91
25 Примеры аллель-частотного спектра двух популяций, которые соответствуют разным демографическим историям двух популяций: а) изоляция, отсутствие миграции; б) между популяциями существовала небольшая миграция; в) сильная миграция между популяциями 91
26 Пример абсолютных и относительных частот гаплотипов 92
27 Пример передачи гаплотипов между поколениями и их вероятности 93
28 Пример двухлокусного гаплотип-частотного спектра, построенного для данных 10 диплоидных особей одной популяции. 94
29 Пример кривых зависимостей значений разных статистик от генетического расстояния между локусами. 95
30 Примеры изменения частот аллелей в модели Райта-Фишера и модели Морана 96
31 Время вычисления значения правдоподобия для *дaдi*, *moments*, *momentsLD* и *momi2* 100
32 Отрицательная функция Розенброка [111] 104
33 Примеры работы методов локальной оптимизации при поиске оптимума функции, изображенной на рисунке (а): (б) метод BFGS, (в) метод Нелдера-Мида, (г) метод Пауэла. 107
34 Пример некоторых моделей из каталога *dadi pipeline*. Источник: [49] 110
35 Пример разработанной расширенной модели демографической истории с параметрами и соответствующие ей демографические истории при разных значениях параметра `Dyn` 115
36 Пример модели $M = \langle \Theta, \Theta_d, \mathcal{E}, \mathfrak{F}, \mathfrak{F}_d \rangle$ расширенного класса . . . 116
37 Структура классов разработанного модуля `variables` 117
38 Структура классов разработанного модуля `models` 118
39 Пример задания расширенной модели 119
40 Схема алгоритма метода, основанного на комбинации генетического алгоритма и локального поиска 123

41 Пример применения оператора мутации разработанного генетического алгоритма 125
42 Пример применения оператора кроссовера разработанного генетического алгоритма 127
43 Плотность нормального распределения $N(\mu, \sigma^2)$ и иллюстрация правила трех сигм 129
44 Примеры плотности усеченного нормального распределения $\hat{N}(\mu, \sigma^2, 0, 1)$ на отрезке $[0, 1]$, у которого среднеквадратичное отклонение σ выбрано в соответствии с разработанным методом . 130
45 Структура классов разработанного модуля `optimizers` 131
46 Реализация разработанного комбинированного метода с помощью модуля `optimizers` 132
47 Применение разработанного комбинированного метода на основе генетического алгоритма, реализованного с помощью модуля `optimizers`, для поиска точки минимума функции Розенброка . 133
48 Пример вывода программы, представленной на рисунке 47 134
49 Пример описание структурированного формата имени набора данных из пакета `deminf_data v1.0.0` 137
50 Сравнение значений правдоподобия, полученных с помощью новых конфигураций и исходной конфигурации генетического алгоритма 141
51 Сравнение генетического алгоритма с настроенными гиперпараметрами с исходной версией на различных наборах данных 142
52 Фрагмент байесовской оптимизации 144
53 Схема алгоритма разработанного комбинированного метода, основанного на баейсовской оптимизации и методе локального поиска 146
54 Обновленная структура классов модуля `optimizers` 150
55 Обновленная структура классов модуля `optimizers` 151
56 Применение разработанного комбинированного метода на основе байесовской оптимизации, реализованного с помощью модуля `optimizers`, для поиска точки минимума функции Розенброка . 152
57 Пример вывода программы, представленной на рисунке 56 153
58 Время вычисления логарифма правдоподобия с помощью *moments* для тестируемых наборов данных из пакета `deminf_data v1.0.0` 155
59 Примеры графиков сходимости двенадцати конфигураций классической байесовской оптимизации для датасетов двух и пяти популяций 157
60 Гистограммы частоты выбора функций ковариации при применении баесовской оптимизации с автоматическим выбором функции ковариации 158

