Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Санкт-Петербургский национальный исследовательский
Академический университет Российской академии наук»
Центр высшего образования

Кафедра математических и информационных технологий

Носкова Екатерина Эдуардовна

Автоматизация процесса вывода совместной демографической истории нескольких популяций из аллель-частотного спектра

Магистерская диссертация

Допущена к защите.
Зав. кафедрой:
д. ф.-м. н., профессор Омельченко А. В.

Научные руководители:
к. т. н. Ульянцев В. И.
Добрынин П. В.

Рецензент:
Авдеев П. В.

Санкт-Петербург
2018

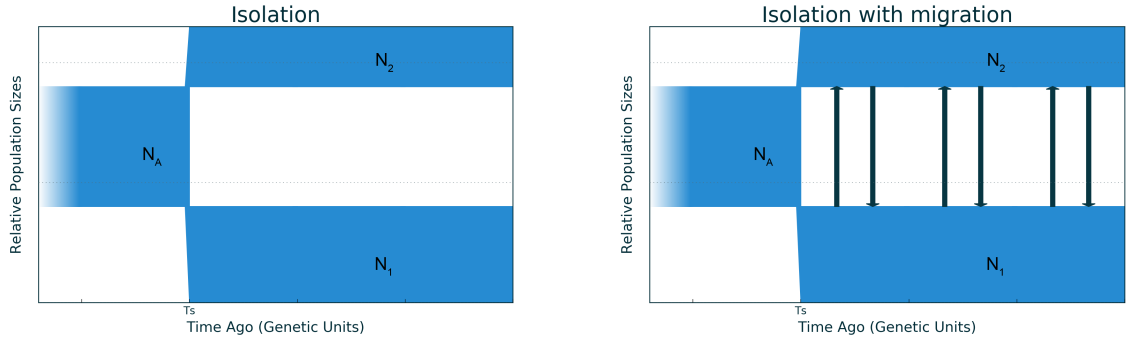
Оглавление

Введение

Понимание роли демографии и отбора в формировании видов и популяций является центральной проблемой популяционной генетики. За последнее время, благодаря развитию технологий секвенирования, происходит накопление данных по геномам особей близких видов, что может помочь пролить свет на решение данной проблемы. Демографическая история или демографическая модель популяций — это история развития этих популяций, которая включает в себя такие события как миграция, разделение популяций и изменения (эффективной) численности. На рисунке 1 приведены примеры основных простейших демографических моделей [34].

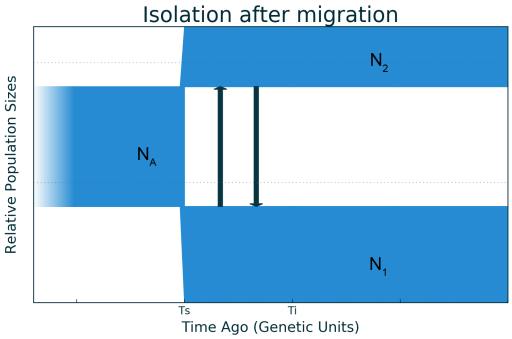
Демографические модели, полученные из генетических данных, играют важную роль в популяционной генетике. Во-первых, они дополняют археологические сведения об исторических событиях, которые не оставили письменных свидетельств, например, темпы и время основных континентальных миграций [24, 10]. Во-вторых, демографические модели способствуют увеличению информации об эволюционных силах и их влиянии на геномы, например, данные о регионах, которые подверглись недавнему отбору [27]. И наконец, они могут быть основой для последующих исследований популяций и медицинских генетических исследований. Такое разнообразие применений привело к появлению большого числа работ, посвященных выводу демографических моделей для популяций людей и других видов: [1, 38, 17, 12, 14, 25, 16, 4, 40].

В идеале мы бы хотели использовать для анализа полные геномы, однако это вычислительно трудная задача, поэтому приходится прибегать к различного вида упрощениям. Аллель-частотный спектр — совместное распределение частот полученных аллелей в популяциях (подробнее в соответствующем разделе основной части работы), является одним из наиболее распространенных и удобных представлений генетической информации. В последнее время было посвящено много работ анализу аллель-частотного спектра и его зависимости от истории развития популяций [38, 1, 17, 39, 26], что привело к появлению различных математических моделей, описывающих данную зависимость, которые в свою очередь привели к появлению различных методов. Примерами реализации этих методов являются *∂a∂i* [16] и *moments* [15], которые основаны на двух моделях: «бесконечного числа сайтов» и Райта-Фишера, и предоставляют возможность симулировать аллель-частотный спектр из заданной демографической модели, а затем вычислить значение правдоподобия — степень схожести полученного спектра со спектром, построенным по реальным данным. Таким образом можно попытаться найти демографическую модель с максимальной величиной правдоподобия. Однако на данный момент не существует программного обеспечения, которое строило бы демографическую модель из данных автоматически [30]. *∂a∂i* и

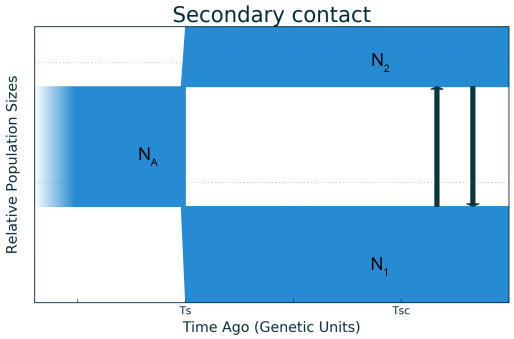


(а) Изоляция

(б) Изоляция с миграцией



(в) Изоляция после миграции



(г) Вторичный контакт

Рисунок 1 – Примеры основных простейших демографических моделей. Во всех существовала одна общая предковая популяция размера N_A, которая затем разделилась T_S времени назад. Получившиеся две популяции существуют по сей день и имеют размеры N_1 и N_2 соответственно. Модель (а), где степени миграций в обоих направлениях равны нулю, соответствует аллопатрическому видообразованию. Остальные модели включают миграции: постоянные (б), перед изоляцией популяций (в), после изоляции (г).

moments имеют ряд существенных недостатков: во-первых, пользователь должен обладать навыками программирования, чтобы ими воспользоваться, во-вторых, они предлагают поиск параметров только для фиксированной пользователем модели, в-третьих, предоставляемые методы оптимизации оказываются неэффективными на практике. Методы, основанные на градиентном спуске, используют численное дифференцирование, так как аналитически градиент не посчитать. Но на практике к сложности с вычислением градиента прибавляется также следующая проблема: оказывается, что оптимизируемая функция имеет сложную структуру (локальные оптимумы, сильное колебание абсолютного значения градиента), что мешает градиентному спуску найти глобальный оптимум. Наличие нескольких локальных оптимумов также вызывает неэффективность и второго класса методов, которые предлагают существующие решения — локального поиска без использования градиента. Однако несмотря на все эти недостатки, *∂a∂i* и *moments* являются эффективным аппаратом, так как могут симулировать аллель-частотный спектр из модели любой сложности.

Другими словами, задача поиска демографической модели по аллель-частотному спектру – это обратная задача, которую можно решать с помощью решения прямых задач приближенными численными методами. Функция от демографической модели, которая показывает сходство с наблюдаемыми данными, не имеет какой-либо аналитической формы, а, кроме того, требует некоторых временных затрат на вычисление. Вместе с тем отсутствие возможности точного и быстрого дифференцирования и сложная структура оптимизируемой функции мотивируют к использованию алгоритмов глобальной оптимизации, к которым относится генетический алгоритм.

Генетический алгоритм — один из наиболее эффективных эвристических методов глобального поиска, впервые был предложен Джоном Холландом в 1975 [13]. Он основан на естественном отборе и совершенствует набор решений задачи с помощью операций «мутации», «скрещивания» и «отбора», чтобы найти лучшее по значению функции адаптации или по-другому целевой функции. Основной областью использования генетического алгоритма является задача оптимизации целевой функции, которая либо является не дифференцируемой, либо ее не удается дифференцировать достаточно эффективным способом, например, когда функция не представима в виде так называемого «closed-form expression». Генетические алгоритмы нашли применение во многих областях и работах в биологии: например, филогенетический анализ [37] или построение генома общего предка [8].

В данной работе предложен метод поиска демографической модели, оптимально соответствующей наблюдаемому аллель-частотному спектру, основанный на генетическом алгоритме с использованием существующих решений для симулирования аллель-частотного спектра из заданной демографической модели, а именно *дадi* и *moments*. Предложенный метод поддерживает до 3х популяций и был реализован в программном обеспечении GADMA (Genetic Algorithm for Demographic Model Analysis), код которого выложен в открытый доступ и может быть найден по ссылке: `https://github.com/ctlab/GADMA`. Эффективность метода была проверена на реальных данных: на геномах современных людей и бабочек *E. gillettii*.

Основная часть работы состоит из трех частей: в первой расположен обзор предметной области: введение в популяционную генетику и существующие методы поиска демографических моделей, во второй описаны основные методы работы: описание генетического алгоритма, и в последней результаты экспериментальных исследований.

1. Обзор предметной области

1.1. Основные понятия и методы популяционной генетики

Популяционная генетика — это раздел генетики, который изучает генетическое разнообразие внутри и между популяциями, и является частью эволюционной биологии. Основными движущими силами эволюции являются естественный отбор, мутагенез, дрейф и поток генов. Центральным понятием популяционной генетики является понятие аллели — вариации генов (локусов генома). Именно распределение частот аллелей и их изменение под влиянием сил эволюции изучает данный раздел биологии. В популяционной генетике существует несколько математических моделей, описывающих эволюцию популяций [5]. Эти модели помогают получать информацию о представляющих интерес величинах. Для моделирования мутации существуют модели «бесконечного числа сайтов» или «бесконечного числа аллелей», а для моделирования поколений — модели Райта-Фишера, Морана или Берроуза-Кокерхэма. Опишем некоторые из них. Модели также часто рассматривают вместе, например, модели «бесконечного числа сайтов» и Райта-Фишера [16], они могут дополнять друг друга.

1.1.1. Модель «бесконечного числа сайтов»

Модель «бесконечного числа сайтов» была впервые предложена Мотоо Кимурой в 1969 году [20]. Она моделирует взаимосвязь мутации и возникновения новых аллелей в популяции, что позволяет рассчитывать гетерозиготность или генетическое разнообразие для оценки генетических расстояний между рассматриваемыми популяциями [35].

Пусть у нас имеется диплоидная популяция размера N, где каждая особь имеет бесконечную последовательность ДНК и рекомбинация отсутствует. Любая мутация происходит в новом сайте, при этом общее число мутантных сайтов на цепочку в расчете на одно поколение является случайной величиной, имеющей распределение Пуассона со средним u.

Таким образом поскольку каждая новая мутация должна произойти в новом сайте, не может быть гомоплазии или обратной мутации. А среднее число новых мутированных сайтов на поколение будет равно:

$$\theta = 4 \cdot N_e \cdot u, \tag{1}$$

где N_e — эффективная численность популяции. «4» возникает в формуле потому, что эффективная численность измеряется только по особям женского пола, то есть особей всего $2N_e$, а также особи диплоидные.

На практике, для поиска θ обычно используют следующую формулу [16]:

$$\theta = 4 \cdot N_e \cdot \mu \cdot L = N_e \cdot \theta_0, \tag{2}$$

где μ — средняя скорость мутации, равная вероятности мутации одного сайта за поколение, L — эффективная длина последовательности, а $\theta_0 = 4u = 4 \cdot \mu \cdot L$ — среднее число новых мутировавших сайтов в одной особи за поколение.

1.1.2. Модель Райта-Фишера (Wright-Fisher)

Одной из самых популярных моделей, описывающих изменение частот аллелей между поколениями, является стохастическая эволюционная модель, которая была введена отдельно Фишером в 1922 году [7] и Райтом в 1931 году [36] и поэтому называется моделью Райта-Фишера, она также была расширена Кимурой в 1955 году [18]. На самом деле существует целый ряд моделей Райта-Фишера, которые учитывают различные факторы, такие как мутация, отбор, существование двух полов, географические факторы, изменения численности популяций и так далее. Сначала рассмотрим простейшую модель, не учитывающую все это, а затем опишем более общую.

Любая модель Райта-Фишера предполагает, что в популяции существует максимум две аллели каждого гена, особи двуполые, их поколения не пересекаются, как, например, у однолетних растений, и что гаметы нового поколения формируются независимо случайным выбором из гамет предыдущего поколения.

Простейшая модель Райта-Фишера. Рассмотрим диплоидную популяцию фиксированного размера N. Предположим, что особи этой популяции двуполые, в каждом локусе может быть только две аллели и поколения не пересекаются. Рассмотрим один локус с двумя аллелями: A_1 и A_2. Между ними не существует селективной разницы, нет мутаций, географических или каких-либо других осложняющих факторов. Пусть $X(t)$ — число генов A_1 в поколении t. Ясно, что в любом поколении t есть $2N$ генов, а следовательно $X(t)$ принимает одно из значений от 0 до $2N$ для любого t: $X(t) \in [0, 1, ..., 2N]$.

Пусть гены из поколения $t+1$, получаются с помощью выборки с повторениями генов поколения t. Это означает, что число $X(t+1)$ является биномиальной случайной величиной с числом испытаний равным $2N$, и вероятностью успеха $\frac{X(t)}{2N}$. То есть вероятность p_{ij} того, что $X(t+1) = j$ при условии, что $X(t) = i$, равна:

$$p_{ij} = P\left(X(t+1) = j | X(t) = i\right) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \tag{3}$$

Важно отметить, что процесс изменения числа аллели A_1 является марковской цепью с матрицей переходов $P = \{p_{ij}\}$ и начальным состоянием $X(0)$.

Общая модель Райта-Фишера. Биномиальную форму простейшей модели Райта-Фишера можно записать в общем виде следующим образом:

$$p_{ij} = \binom{2N}{j} \psi_i^j \left(1 - \psi_i\right)^{2N-j}, \tag{4}$$

где ψ_i — вероятность успеха, то есть вероятность получить аллель A_1. Данная запись определяет общую модель Райта-Фишера, в которой ψ_i формируется учитыванием различных осложняющих факторов. Приведем пример для модели с мутацией и отбором.

Если у нас есть мутация и аллель A_1 мутирует в аллель A_2 с вероятностью u, а аллель A_2 мутирует в A_1 с вероятностью v, то ψ будет:

$$\psi_i = \frac{i}{2N}(1-u) + \left(1 - \frac{i}{2N}\right) v \tag{5}$$

Первое слагаемое соответствует вероятности выбрать гамету с аллелью A_1 и не мутировать, второе — вероятности выбрать гамету с A_2 и мутировать ее в A_1.

Теперь допустим существование отбора. Пусть для каждой аллели A_1, A_2 у нас имеются значения приспособленности ω_1, ω_2 соответственно, которые определяют отбор: менее приспособленные особи имеют меньший шанс дать потомство. Для простоты мы рассматриваем отбор каждой аллели. Случай отбора генотипов, то есть когда приспособленности гомозигот и гетерозигот отличаются, аналогичен. Относительная частота аллели A_1 в текущей популяции равна $x = i/2N$, тогда:

Аллель	A_1	A_2
Приспособленность	ω_1	ω_2
Относительная частота	x	$1-x$

Вероятность получить аллели A_1 и A_2 в следующем поколении равны $\frac{\omega_1}{\overline{\omega}}x$ и $\frac{\omega_2}{\overline{\omega}}(1-x)$ соответственно, где $\overline{\omega} = \omega_1 x + \omega_2(1-x)$ — средняя приспособленность в популяции. А вероятность успеха в общей модели Райта-Фишера с учетом мутации и естественного отбора будет равна:

$$\psi_i = \frac{1}{\overline{\omega}} \left\{ \omega_1 \frac{i}{2N} (1-u) + \omega_2 \left(1 - \frac{i}{2N}\right) v \right\} \tag{6}$$

1.1.3. Применение теории диффузии в популяционной генетики

Как говорилось выше математические модели популяционной генетики помогают получить информацию об эволюции и развитии видов и популяций. Например, если у нас есть популяция с определенным числом аллелей, они могут помочь ответить на вопрос, сколько времени в среднем требуется, чтобы осталась только одна аллель (фиксация аллели) и возможно ли такое вообще (есть ли равновесие)?

Однако чем больше факторов мы учитываем, тем сложнее становится модель и тем труднее явно вывести интересующие величины. Поэтому вместо этого можно аппроксимировать эти величины с помощью простых, достаточно точных выражений, в чем поможет общий метод, основанный на приближении дискретного процесса непрерывным диффузионным.

Пусть у нас имеется простейшая модель Райта-Фишера: диплоидная популяция размера N в каждом поколении, поколения не пересекаются, а процесс изменения частот аллелей — марковская цепь. От частот можно легко перейти к рассмотрению относительных частот, поделив на $2N$. Рассмотрим две аллели одного локуса: A_1 и A_2, которые имеют относительные частоты x и $(1-x)$ соответственно. Пусть $\phi(p,x,t)$ — условная плотность вероятности того, что относительная частота аллели A_1 равна x в момент времени t при условии, что начальная частота в момент $t=0$ была равна p. Эта плотность также задает переходную вероятность того, что частота аллели изменится с p на x за время t. При фиксированной p, ее можно опустить и $\phi(p,x,t)$ превратится в $\phi(x,t)$.

Тогда для $\phi(x,t)$ можно вывести следующее уравнение [19]:

$$\frac{\partial\phi(x,t)}{\partial t} = -\frac{\partial}{\partial x}\{M(x)\phi(x,t)\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{V(x)\phi(x,t)\}, \tag{7}$$

где $M(x)$, $V(x)$ — среднее и дисперсия изменения относительных частот аллели A_1 за одно поколение. Это уравнение называется прямым уравнением Колмогорова в математике [21] или уравнением Фоккера-Планка в физике.

Если рассмотреть общую модель Райта-Фишера, то уравнение Фоккера-Планка будет верно и иметь различный вид при учитывании влияния разных факторов [19]. Например, для нескольких популяций с миграцией, отбором и изменением численности оно будет следующим [16]:

$$\frac{\partial f(x;t)}{\partial\tau} = -\sum_{i=1,..,P}\frac{\partial}{\partial x}\left(\gamma_i x_i(1-x_i) + \sum_{j=1,...P} M_{ij}(x_i - x_j)\right) f(x;t) + \\ + \frac{1}{2}\sum_{i=1,..,P}\frac{\partial^2}{\partial x^2}\frac{x_i(1-x_i)}{\nu_i} f(x;t), \tag{8}$$

где время в единицах $\tau = \frac{t}{2N_{ref}}$, а N_{ref} — численность некоторой референсной популяции (обычно размер N_A общей предковой); $\nu_i = \frac{N_i}{N_{ref}}$ — относительная численность популяции; $M_{ij} = 2N_{ref}m_{ij}$ — относительные темпы миграции, а m_{ij} — доля особей в популяции i, которые пришли из популяции j; $\gamma_i = 2N_{ref}s_i$ — относительные силы естественного отбора.

Данное уравнение не имеет решения в аналитическом виде и его считают численно. Более того, если принять в рассмотрение модель «бесконечного числа сайтов», то можно ввести мутацию, внедрив возмущение пропорциональное величине потока $\theta =$

$N_r ef\theta_0$ на малые частоты в каждый момент времени [16]. Уравнение Фоккера-Планка позволяет приближенно вычислять многие интересующие аспекты популяционной генетики, в частности, ожидаемый аллель-частотный спектр, о котором пойдет речь дальше.

1.1.4. Аллель-частотный спектр

Одним из наиболее популярных и удобных представлений генетической информации является аллель-частотный спектр (Allele Frequency Spectrum). Для его построения используют, так называемые, полученные аллели (derived allele) — аллели, которые отличаются от референсной аллели древнего предка, а следовательно были получены или приобретены в ходе эволюции. Довольно часто в качестве локусов аллелей берут отдельные позиции в геноме. Тогда полученные аллели — это единичные нуклеотидные варианты (Single Nucleotide Variants, SNV's).

Определение 1.1. Аллель-частотный спектр (АЧС) N популяций — это совместное распределение частот полученных аллелей у N популяций.

По-другому, аллель-частотный спектр N популяций является гистограммой размерности N, где оси соответствуют популяциям и каждый элемент содержит число локусов, на которых полученная аллель, встретилась определенное число раз. Например, если у нас две популяции, то их аллель-частотный спектр будет двумерной матрицей A, где элемент $A[i, j]$ на позиции i, j является числом полученных аллелей, которые встретились у i хромосом первой популяции и у j хромосом второй.

Размер матрицы обусловлен числом хромосом в популяциях, которых мы взяли для построения: если у нас n_1 хромосом первой популяции, n_2 хромосом второй, ..., n_P хромосом P-ой популяции, то АЧС будет P-мерной матрицей размерности $(n_1 + 1) \times (n_2 + 1) \times \cdots \times (n_p + 1)$. Единица добавляется к числу хромосом, так как аллель может встретится и у 0 хромосом. Примеры аллель-частотных спектров двух популяций представлены на рисунке 2.

Аллель-частотный спектр предполагает независимость локусов, по которым его строят, что в случае SNV не всегда истинно. Таким образом АЧС теряет информацию о связанности и позиции аллелей. К сожалению, такие упрощения приводят к еще одному важному недостатку аллель-частотных спектров. Было показано, что в случае одной изолированной популяции несколько демографических моделей могут соответствовать одному АЧС [26]. В случае нескольких популяций такого примера нет, но и обратного не доказано. Поэтому мы должны учитывать это и предполагать, что наш метод — лишь один из доступных способов получить информацию о развитии видов и популяций.

В случае модели Райта-Фишера и «бесконечного числа сайтов» ожидаемый аллель-частотный спектр P популяций размера $(n_1 + 1)(n_2 + 2) \cdots (n_P + 1)$ может

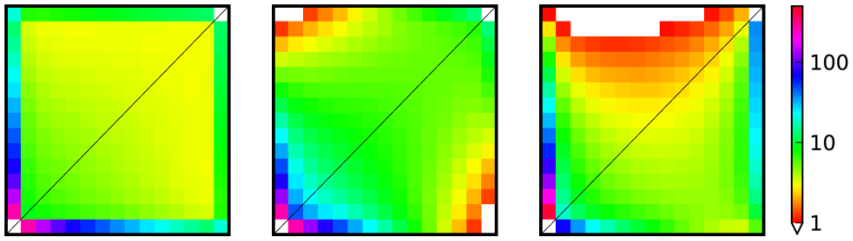


Рисунок 2 – Примеры аллель-частотного спектра двух популяций. Рисунок позаимствован из работы [16].