61 Примеры графиков сходимости двух конфигураций байесовской оптимизации с автоматическим выбором функции ковариации и четырех наилучших конфигураций классической байесовской оптимизации для наборов данных трех и четырех популяций 159
62 Примеры графиков сходимости ансамблевого метода байесовской оптимизации и двух конфигураций метода с автоматическим выбором функции ковариации для наборов данных четырех и пяти популяций 160
63 Демографическая история одной популяции, которая была использована для симуляции данных и симулированные генетические данные в виде аллель-частотного спектра 162
64 Используемые модели для сравнения методов на симулированных данных одной популяции 163
65 Демографические истории, полученные путем настройки параметров модели 1 разными методами 164
66 Демографические истории, полученные путем настройки параметров модели 2 методом GA+P 165
67 Демографическая история двух популяции, которая была использована для симуляции данных и симулированные генетические данные в виде аллель-частотного спектра 165
68 Используемые модели для сравнения методов на симулированных данных двух популяций 166
69 Демографические истории двух популяций, полученные путем настройки параметров модели 1 разными методами 167
70 Демографическая история двух популяций, полученная путем настройки параметров модели 2 методом GA+P 167
71 Демографическая история трех популяции, которая была использована для симуляции данных и симулированные генетические данные в виде аллель-частотного спектра 168
72 Используемые модели для сравнения методов на симулированных данных трех популяций 168
73 Демографические истории трех популяций, полученные путем настройки параметров модели 1 разными методами 169
74 Демографическая история трех популяций, полученная путем настройки параметров модели 2 методом GA+P 170
75 Географическое расположение образцов генетических данных. Источник: [49] 170
76 Генетические данные различных пар популяций кошачьей лягушки в виде аллель-частотных спектров 171
77 Полученные демографические истории для различных пар популяций кошачьей лягушки 172
78 Генетические данные двух популяций американской пумы (а) и рассматриваемые модели демографической истории (б, в) 173

79 Полученные демографические истории двух популяций американской пумы для (а) модели 1 без инбридинга, (б) модели 2 с инбридингом 175
80 Генетические данные популяции огородной капусты (а) и рассматриваемые модели демографической истории (б, в) 176
81 Полученные демографические истории одной популяции огородной капусты для (а) модели 1 без инбридинга, (б) модели 2 с инбридингом 178
82 Демографическая история двух популяций орангутангов, использованная для симуляции генетических данных с помощью *stdpopsim* [6] 180
83 Модели демографической истории двух популяций орангутангов: (а) ORAN-NOMIG, (б) ORAN-MIG, (в) ORAN-STRUCT-NOMIG, (г) ORAN-STRUCT-MIG, (д) ORAN-PULSE-1, (е) ORAN-PULSE-3, (ж) ORAN-PULSE-7 181
84 Результаты настроенной модели ORAN-NOMIG для методов вычисления правдоподобия, основанных на (а) аллель-частотном спектре, (б) статистиках неравномерного сцепления генов 182
85 Результаты настроенной модели ORAN-STRUCT-NOMIG для методов вычисления правдоподобия, реализованных в (а) *дaдi*, (б) *moments*, (в) *momi2*, (г) *momentsLD* 183
86 Результаты настроенных моделей для метода вычисления правдоподобия, реализованного в *momi2*: (а) модель ORAN-NOMIG, (б) модель ORAN-PULSE-1, (в) модель ORAN-PULSE-3, (г) модель ORAN-PULSE-7 183
87 Географическое расположение образцов генетических данных. Источник: [4] 185
88 Генетические данные трех популяций современного человека на территории Российской федерации в виде аллель-частотного спектра 186
89 Рассматриваемая модель расширенного класса 187
90 Полученная демографическая история трех популяций современного человека 187
91 Сходимость рассматриваемых методов настройки параметров модели демографической истории двух популяций по (а) итерациям, (б) времени 189
92 Сходимость рассматриваемых методов настройки параметров модели демографической истории трех популяций по (а) итерациям, (б) времени 190
93 Сходимость рассматриваемых методов настройки параметров модели демографической истории четырех популяций по (а) итерациям, (б) времени 190