быть найден с использованием решения $\phi(x,t)$ уравнения 8 диффузии следующим образом [16]:

$$M[d_1,d_2,\cdots,d_P]=\int_0^1\cdots\int_0^1\prod_{i=1,2,\cdots P}\binom{n_i}{d_i}x_i^{d_i}(1-x_i^{n_i-d_i})\phi(x_1,x_2,\cdots,x_P)dx_i, \qquad (9)$$

1.2. Существующие методы вывода сложных демографических историй

Существует два типа методов для вывода совместной демографических моделей нескольких популяций: первый основан на данных по гаплотипам и второй на аллель-частотных спектрах. Первый требует довольно аккуратной дополнительной предобработки данных, так называемого фазирования (phasing), а вывод демографической модели вычислительно сложен, что является довольно важными недостатками таких методов. Для хорошего фазирования требуется правильно выбранная эталонная популяция, которая родственно близка особям, которых мы взяли для анализа, что иногда бывает невозможно. Также может потребоваться генотипировать семьи или большое число особей, что увеличивает стоимость всего проекта.

1.2.1. Методы, основанные на данных гаплотипов

Основным методом определения демографической истории популяции на основе данных о гаплотипах является Методы Монте Карло по схеме марковской цепи (Markov Chain Monte Carlo, MCMC) для оценки коалесцентивной истории диплоидных геномов. Эти методы используют разные модели построения истории коалесценции и реализуются в нескольких популярных программных решениях, например, парная последовательная марковская коалесценция (pairwise sequentially Markovian coalescent, PSMC [23]), множественная последовательная марковская коалесценция (multiple sequentially Markovian coalescent, MSMC [29]) и анализ правдоподобия с помощью алгоритма Метрополиса с использованием случайной

коалесценции (Likelihood Analysis with Metropolis Algorithm using Random Coalescence, LAMARC [22]). PSMC по геному особи строит историю для ее популяции, в то время как MSMC и LAMARC выводят истории нескольких. Однако MSMC требует большого количества оперативной памяти для запуска и дает коэффициенты перекрестной коалесценции вместо темпов миграций, которые нас интересуют. LAMARC же является временно-сложным: для оценки скорости рекомбинации 60-ти митохондриальных ДНК длины каждая 16 тысяч оснований требует 3-4 недель [22], что показывает его неприменимость к полногеномным данным.

1.2.2. Методы, основанные на аллель-частотном спектре

Существующие решения для вывода демографической истории нескольких популяций из наблюдаемого аллель-частотного спектра основаны на максимизации правдоподобия: пользователь задает демографическую модель и ее параметры, затем из нее симулируется ожидаемый аллель-частотный спектр и считается правдоподобие — меру схожести спектров. Модель с наибольшим правдоподобием является решением поставленной задачи. Для ее поиска можно использовать оптимизации параметров, однако существующие программные решения предлагают методы локального поиска, которые надо запускать много раз для поиска глобального максимума.

Существует два популярных метода для симуляции аллель-частотного спектра из заданной пользователем демографической модели: основанный на дифференциальном уравнении диффузии (∂a∂i [16]) и основанный на непрерывной последовательной марковской коалисцентной аппроксимации (fastsimcoal2 [6]). ∂a∂i является пакетом для популярного языка Python и численно решает дифференциальное уравнение (Partial Differential Equation, PDE) диффузии 8, которое не имеет решения в аналитическом виде, для получения соответствующего аллель-частотного спектра. Однако в случае анализа моделей с большим числом параметров, популяций или большой размерности спектра PDE приводит к вычислительным трудностям, из-за чего ∂a∂i поддерживает только до трех популяций. Fastsimcoal2 может обрабатывать любое количество популяций и любую демографическую модель, но он зачастую бывает вычислительно трудным, поскольку fastsimcoal2 несколько раз подряд имитирует аллель-частотный спектр для модели, чтобы получить стабильный.

Совсем недавно был представлен и реализован (*moments* [15]) новый подход, основанный на аппроксимации моментов случайного процесса, что быстрее и стабильнее, чем решение дифференциального уравнения в ∂a∂i. *moments* также является пакетом языка Python, обладает тем же интерфейсом, что ∂a∂i, и представляют собой компромисс между скоростью и точностью, а также может поддерживать до пяти популяций. Однако также как и ∂a∂i он использует численные

методы.

Для симулирования аллель-частотного спектра из заданной демографической модели были выбраны два программных решения: самый популярный — ∂a∂i, и новый — *moments*, с тем же интерфейсом. Они предлагают возможность вычислить правдоподобие и максимизировать его локальной оптимизацией параметров для фиксированной пользователем модели.

Не учитывая того, что оптимизации ограничены предложенной им моделью, на практике они все равно оказываются неэффективными: во-первых, для вычисления градиента они используют численное дифференцирование, вычислительная сложность которого высока, так как высока вычислительная сложность самой функции, а, во-вторых, пространство поиска имеет сложную структуру с наличием нескольких локальных минимумов и колебаниями абсолютного значения градиента, которая мешает найти глобальный оптимум. Более того, с увеличением числа параметров увеличивается и сложность этой структуры , и вместе с тем падает эффективность существующих алгоритмов оптимизации. Такое поведение приводит к дополнительным ограничениям на рассматриваемые модели, в частности, к уменьшению числа параметров и фиксированию законов изменения численности популяций. Примеры ограничений можно посмотреть в разделе «Экспериментальные исследования», где приведены существующие модели из статьей для сравнения.

1.2.3. Сравнение демографических моделей с помощью правдоподобия

Допустим мы умеем считать ожидаемый аллель-частотный спектр M из заданной демографической модели с параметрами. У нас имеется P популяций, размеров n_1, n_2, ..., n_P соответственно.

Предположим, что каждый элемент аллель-частотного спектра $S[d_1, .., d_P]$ — это независимая Пуассоновская величина [28] со средним равным $M[d_1, .., d_P]$. Тогда мы можем посчитать правдоподобие — вероятность получить наблюдаемый спектр S, если ожидаемый M, как произведение $(n_1 + 1)(n_2 + 1)...(n_P + 1)$ Пуассоновских правдоподобий:

$$\mathcal{L}(M|S) = \prod_{i=1,\cdots P} \prod_{d_i=1,\cdots n_i} \frac{e^{-M[d_1,\cdots,d_P]} M[d_1,\cdots,d_P]^{S[d_1,\cdots,d_P]}}{S[d_1,\cdots,d_P]!} \tag{10}$$

∂a∂i и *moments* предлагают сравнивать модели по логарифму от этого значения $\log(\mathcal{L}(M|S))$. Заметим, что так как $\mathcal{L}(M|S) \in [0, 1]$, то $\log(\mathcal{L}(M|S)) \in [-\inf, 0]$ и чем ближе к 0 тем лучше. В данной работе $\log(\mathcal{L}(M|S))$ был выбран в качестве целевой функции генетического алгоритма, о котором пойдет речь дальше.

1.2.4. Информационный критерий Акаике и Байесовский информационный критерий

С увеличением числа параметров модели мы можем ее переобучить: модель с большим числом параметров сможет подобрать модель лучше, чем модель с меньшим числом, но будет при этом хуже соответствовать действительности, например, из-за ошибок в данных.

Для сравнения моделей с разным числом параметров используют информационный критерий Акаике (AIC), который был предложен Хиротугу Акаике в 1971 году [2]:

$$AIC(M, S) = 2 \cdot k - 2 \cdot \log(\mathcal{L}(M|S)), \tag{11}$$

где k — число параметров модели, а $\log(\mathcal{L}(M|S)$ — значение логарифмической функции правдоподобия.

Байесовский информационный критерий (BIC) [32] является наиболее часто используемой модификацией AIC, он имеет больший штраф на количество параметров:

$$BIC(M, S) = \log(n) \cdot k - 2 \cdot \log(\mathcal{L}(M|S)), \tag{12}$$

где n — объем выборки. В случае аллель-частотного спектра n — это число элементов спектра. Чем BIC и AIC меньше, чем модель лучше подходит.

1.3. Постановка задачи

Рассмотрим функцию $f(\Theta, A, C)$, которая принимает на вход параметры $\Theta = \{\theta^k\}_{k=1}^{N_\Theta}$, $\theta^k \in \mathbb{R}$, аллель-частотный спектр $A \in \mathbb{R}^{P \times P}$, множество констант $C = \{c_k\}_{k=1}^{N_C}$ и возвращает меру соответствия параметров Θ аллель-частотному спектру A.

Функция $f(\Theta, A, C)$ строит демографическую модель по параметрам Θ, которые ее определяют, вычисляет по ней ожидаемый аллель-частотный спектр M с учетом констант C и затем определяет степень сходства между M и наблюдаемым A по формуле 10. Константами могут являться различные параметры алгоритмов вычисления ожидаемого спектра, например, размеры сетки для численного решения дифференциального уравнения, или параметры моделей, например, среднее число новых мутированных сайтов в особи за одно поколение θ_0 или время t_g на одно поколение. Функция f может иметь различные детали реализации, например, в данной работе для этого были выбраны ∂a∂i и *moments*.

Целью данной работы является разработка алгоритма поиска демографической модели, лучше всего соответствующей данному аллель-частотному спектру. Формально задачу можно сформулировать следующим образом:

Вход

- $A \in \mathbb{R}^{P \times P}$ — P-мерная матрица, $P \in \{2, 3\}$.
- $C = \{c_k\}_{k=1}^{N_C}$ — множество констант.

Выход

- Набор $\Theta \in \mathbb{R}^{N_\Theta}$ значений, который максимизирует значение функции f:

$$\Theta : \quad f(\Theta, A, C) \to \max$$

Существуют приближенные решения данной задачи с дополнительным входом — фиксированной демографической моделью, с помощью различных алгоритмов локального поиска, однако на практике, как говорилось выше, эти алгоритмы оказываются неэффективными. В данной работе представлен новый алгоритм приближенного решения данной более общей задачи автоматически с использованием одного из самых эффективных методов глобальной оптимизации — генетического алгоритма.

1.4. Выводы по разделу «Обзор предметной области»

В данном разделе были описаны основные понятия, термины, математические модели и существующие теоретические результаты популяционной генетики. Были описаны основные методы поиска демографической истории по генетическим данным и их недостатки. Методы, основанные на анализе данных гаплотипов, вычислительно сложны и, кроме того, требуют аккуратной предобработки данных. В противоположность им, существуют методы, основанные на аллель-частотном спектре, которые в последнее время получили широкое распространение, лишены этих недостатков. Однако они требуют от пользователя навыков программирования и предлагают алгоритмы поиска параметров, которые применимы только для фиксированного числа параметров и на практике терпят неудачу в поиске глобального оптимума.

2. Предлагаемый метод поиска демографической модели по аллель-частотному спектру

2.1. Представление демографической модели

Перед описанием основного разработанного алгоритма, приведем разработанное представление демографической модели.

Пусть разделение популяции — это разделение одной популяции ровно на две новые, популяции не вымирают и не аггрегируются. Тогда количество разделений популяций напрямую зависит от количества рассматриваемых популяций, то есть на один меньше.

Представим демографическую модель как последовательность временных периодов и разделений популяций, которые имеют определенное число параметров. Зафиксируем порядок популяций: первая — самая древняя, последняя — самая недавно образовавшаяся. Этот порядок обычно известен, например, для людей, либо его можно перебрать. Тогда если число популяций не больше трех, то каждое разделение будет делить последнюю образовавшуюся популяцию. Таким образом у него будет только один параметр — доля, в которой численность популяции делится.

Следующим важным шагом является концепция «периода времени». Во-первых, это отрезок времени, в течение которого для каждой популяции поддерживается определенный закон изменения эффективного размера. Мы рассматриваем три основных закона, которые также изображены на рисунке 3: внезапный, линейный и экспоненциальный. Во-вторых, параметры темпов миграций между популяциями постоянны в течение этого отрезка времени. Таким образом, каждый временной период имеет следующие параметры:

- время,
- размеры численности популяций в конце этого периода,
- законы изменения численности популяций,
- миграции между популяциями, если их больше одной.

А размеры численности популяций в начале периода равны численностями популяций в конце предыдущего периода.

Первый период является особенным: мы считаем, что он длится с самого начала существования общей предковой популяции, и закон изменения ее размера внезапный [16]. Поэтому для этого периода единственным параметром является размер популяции предков.

Заметим, что количество разделений определяется числом рассматриваемых популяций, но число периодов можно менять и таким образом менять число

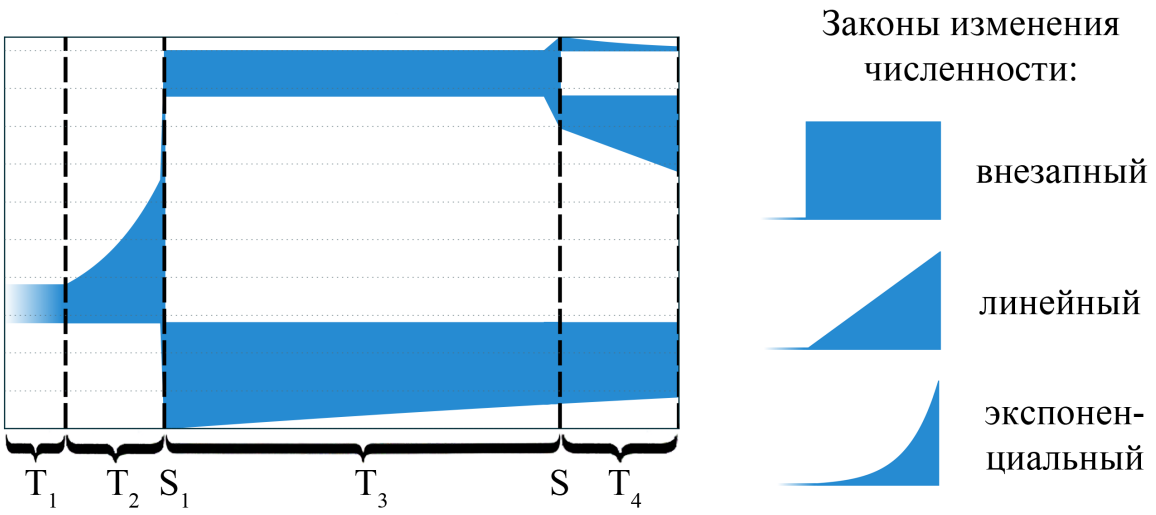


Рисунок 3 – Структура демографической модели: T_i — временные периоды, S_i — разделения популяций, и рассматриваемые законы изменения численности. Структура представленной модели будет (2,1,1).

параметров демографической модели, ее подробность и сложность. Тогда определим понятие структуры демографической модели:

Определение 2.1. Структура модели — число периодов до и после одного разделения в случае двух популяций, в случае трех популяций — число периодов до первого, между первым и вторым разделением и после второго разделения.

Например, мы наблюдаем три популяции, вначале была популяция предков, и она когда-то начала расти, затем произошло разделение, и в течение одного периода времени изменились численности новых двух популяций, затем произошло разделение второй популяции, и теперь три текущие популяции изменялись в течение одного периода времени, структура такой модели будет (2,1,1). Модель этой структуры представлена на рисунке 3. Назовем структуры (1,1), (1,1,1) простейшими.

Более формально, структура модели — это последовательность вида $S^* = \{s_i^*\}_{i=1}^P$, $s_i \in \mathbb{N}$, где $P \in \{2, 3\}$ — число популяций. При этом число параметров $N_\Theta(S^*)$ демографической модели структуры S^* будет определятся следующим образом:

$$N_\Theta(S^*) = (P-1) + \sum_{i=0}^{P} N_\theta^i(s_i^*), \quad \text{где} \quad N_\Theta^i(s_i^*) = \begin{cases} 1 + 3(s_1^* - 1), & \text{если } i = 1, \\ 7s_2^*, & \text{если } i = 2, \\ 13s_3^*, & \text{если } i = 3. \end{cases}$$

Слагаемое $(P-1)$ соответствует числу параметров разделений, а $\sum_{i=0}^{P} N_\Theta^i(s_i^*)$ — числу параметров временных периодов.

Таким образом мы можем однозначно интерпретировать демографическую модель по списку параметров и ее структуре, зафиксировав для каждого временного периода

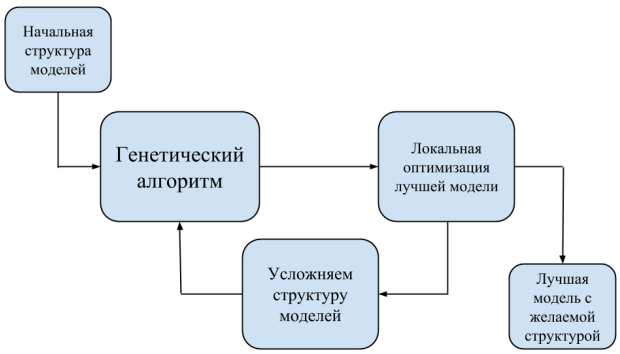


Рисунок 4 – Схема основного алгоритма.

определенный порядок параметров.

2.2. Основной алгоритм

Опишем основной разработанный алгоритм для поиска демографической модели по аллель-частотному спектру. Предполагается, что у нас известны некая начальная и конечная — желаемая, структуры модели, которые определяют число параметров, а следовательно и сложность модели.

Сначала подбираем демографическую модель простой начальной структуры с помощью генетического алгоритма и локальных оптимизаций, затем усложняем ее, деля один из периодов времени на два, и запускаем генетический алгоритм и локальную оптимизацию на этих моделях с новой более сложной структурой. Делаем так до тех пор, пока не достигнем желаемой структуры. Такой алгоритм позволяет подобрать сначала более простую модель и потом усложнить ее до более подробной. Если попробовать подбирать сразу сложную модель, то могут возникнуть сложности из-за большого числа параметров. Также в генетическом алгоритме фиксируется структура модели и число ее параметров, что значительно упрощает операцию скрещивания. Между генетическим алгоритмом и усложнением модели есть еще один шаг: попытка улучшить лучшую модель генетического алгоритма с помощью локального поиска, который обеспечивает более аккуратную оптимизацию параметров. Схема основного алгоритма представлена на рисунке 4, а псевдокод в листинге 1.

Чтобы избежать переобучения в предложенный метод можно добавить сравнение моделей по функциям BIC или AIC. Очевидно, что достаточно сравнивать лишь финальные для каждой структуры модели, так как количество параметров между увеличениями структуры не меняется, а значит и значения BIC и AIC зависят только от значений правдоподобия. В реализации метода был выбран BIC, так как он накладывает больший штраф на число параметров, и в случае, если модели с лучшими значениями правдоподобия и BIC не совпали, пользователю сообщается о

переобучении.

Листинг 1 – Основной алгоритм поиска демографической модели структуры S^F по аллель-частотному спектру A.

```
1: function FindDemographicModel(S^I, S^F, A, C, P_GA)
2:     S* ← S^I                                  ▷ Текущая структура моделей
                  ▷ Строим множество случайных моделей начальной структуры
3:     N ← GetSizeOfGenerationInGA(P_GA)
4:     X ← { GenerateRandomModel((S*) }_{i=1}^N
5:     while ∃k :  s*_k ≤ s^F_k do          ▷ Пока не достигнем финальной структуры
6:         X ← GeneticAlgorithm(X, A, C, P_GA)
7:         X[0] ← LocalSearch(X[0], A)
8:         S*, X ← IncreaseStructure(X, S^F)
9:     end while
10:    return X[0]
11: end function
```

2.3. Генетический алгоритм

Генетический алгоритм — один из наиболее эффективных эвристических методов. Он основан на принципах эволюции, что влечет коллизию обозначений с нашей задачей. Основной идеей является симуляция естественного отбора, который оставляет самую адаптированную особь. Целью алгоритма является поиск решения задачи, которое имеет максимальное или минимальное значение целевой функции. В начале создается множество фиксированного размера случайных решений, которые называются **особями**, а само множество **поколением**. Каждой особи присваивается значение ее **приспособленности**, которое определяется значением целевой функции. После этого итеративно строятся новые поколения с помощью мутаций, скрещивания и отбора наиболее приспособленных особей. Все эти операции могут быть как детерминированными, так и случайными, а их порядок может меняться в разных реализациях. В нашем случае особи — это демографические модели одинаковой структуры, а функция приспособленности — логарифм правдоподобия $\log(\mathcal{L}(M|S))$, который был описан выше.

Для формирования нового поколения демографических моделей выбираются несколько наиболее адаптированных моделей из предыдущего, несколько мутированных, несколько скрещенных и несколько случайных. Чтобы определить наиболее адаптированные модели, они сортируются по значению целевой функции. Выбор моделей для мутации или скрещивания случайный, но вероятность выбора прямо пропорциональна значению адаптации: чем лучше модель, тем вероятнее ее выбрать. Генетический алгоритм останавливается, когда больше не может получить более хорошую по значению целевой функции демографическую модель в течение

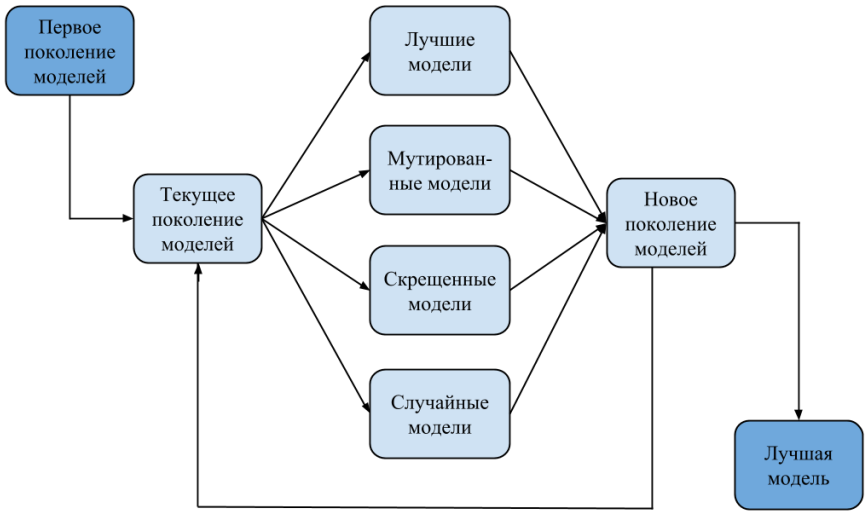


Рисунок 5 – Схема генетического алгоритма.