94 Сходимость рассматриваемых методов настройки параметров модели демографической истории пяти популяций по (а) итерациям, (б) времени 190
95 Модели демографической истории современного человека (а) четыре популяции, (б) пять популяций 192
96 Демографические истории для модели 1 четырех популяций современного человека: (а) базовая история, полученная с помощью метода Пауэлла с перезапусками в работе [55], (б) лучшая история, полученная с помощью разработанного метода BO Ensemble, (в) альтернативная история, полученная с помощью разработанного метода BO Ensemble 193
97 Демографические истории для модели 2 пяти популяций современного человека: (а) базовая история, полученная с помощью метода Пауэлла с перезапусками в работе [55], (б) история, полученная настройкой четырех параметров с помощью разработанного метода BO Ensemble, (в) лучшая история, полученная настройкой 21 параметра с помощью разработанного метода BO Ensemble, (г) альтернативная история, полученная настройкой 21 параметра с помощью разработанного метода BO Ensemble 194
98 Блок-схема разработанного метода автоматического перебора моделей расширенного класса с разным числом параметров 198
99 Пример модели трех популяций, которая соответствует ограничению (2,1,1) 200
100 Пример модели двух популяций, соответствующей ограничению (2,1,0) 200
101 Демографические истории для модели, представленной на рисунке 100, при разных значениях ее параметров 200
102 Два класса, реализованных для разработанного метода автоматического перебора моделей 203
103 Генетические данные в виде аллель-частотного спектра и демографическая история, полученная ранее в работе [45] 205
104 Демографическая история, полученная разработанным методом . 206
105 Полученные демографические истории для различных пар популяций кошачьей лягушки 207
106 Географическое расположение образцов генетических данных. Источник: [5] 208
107 Генетические данные в виде аллель-частотных спектров для (а) двух популяций, (б) и (в) трех популяций 209
108 Схема вывода демографической истории популяций голубой акулы 210
109 Демографическая история двух популяций голубой акулы 211
110 Демографическая история трех популяций голубой акулы 212
111 Структура программного комплекса GADMA 216
112 Структура классов модуля `data` программного комплекса GADMA 217

113 Структура классов модуля `engines` программного комплекса GADMA 218
114 Пример входного файла с опциями запуска для программного комплекса GADMA 221
115 Пример структуры директории с результатами запуска 222
116 Пример визуального представления демографической истории популяций и использованных статистик генетических данных, созданное GADMA 223
117 Пример вывода GADMA в командной строке 223
118 Пример отчета о покрытии программного кода GADMA тестами на сервисе *CodeCov* 226
119 Иерархическая структура каталога библиотеки *stdpopsim*, интерфейс прикладного программирования (API) и интерфейс командной строки. Источник: [6] 227
120 Пример тестового и визуального представления демографической истории, полученных с применением *demes*. Источник: [8] 229

Список таблиц

P.1 Результаты экспериментальных исследований сравнения методов настройки параметров на симулированных данных трех популяций 22

P.2 Результаты 100 запусков различных методов для поиска параметров модели с инбридингом для вывода демографической истории двух популяций пум 23

S.1 Results of experimental studies for comparing parameter tuning methods on simulated data of three populations 43

S.2 Results of 100 repeats of different methods for parameter tuning in case of model 2 with inbreeding for two puma populations 44

1 Существующие программные средства для вывода демографической истории популяций по генетическим данным 72

2 Краткое описание и обозначения гиперпараметров разработанного генетического алгоритма. 137

3 Начальные значения и область значений, используемые для оптимизации гиперпараметров разработанного генетического алгоритма. Первые два гиперпараметра `gen_size` и `n_init_const` являются целыми числами и имеют дискретную область значений. Остальные восемь гиперпараметров являются непрерывными. 138