нескольких итераций. Схема генетического алгоритма представлена на рисунке 5, а псевдокод можно найти в листинге 2.

2.3.1. Мутация демографической модели

Мутация демографической модели — это изменение некоторого числа ее параметров. У нас имеется два параметра алгоритма: так называемые, степень и сила мутации. Число параметров для изменения случайно выбирается из биномиального распределения со средним, равным силе мутации. А то, насколько изменится каждый параметр, определяется знаком ($+1$ или -1 равновероятно) и долей изменения, которая случайно выбирается из нормального распределения со средним, равным степени мутации, и дисперсией, равной половине среднего.

Вероятность выбора параметров прямо пропорциональна их весам, которые в начальный момент алгоритма равны (то есть выбор равновероятен), и затем увеличиваются в том случае, если произошла мутация соответствующего параметра, которая привела к улучшению модели. Графическое представление мутации одного параметра модели представлено на рисунке 6, а псевдокод в листинге 3.

2.3.2. Адаптивные степень и сила мутации

На начальных итерациях сильные изменения большого числа параметров, гораздо эффективнее малых изменений малого числа, тогда как при приближении к оптимальному решению наоборот. Поэтому степень и сила мутации могут быть адаптивными, то есть меняется в течение работы алгоритма. Существует несколько способов сделать изменчивую величину, одним из самых популярных является алгоритм «одной пятой» [31]. Применим его для степени мутации: на каждой итерации, если получили «успешное» решение, то есть оно стало лучше после

Листинг 2 – Генетический алгоритм для поиска демографической модели фиксированной структуры.

```
1: function GeneticAlgorithm(X_0, A, C, P_GA)
2:     S* =GetCurStructure(P_GA))
3:     SetInitialValues(P_GA)
4:     X ← SortByf(X_0, S*, A, C)                          ▷ Первое поколение моделей
5:     N ← GetSizeOfGenerationInGA(P_GA)
6:     N_B ← GetNumberOfBestModels(P_GA)
7:     N_M ← GetNumberOfMutatedModels(P_GA)
8:     N_C ← GetNumberOfCrossedModels(P_GA)
9:     N_R ← N − N_B − N_M − N_C
10:    T ← 0                                               ▷ Число итераций без улучшения
11:    T_max ← GetMaxItWithoutChanges(P_GA)
12:    while T ≤ T_max do
13:        X' ← [ ]                                        ▷ Строим новое поколение моделей
14:        ω ← {f(x^(i), S*, A, C)}_{i=1}^N                ▷ Веса для дискретного распределения
15:        for k ← 1..N_B do                               ▷ Добавляем лучшие модели
16:            X'.Add(X[k])
17:        end for
18:        for k ← 1..N_M do                               ▷ Добавляем мутированные модели
19:            j ← DiscreteRandom({i}_{i=1}^N, ω)
20:            X'.Add(Mutate(X[j], P_GA))
21:        end for
22:        for k ← 1..N_C do                               ▷ Добавляем скрещенные модели
23:            j_1 ← DiscreteRandom({i}_{i=1}^N, ω)
24:            j_2 ← DiscreteRandom({i}_{i=1}^N, ω)
25:            X'.Add(Cross(X[j_1], X[j_2], P_GA))
26:        end for
27:        for k ← 1..N_R do                               ▷ Добавляем случайные модели
28:            X'.Add(GenerateRandomModel(S*))
29:        end for
30:        X' ←SortByf(X', S*, A, C)
31:        ε ← GetEpsilon(P_GA)
32:        if |f(X'[0], S*, A, C) − f(X[0], S*, A, C)| ≤ ε then
33:            T ← T + 1
34:            if X'[0].lastOperation == "Mutation" then
35:                b ← True
36:            else
37:                b ← False
38:            end if
39:        else
40:            T ← 0
41:            b ← False
42:        end if
43:        UpdateValue(b, P_GA, ''strength'')               ▷ Адаптивная сила мутации
44:        X ← X'
45:    end while
46:    return X
47: end function
```

Листинг 3 – Процедура мутации демографической модели для генетического алгоритма.

```
1: function Mutate(x, P_GA)
2:     μ_s ← GetCurMeanMutStrength(P_GA)   ▷ Число параметров для изменения
3:     μ_r ← GetCurMeanMutRate(P_GA)
4:     ω ← GetParamsWeights(x)
5:     L ← GetNumberOfParams(x)
6:     l ← BinomialRandom(μ_s, L)          ▷ Число параметров для изменения
7:     inds ← DiscreteRandom({i}_{i=1}^{L}, ω, size = l)   ▷ Индексы параметров для
   изменения
8:     x' =Copy(x)
9:     for i ∈ inds do                     ▷ Изменяем каждый параметр
10:        σ ← UniformRandom({1, −1})
11:        λ ← TruncNormRandom(μ_r, μ_r/2, 0.0, 1.0)
12:        x'.params[i] ← (1 + σ · λ) · x'.params[i]
13:    end for
14:    if f(x', S*, A, C) − f(x, S*, A, C) ≤ ϵ then
15:        b ← True
16:    else
17:        b ← False
18:    end if
19:    UpdateValue(b, P_GA, "rate")        ▷ Адаптивная степень мутации
20:    x.lastOperation ←"Mutation"
21:    return x'
22: end function
```

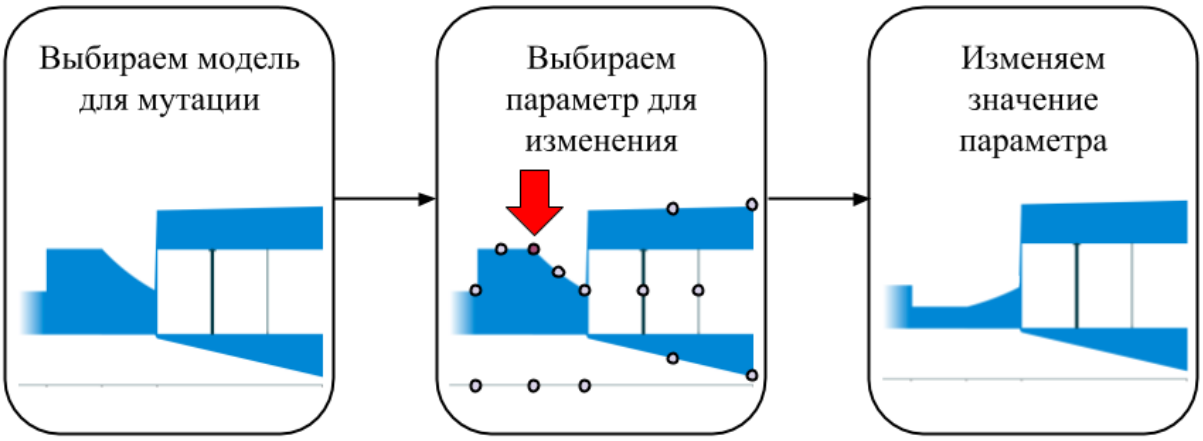


Рисунок 6 – Мутация одного параметра демографической модели.

мутации, то умножаем степень мутации на константу $C \in [1, 2]$, если нет, то делим на корень четвертой степени из C, уменьшая степень. В случае силы мутации для «успешности» решения необходимо дополнительно проверить стало ли решение лучшим за все время работы алгоритма.

Таким образом часто получая новое лучшее решение с помощью мутации, что происходит в начале работы генетического алгоритма, мы будем увеличивать число параметров, которые меняются при мутации, и степень того насколько они меняются, а затем с приближением к оптимуму и уменьшением частоты, они будут снижаться и приводить к более аккуратному поиску. Увеличение числа параметров для изменения приводит к более эффективному скрещиванию. В то же время из более ослабленных условий успешности решения для адаптивной степени мутации следует ее более частые изменения, по сравнению в силой мутации, что делает мутацию более эффективной. Псевдокод функции обновления значений силы и степени мутаций представлен в листинге 4.

2.3.3. Скрещивание демографических моделей

Чтобы скрестить две демографические модели, мы представим обе модели как последовательность параметров и получим их потомка, выбирая каждый его параметр случайным образом равновероятно от одного, либо второго предка. Так как структура моделей не изменяется в течение работы генетического алгоритма, то число параметров у всех моделей будет одинаковое и их можно однозначно интерпретировать по ним и легко скрещивать таким образом.

Псевдокод соответствующей функции представлен в листинге 5. Графическое изображение скрещивания двух моделей можно найти на рисунке 7.

Листинг 4 – Функция обновления силы и степени мутации по алгоритму «одной пятой».

```
1: function UpdateValue(b, P_GA, s)
2:     if s ==“rate” then
3:         μ ← GetCurMeanMutRate(P_GA))
4:         c ← GetConstOfMeanMutRate(P_GA))
5:     else// s ==“strength”
6:         μ ← GetCurMeanMutStrength(P_GA))
7:         c ← GetConstOfMeanMutStrength(P_GA))
8:     end if
9:     if b == True then                          ▷ Правило «одной пятой»
10:        μ ← c · μ
11:    else
12:        μ ← 1/c^0.25 · μ
13:    end if
14:    if s == “rate” then
15:        SetConstOfMeanMutRate(P_GA, μ)
16:    else// s ==“strength”
17:        SetConstOfMeanMutStrength(P_GA, μ)
18:    end if
19: end function
```

Листинг 5 – Процедура скрещивания демографических моделей для генетического алгоритма.

```
1: function Cross(x_1, x_2, P_GA)
2:     L ← GetNumberOfParams(x)
3:     B ← UniformRandom({False, True}^L)
4:     x ← Copy(x_1)
5:     for i ← 1..L do          ▷ Берем нужные параметры от второго родителя
6:         if B[i] then
7:             x.params[i] = x_2.params[i]
8:         end if
9:     end for
10:    x.lastOperation ← “Cross”
11:    return x
12: end function
```

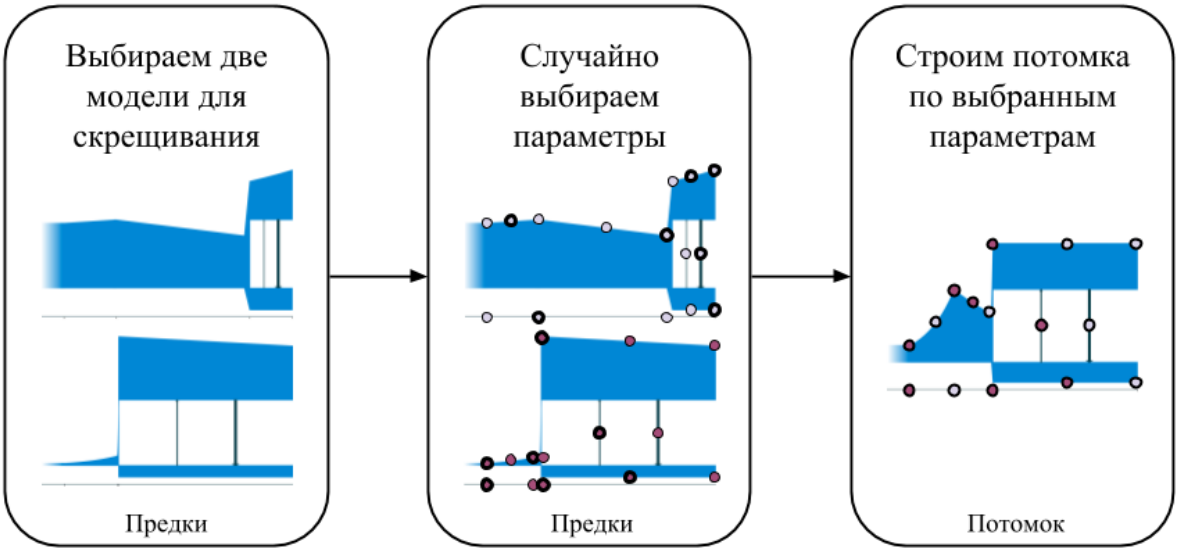


Рисунок 7 – Скрещивание двух демографической модели.