4 Значения гиперпараметров генетического алгоритма после каждой попытки оптимизации с помощью SMAC 140

5 Темпы ускорения генетического алгоритма при использовании новых конфигураций по сравнению с исходной версией 141

6 Результаты 50 повторов тестируемых методов настройки параметров для моделей демографической истории одной популяции . . . 164

7 Результаты 50 повторов тестируемых методов настройки параметров для моделей демографической истории двух популяций . . . 166

8 Результаты 10 повторов тестируемых методов настройки параметров для моделей демографической истории трех популяций 169

9 Результаты 100 повторов различных методов для поиска параметров модели 1 без инбридинга демографической истории двух популяций пум 174

10 Результаты 100 повторов различных методов для поиска параметров модели 2 с инбридингом демографической истории двух популяций пум 174

11 Результаты 100 повторов различных методов для поиска параметров модели 1 без инбридинга демографической истории одной популяции огородной капусты 177

12 Результаты 100 повторов различных методов для поиска параметров модели 2 с инбридингом демографической истории одной популяции огородной капусты 177

13 Среднее время одного вычисления правдоподобия при использовании метода GA+NM и различных методов вычисления правдоподобия 180
14 Среднее число вычислений правдоподобия, которое потребовалось для настройки параметров рассматриваемых моделей с использованием метода GA+NM и различных методов вычисления правдоподобия 182
15 Список опций входного файла программного комплекса GADMA 220
16 Ограничения программного комплекса GADMA при использовании разных движков 224

Приложение А. Благодарности

Я хочу выразить благодарность своему научному руководителю, Владимиру Игоревичу Ульянцеву, который согласился взять меня в ученики шесть лет назад, увидев потенциал в моих исследованиях, направлял меня все эти годы и привел к защите этой диссертации. Отдельно хочу поблагодарить моего коллегу и соавтора Павла Владимировича Добрынина, который невольно стал основоположником направления моих исследований, за его поддержку, помощь и советы на всех этапах моей научной работы. Я признательна профессору Анатолию Абрамовичу Шалыто за неоценимую помощь при подготовке этой диссертации и моему соавтору Клаусу Копфли за веру в меня и терпение при написании текстов статей на английском языке.

Я благодарна коллегам из международного научного центра «Компьютерные технологии» и с факультета информационных технологий и программирования Никите Алексееву, Анне Жук и Антону Замятину за поддержку на протяжении различных периодов подготовки диссертации. Также я благодарна международному консорциуму PopSim, который стимулировал мои исследования на протяжении многих лет.

Я благодарна моей семье и друзьям. Особенно я хочу поблагодарить моего мужа, Вячеслава Боровицкого, за постоянную поддержку и твердую веру в меня. Его участие в обсуждении научных тем и обмен идеями стали для меня вдохновением, давая новые перспективы и направления в моей работе. Я благодарна своей маме Елене Хамитовой, папе Эдуарду Носкову, сестре Марии Зайцевой и свекрови Наталии Боровицкой за оказанную поддержку, теплоту и любовь. Сложно выразить словами насколько это было и остается важным для меня.

Благодарю всех, кто был частью этого пути, но не был упомянут выше!

Приложение Б. Награды автора, полученные во время работы над диссертацией

17th Annual Human-Competitive Results Award ("Humies") 2020 BRONZE AWARD

US $2,000

To

Ekaterina Noskova, Vladimir Ulyantsev, Klaus-Peter Koepfli, Stephen J. O'Brien, Pavel Dobrynin

For entry entitled:

"gadma: genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data"

In Cancún as well as in the virtual space

sigevo

acm

SYSTEMS BIOLOGY
FELLOWSHIP PROGRAM

Systems Biology Program Winner Certificate

This Certificate is proudly presented to
Ekaterina Noskova
for the project
"Computational methods for unsupervised demographic inference of multiple populations from genomic data"

Skolkovo
Institute of Science
and Technology

Skoltech

Konstantin
Severinov,

Skoltech
Professor