2.4. Локальные оптимизации

Локальные оптимизации отличаются эффективностью при условии, что начальное решение близко к оптимальному. Они более аккуратно подбирают параметры, чем генетический алгоритм, и могут существенно улучшить его результат.

∂a∂i, а с ним и *moments*, предоставляют следующий выбор алгоритмов локальной оптимизации:

- Алгоритм Бройдена — Флетчера — Гольдфарба — Шанно (BFGS).
- L-BFGS-B — модификация BFGS алгоритма, которая лучше работает, когда оптимальные параметры близки к граничным.
- Метод Нелдера—Мида или симплекс-метод.
- Метод Пауэлла.

Метод Пауэлла был предложен авторами *moments* и был отмечен как наиболее эффективный, поэтому он был адаптирован для использования с *∂a∂i* и для наших экспериментальных исследований был выбран в качестве алгоритма локального поиска. В реализации нашего метода пользователю предоставляется возможность выбора алгоритма локальной оптимизации.

2.5. Усложнение структуры демографической модели

Нам необходимо уметь усложнять структуру демографической модели для достижения нужной. Для этого выберем период и разделим его на два, разделив

посередине его времени. Период выбираем случайно из расчета на то, чтобы новая структура не стала больше требуемой по одному из значений. Величины параметров полученных периодов вычисляются по родительскому: численности популяций первого равна численностям посередине времени родителя, а второго равна численностям в конце, время обоих будет равно половине времени родителя, миграции и законы изменения численности останутся прежними. Тогда по своей сути модель не изменится, появится больше параметров, а вместе с тем не поменяется и величина правдоподобия. Псевдокод функции усложнения структуры представлен в листинге 6.

2.6. Выводы по разделу «Предлагаемый метод поиска демографической модели по аллель-частотному спектру»

В данном разделе описан основной разработанный алгоритм. Он итеративный и сначала подбирает модель более простой структуры, затем усложняет ее для получения более подробной. Также в разделе присутствует описание генетического алгоритма, который является фундаментом основного метода, а вместе с ним и операций мутации и скрещивания для демографических моделей.

В отличие от существующих решений разработанный метод использует алгоритм глобальной оптимизации — генетический алгоритм, и не ограничен закрепленной пользователем моделью. Он подходит для поиска большого числа параметров и позволяет искать демографическую модель определенной структуры со всеми доступными параметрами, включая законы изменения численности, которые могут быть внезапными, линейными и экспоненциальными. Последние два больше соответствуют реальным функциям изменения размера популяций. Таким образом предложенный алгоритм позволяет автоматически подбирать демографическую модель для аллель-частотного спектра, используя единственное несильное ограничение — структуру.

Листинг 6 – Алгоритм усложнения структуры набора демографических моделей одной структуры.

```
1: function IncreaseStructure(X, S^F)
2:     S* ← GetCurStructure(X)
3:     D ← {S^F[i] − S*[i]}_{i=1}^{P}                ▷ Ищем где можно увеличить структуру
4:     Q ← DiscreteRandom({i}_{i=1}^{P}, D)
5:     d ← Σ_{i=1}^{Q} S*[i] + UniformRandom({i}_{i=1}^{S*[Q]})   ▷ Выбираем индекс периода
6:     S*[Q] ← S*[Q] + 1                                ▷ Новая структура
7:     for x ∈ X do          ▷ Если первый период, то добавляем новый после него
8:         p ← x.timePeriods[d]
9:         if d == 1 then
10:            t ← UniformRandom(0.0, x.maxTime)
11:            N_A ← GetSizesOfPops(p)
12:            p_new ← timePeriod(t, N_A,"sudden")
13:            x.insertTimePeriod(p_new, d + 1)
14:        else
15:            p_pred ← x.timePeriods[d − 1]
16:            p_new ← Copy(p)
17:            N ← Array(size = Q)      ▷ Вычисляем размеры популяций посередине
   периода p
18:            g =getGrowthTypes(x.timePeriods[d])
19:            for k ← 1..Q do
20:                if g[k] ==“sudden” then
21:                    N[k] ← GetSizesOfPops(p)
22:                else
23:                    t ← GetTime(p)
24:                    a ← GetSizesOfPops(p_pred)[k]
25:                    a ← GetSizesOfPops(p)[k]
26:                    if g[k] == ”linear” then
27:                        N[k] ← (a+b)/2
28:                    else//g[k] == ”exponential”
29:                        N[k] ← √(a · b)
30:                    end if
31:                end if
32:            end for
33:            SetSizesOfPops(p_new, N)
34:            t_new = p.GetTime/2
35:            SetTime(p, t_new)    ▷ Уменьшаем время у выбранного периода пополам
36:            SetTime(p_new, t_new)
37:            x.insertTimePeriod(p_new, d)          ▷ Добавляем новый период перед p
38:        end if
39:    end for
40:    return S*, X
41: end function
```

3. Экспериментальные исследования

Все данные, параметры и результаты запусков доступны в репозитории по адресу: `https://bitbucket.org/noscode/gadma_results/src/master/`

3.1. Популяции современных людей

Демографическая история популяций людей вызывает большой интерес. Довольно часто рассматривают демографическую модель «выхода людей из Африки» для следующих трех популяций [16, 15, 4, 11]:

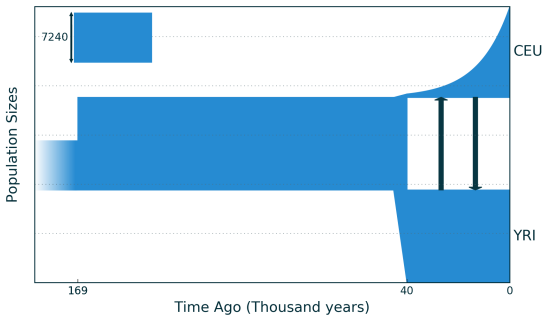
- YRI — люди народа Йоруба из города Ибадан, Нигерия,
- CEU — жители штата Юта с предками из северной и западной Европы,
- CHB — китайцы народа Хань из Пекина, Китай.

Геномы для этих популяций доступны в нескольких базах данных публичных проектов, например, Environmental Genome Project [33] или 1000 Genome Project [3].

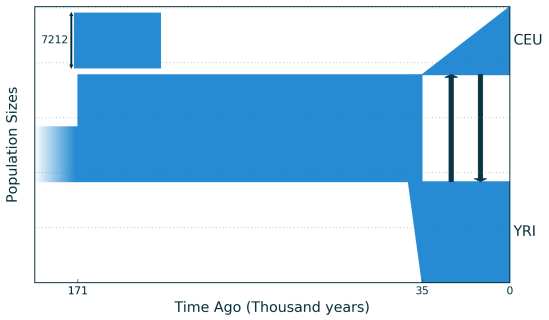
Для выявления эффективности метода был выбран аллель-частотный спектр из статьи 2009 года Райана Гутенкунста и др. [16], в которой описывается $\partial a \partial i$ и представлены демографические модели для двух (YRI, CEU) и трех популяций, выведенные из этого спектра. Однако эти модели, которые также изображены на рисунках 8a и 9a были получены в результате большого числа запусков локальной оптимизации и также имеют ряд ограничений на число параметров: размер популяции YRI не меняется после первого роста численности предковой популяции, миграции симметричны, законы изменения численности зафиксированы внезапными, кроме последнего временного периода для CEU, CHB, где происходит экспоненциальный рост. Всего у модели для двух популяций 6 параметров, а для трех — 14, которые приведены в таблицах 1 и 3.

В статье Р. Гуненкунста и др. [16] аллель-частотный спектр размера $21 \times 21 \times 21$ был построен на базе данных Environmental Genome Project [33] (секвенирование по Сенгеру). Были рассмотрены все биаллельные SNV некодирующих регионов 219 генов (общей длины 5.01 Mb). Эффективная длина используемой последовательности, то есть после фильтрования неподходящих вариантов и учета ошибок выравнивания, получилась равной $L = 4.04 \times 10^6$. В статье было отмечено, что данные по вариантам согласуются в результатах секвенирования следующего поколения и секвенировани по Сенгеру на 99.5%, поэтому никакой корректировки для учета ошибок секвенирования не было проведено [16].

Для оценки нейтральной скорости мутации авторы статьи использовали информацию о разделении людей и шимпанзе. Они выбрали время жизни одного



(а) Модель из дополнительных материалов статьи [16].



(б) Полученная модель (модель 3) с лучшим значением правдоподобия.

Рисунок 8 – Демографические модели для популяций YRI, CEU. **(а)** существующая демографическая модель; **(б)** Модель 3 с 12 параметрами. Модели 1 и 2 не представлены, так как их изображения не сильно отличаются от (а).

поколения людей равным $t_g = 25$ годам и оценили скорость мутации как $\mu = 2.35 \times 10^{-8}$. Мы также взяли эти значения для нашего анализа и вычислили величину средней частоты возникновения мутации в одной особи за поколение (подробнее в Обзоре предметной области) как $\theta_0 = 4\mu L = 0.37976$.

Размер поколения демографических моделей генетического алгоритмы был выбран равным 10, сила и степень мутации, равными 0.2, а пропорции лучших, мутированных, скрещенных и случайных моделей в новом поколении, равными 0.2 : 0.3 : 0.3 : 0.2. Сила и степень мутации были адаптивными с константами 1.05 и 1.02 соответственно. Аллель-частотный спектр симулировался с помощью $\partial a \partial i$ с размером сетки $G = \{40, 50, 60\}$, величина значения правдоподобия при сравнении считалась значимой до 2го знака после запятой и генетический алгоритм останавливался после 100 итераций без улучшений. В качестве локального поиска был выбран метод Пауэлла. Финальные структуры моделей соответствовали структурам моделей из статьи и были равны начальным, кроме тех случаев где об этом явно сказано.

3.1.1. Две популяции

Дополнительные материалы к статье [16] предоставляют вывод демографической модели для двух популяций: YRI и CEU. Там же приведены параметры с максимальным значением правдоподобия, полученные для этой модели, которые можно найти в таблице 1. Были выведены три демографические модели по тому же спектру: с такими же параметрами, что и в статье (модель 1) и две со всеми возможными двенадцатью параметрами (модели 2, 3). Модели 2 и 3 отличались начальной структурой: модель 2 имела структуру (1,1), которая затем расширялась до (2,1) в процессе работы алгоритма, и модель 3 имела сразу равную финальной — (2,1). Метод был запущен 10 раз для каждой модели из трех и значения лучших параметров

приведены в таблице 1, а некоторые модели изображены на рисунке 8. Все три модели имеют правдоподобие лучше, чем у существующей, параметры первой и второй несущественно отличаются, а третья имеет самое лучшее значение правдоподобия. Модель 3 показывает более низкую численность популяции европейцев, более высокий темп их роста (с 25 особей до 9 тысяч) и меньшее время разделения, чем модель из статьи и модели 1, 2. Миграции подбирались несимметричными, однако они практически равны между собой и совпадают со значениями между моделями.

Для демонстрации неэффективности методов локальной оптимизации для модели из статьи [16] был запущен 100 раз один из предложенных $\partial a \partial$i методов — BFGS. В каждом запуске начальные значения параметров выбирались случайным образом. Самое лучшее значение логарифма правдоподобия было равно -1629.24, что, к сожалению, довольно далеко от оптимального -1066.35, предложенного статьей. Среднее время одного запуска оптимизации составило около 21 минуты.

Для сравнения запусков с различными начальными структурами были вычислены некоторые характеристики, которые приведены в таблице 2. Запуски с простой начальной структурой, показывают более стабильный по значению правдоподобия результат, однако имеют более долгое время работы. Также все полученные модели для запусков с простой структурой имеют те же законы изменения численности и близкие параметры, что и существующая модель из статьи, что неверно для результата поиска сразу сложных. В то же время, хоть запуски со сложной структурой дали модель с лучшим правдоподобием, но она отличается от остальных по параметрам, что возможно говорит о ее несостоятельности.

Таблица 1 – Параметры для различных демографических моделей с максимальным значением правдоподобия для двух популяций современных людей. Темпы миграций указаны на одно поколение.
[2]Линейный рост

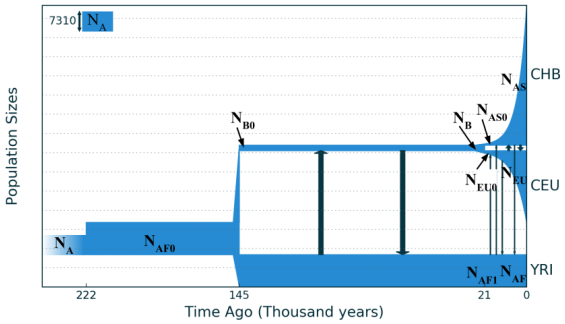
	Модель из доп. материалов статьи [16]	Модель, полученная GADMA (1)	Модель, полученная GADMA (2)	Модель, полученная GADMA (3)
Число параметров:	6	6	12	12
Max log likelihood:	−1066.35	−1066.28	−1065.87	**−1065.15**
AIC:	2144.70	**2144.56**	2155.74	2154.30
BIC:	2168.65	**2168.51**	2203.64	2202.20
Значения параметров:				
N_A	7240	7230	7200	7210
N_{AF0}	13620	13580	13400	14000
N_{EU0}	515	530	560	25
N_{EU}	13360	12400	12100	8950[2]
N_{AF}	$(=N_{AF0})$	$(=N_{AF})$	13500	13300
$m_{AF-EU}(\times 10^{-5})$	25	25	28.4	23.2
$m_{EU-AF}(\times 10^{-5})$	$(=m_{AF-EU})$	$(=m_{AF-EU})$	24.4	24.2
T_{AF} (kya)	168.5	171.5	176.5	171.1
T_{AF-EU} (kya)	40	40.8	42.3	34.8

Таблица 2 – Сравнение запусков с разными начальными структурами: более простой (10 запусков) и сразу сложной (10 запусков) при поиске демографических моделей двух популяций.
$\log LL$ — логарифм правдоподобия.

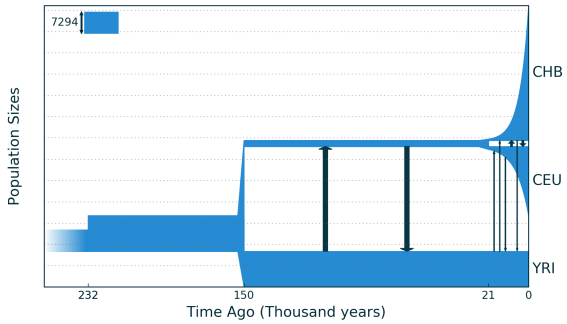
	Начальная структура	Среднее время итерации, с	Среднее число итераций	Среднее значение $\log LL$	Среднее отклонение $\log LL$
Модель 2	(1,1)	4.06	2938	**−1066.39**	**0.22**
Модель 3	(2,1)	**3.77**	**1400**	−1071.16	14.35

3.1.2. Три популяции

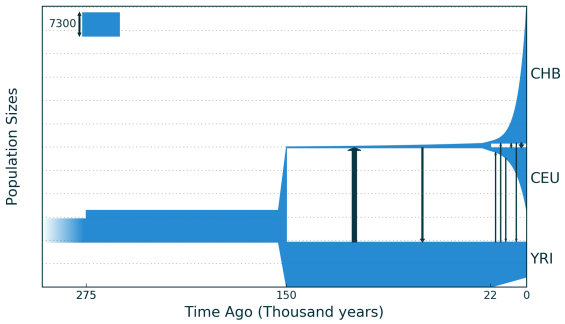
Были найдены различные демографические модели и для трех популяций из аллель-частотного спектра, полученного в работе [16]. Первая модель (модель 1) имела те же параметры, что в статье, и были получены их значения, представленные в таблице 3 и дающие лучшее значение правдоподобия, чем существующие. Однако время выхода людей из Африки, то есть разделение африканской популяции (YRI) и евразийской (общей популяции CEU и CHB), было отодвинуто на 400 тысяч лет назад, что не подтверждается никакими другими исследованиями. Поэтому были использованы экспертные данные, которые ограничили это время 150 тыс. лет назад. Была подобрана демографическая модель (модель 2) с данным ограничением и она также имеет лучшее значение правдоподобия, чем в статье. Далее была выведена



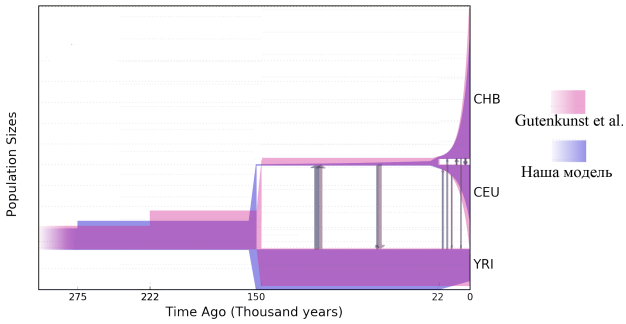
(а) Демографическая модель из статьи Р. Гутенкунста и др. [16] (14 параметров).



(б) Демографическая модель, полученная с помощью GADMA (14 параметров)*.



(в) Демографическая модель, полученная с помощью GADMA (26 параметров)*.



(г) Сравнение модели из статьи [16] и нашей модели с 26 параметрами.

Рисунок 9 – Демографические модели для популяций YRI, CEU и CHB. **(а)** существующая демографическая модель; **(б)** Модель 1 с теми же параметрами, полученная нашим методом*; **(в)** Модель 2 без ограничений на параметры*; **(г)** сравнение существующей модели и последней полученной нашим методом. *__(б)__ и __(в)__ имели ограничение на время первого разделения (не более 150 тыс. лет назад).

модель со всеми доступными параметрами (26 штук). Тенденция раннего выхода из Африки также сохранилась (модель не приведена), поэтому экспертные данные были применены и тут и получена модель (модель 3) не только с самой лучшей величиной правдоподобия, но и с лучшим значением AIC.

Параметры всех полученных моделей приведены в таблице 3, а модели изображены на рисунке 9.

Как и в случае двух популяций, был протестирован алгоритм локальной оптимизации $\partial a \partial i$— BFGS. Он был запущен 100 раз из случайно выбранных начальных параметров и самое лучшее значение логарифма правдоподобия получилось равным -6323.99, что все равно далеко от оптимального -6326.89, хотя и ближе чем получилось у двух популяций. Этот пример показывает сложность поиска параметров с помощью существующих алгоритмов оптимизации.

Значения правдоподобия у существующей модели и моделях 1,2 близки между собой, как и их параметры, не считая времени разделения у модели 1. В модели 3,

Таблица 3 – Параметры для различных демографических моделей с максимальным значением правдоподобия для трех популяций. Темпы миграций указаны на одно поколение.
[2]Линейный рост
[3]Экспоненциальный рост

	Модель из статьи [16]	Модель, полученная GADMA (1)	Модель, полученная GADMA (2)	Модель, полученная GADMA (3)
Число параметров:	13	13	14	26
Max log likelihood:	−6316.89	−6314.40	−6315.86	**−6288.36**
AIC:	12659.78	12654.80	12657.72	**12628.72**
BIC:	12750.61	**12745.63**	12748.55	12806.00
Значения параметров:				
N_A	7300	6000	7300	7300
N_{AF0}	12300	11840	12200	9900
N_{B0}	2100	2050	2070	280
N_B	$(=N_{B0})$	$(=N_{B0})$	$(=N_{B0})$	1450[3]
N_{AF1}	$(=N_{AF0})$	$(=N_{AF0})$	$(=N_{AF0})$	14000
N_{AF}	$(=N_{AF0})$	$(=N_{AF0})$	$(=N_{AF0})$	11000[2]
N_{EU0}	1000	930	950	890
N_{EU}	27300	24700	23700	19600
N_{AS0}	510	500	510	560
N_{AS}	53200	50000	46200	42200
m_{AF-B} $(\times 10^{-5})$	25	26.2	25.4	68
m_{B-AF} $(\times 10^{-5})$	$(=m_{AF-B})$	$(=m_{AF-B})$	$(=m_{AF-B})$	230
m_{AF-EU} $(\times 10^{-5})$	3.0	3.1	3.1	17.8
m_{EU-AF} $(\times 10^{-5})$	$(=m_{AF-EU})$	$(=m_{AF-EU})$	$(=m_{AF-EU})$	11.3
m_{AF-AS} $(\times 10^{-5})$	1.9	1.9	2.0	3.7
m_{AS-AF} $(\times 10^{-5})$	$(=m_{AF-AS})$	$(=m_{AF-AS})$	$(=m_{AF-AS})$	8.3
m_{EU-AS} $(\times 10^{-5})$	9.6	10.2	10.2	75.4
m_{AS-EU} $(\times 10^{-5})$	$(=m_{AS-EU})$	$(=m_{AS-EU})$	$(=m_{AS-EU})$	27.4
T_{AF} (kya)	220	570	232	274.8
T_B (kya)	140	400	150	149.8
T_{EU-AS} (kya)	21.2	21.0	21.1	22.4

которая имеет 26 параметров и самое лучшее значение AIC, некоторые параметры также довольно похожи на значения в других моделях. Исключения составляют темпы миграций и численность евразийской популяции, которая экспоненциально растет с 200 особей до 1500 после разделения предковой популяции. Для сравнения в других моделях эта численность — константа, равная 2000 особям, что выглядит менее естественным, чем экспоненциальный рост. Довольно существенно отличаются темпы миграций: они более высокие. Модель показывает, что самые большие миграции происходили между африканской и евразийской популяциями, а после разделения последней между европейской и азиатской. Более того, чем популяции географически дальше, тем темпы миграций меньше, что было выражено и в других моделях, но не

так ярко.

Алгоритм был запущен 10 раз для каждой из трех моделей и в таблице 3 представлены лучшие решения. Также были проведены запуски (также 10 штук) с использованием *moments* с целью сравнить его эффективность с ∂a∂i. Авторы *moments* в своей статье делали аналогичное сравнение на симулированных данных, которые показывают неоднозначные результаты. По результатам 20 запусков (10 с ∂a∂i, 10 с *moments*) в таблице 4 представлены различные характеристики времени и стабильности результата по значению правдоподобия. Для корректного сравнения величин правдоподобия, у полученных с использованием *moments* моделей они были пересчитаны для ∂a∂i с сеткой $G = \{40, 50, 60\}$. *moments* показал свою скорость: быстрее ∂a∂i почти в 6 раз, и аккуратность: среднее и дисперсия правдоподобия полученных моделей меньше.

Таблица 4 – Сравнение запусков с использованием ∂a∂i и *moments* при поиске демографических моделей трех популяций с 26 параметрами. Значение логарифма правдоподобия считалось с использованием ∂a∂i с сеткой $G = \{40, 50, 60\}$.
$\log LL$ — логарифм правдоподобия.

	Среднее время итерации, с	Среднее число итераций	Среднее значение $\log LL$	Среднее отклонение $\log LL$
∂a∂i	105.90	6000	−6310.82	31.39
moments	**18.44**	**4700**	**−6305.21**	**12.72**

3.2. Бабочки *Euphydryas gillettii*

В 2013 году была опубликована статья Раджива Маккоя и др. [9], посвященная демографической истории бабочек *Euphydryas gillettii* с помощью ∂a∂i. У этих бабочек существует естественная популяция, обитающая на лугах восточного склона северных Скалистых гор в штате Вайоминг (WY), однако в 1977 году (за 33 года до взятия образцов) часть особей была отловлена и была намеренно создана другая популяция на полях биологической лаборатории Скалистых гор в городе Готик штата Колорадо (CO), численность которой затем проверяли каждый год. Рисунок 10 взят из статьи [9] и показывает историю популяций и замеры численности бабочек из Колорадо, откуда видно довольно сильные колебания размера популяции, включающие сильный ботлнек в 25 особей. Две популяции были отделены засушливым Бассейном Большого Разрыва, а бабочки очень чувствительны к условиям среды обитания, что исключает поток генов между популяциями.

Для построения аллель-частотного спектра авторы отсеквенировали РНК восьми особей популяции в штате Вайоминг (WY) и восьми особей популяции штата Колорадо (CO) и нашли единичные нуклеотидные варианты (SNV). На основе

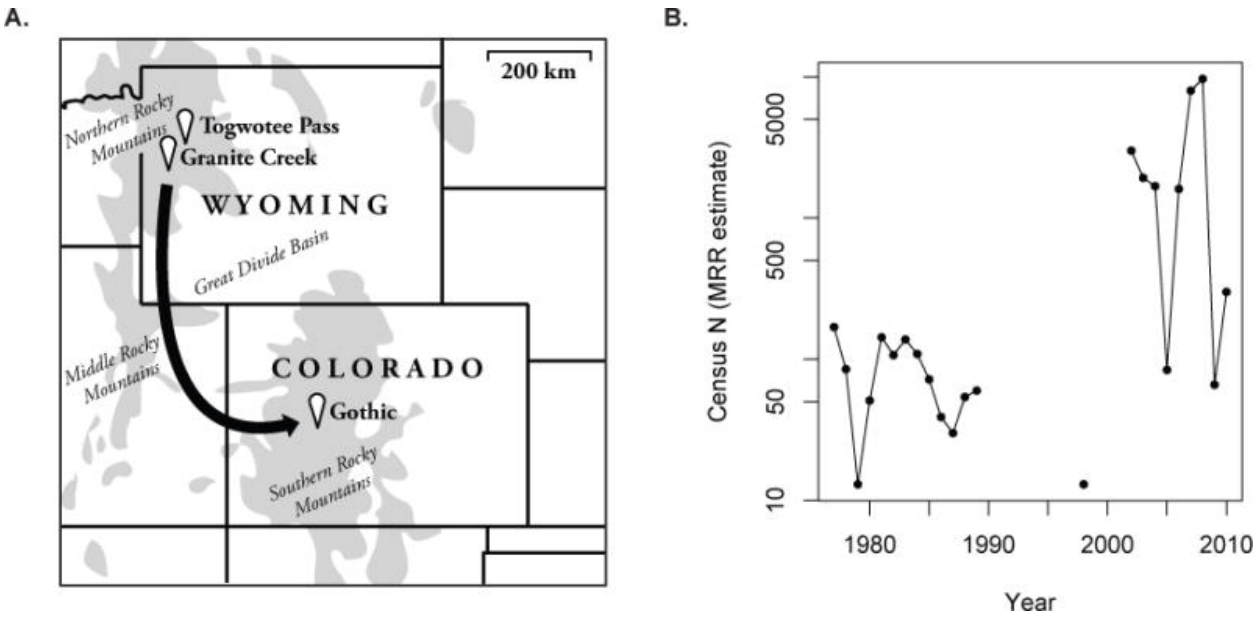


Рисунок 10 – Рисунок из статьи [9] описывающий историю популяций бабочек из штатов Вайоминг и Колорадо. (A) Карта местности и перемещение бабочек из естесственной популяции в Вайоминге в биологическую лаболаторию штата Колорадо для формирования новой популяции. (B) Mark-release-recapture оценка численности популяции бабочек в штате Колорадо. Средняя численность особей по этим оценкам была равна 34 особям. Ось ординат представлена в логарифмической шкале, чтобы лучше показать колебания на маленьких численностях.

найденных вариантов были построены два аллель-частотных спектра размера 13×13: по синонимичным заменам и по всем: синонимичным, несинонимичным и нуклеотидным вариантам из нетранслируемых областей.

Были рассмотрены три демографические модели для аллель-частотного спектра, построенного из синонимичных замен: без миграций (тип A), с однонаправленной мутацией из CO в WY (тип B1) и двунаправленной мутацией (тип B2). Для всех трех моделей, по одной каждого типа, было вычислено значение AIC, которое было минимальным для типа модели без миграций (A), поэтому она была выбрана для аллель-частотного спектра из всех SNV и были найдены ее параметры дающие на нем максимальное правдоподобие. Без учета миграций модели имели следующую структуру: была одна популяция размера N_A, которая в какой-то момент времени разделилась на две, численность которых не менялась в дальнейшем (внезапный закон изменения численности). Все параметры вычислялись относительно N_A, которая считалась равной 1 и вычислялась подгонкой параметра θ_0, и имели следующие обозначения: η_{WY}, η_{CO} — относительные численности популяций на данный момент; τ_{SPLIT} — время разделения; M_{WY-CO}, M_{CO-WY} — отмасштабированные темпы миграций. Для наших моделей в таблице приведен еще один параметр η_{WY0} — доля численности предковой популяции, которая формирует популяцию WY, популяцию CO соответственно формирует доля равная $1 - \eta_{WY0}$.

Однако в случае, когда закон изменения популяции внезапный, численность после разделения будет равна численности в настоящий момент времени.

Все данные по SNV для аллель-частотных спектров и скрипты для поиска параметров моделей лежат в открытом доступе в репозитории авторов статьи и могут быть найдены по ссылке: `https://github.com/rmccoy7541/egillettii-rnaseq/tree/master/data`.

Для поиска демографических моделей бабочек с помощью нашего метода и сравнения с существующими моделями, мы взяли те же два аллель-частотных спектра, построенных из всех замен и только из синонимичных, что в статье [9]. Были рассмотрены модели с такой же структурой — (1,1), но с поиском всех доступных параметров. Таким образом добавились еще 5 параметров: размер предковой популяции, отношение в котором делилась ее численность при разделении и законы изменения численности (3 штуки). Для симулирования аллель-частотного спектра при поиске был использован *moments*, который показал свою скорость на предыдущих данных, однако для вычисления значения правдоподобия был взят и ∂a∂i с размерами сеток $G = \{32, 42, 52\}$ для корректного сравнения с существующими моделями.

Размер поколения демографических моделей в генетическом алгоритме был выбран равным 10, сила и степень мутации, равными 0.2 с константами 1.0 и 1.02 соответственно, а пропорции лучших, мутированных, скрещенных и случайных моделей в новом поколении, равными 0.2 : 0.3 : 0.3 : 0.2. Величина значения правдоподобия при сравнении считалась значимой до второго знака после запятой и генетический алгоритм останавливался после 100 итераций без улучшений. В качестве локального поиска также был выбран метод Пауэлла. Так как структура модели из статьи соответствует простейшей структуре (1,1), она была выбрана в качестве начальной и конечной структур.

Для спектра, построенного по синонимичным заменам, для которого были найдены параметры трех моделей типов A, B1, B2, были запущены поиски параметров моделей с миграцией и без (назовем их также B2 и A). В результате 50 запусков для каждой модели были получены различные локальные максимумы. Три из них для модели B2 приведены в таблице 5, все они лучше по значению правдоподобия, чем существующие модели. Для модели A без миграций получившиеся модели не так сильно отличаются друг от друга, поэтому в таблице приведена только одна лучшая.

Такой же результат был получен и для аллель-частотного спектра, построенного по всем SNV. В таблице 6 приведены параметры лучшей модели для типа A и трех моделей для B2, которые также изображены на рисунке 11.

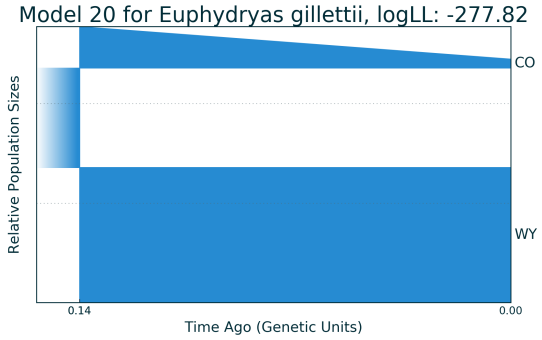
Одним из результатов статьи [9] стал вывод о неприменимости модели с миграциями к истории без миграций. Однако была найдена ошибка при составлении таблицы с параметрами и этот вывод был признан неверным. Более того, оказалось

Таблица 5 – Демографические модели для аллель-частотного спектра из синонимичных SNV.
[1] Внезапное изменение численности
[2] Линейное изменение численности

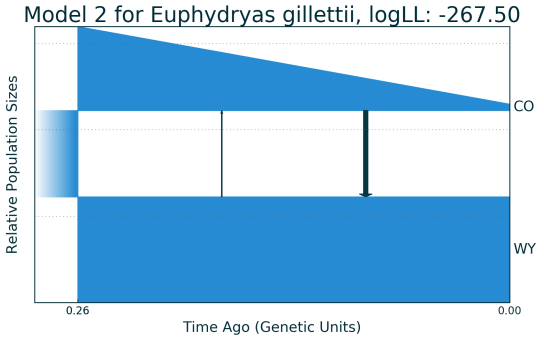
	A [9]	B1 [9]	B2 [9]	A	B2 (1)	B2 (2)	B2 (3)
Log likelihood (∂a∂i)	−211.63	−210.83	−210.80	−211.42	**−205.94**	−210.59	−210.80
Log likelihood (*moments*)	−211.63	−210.83	−210.80	−211.09	**−205.92**	−210.54	−210.76
AIC (∂a∂i)	429.26	429.66	431.60	432.84	**425.88**	435.18	435.60
Значения параметров:							
η_{WY0}	NA	NA	NA	0.79	0.0015	0.999	$(= \eta_{WY})$
η_{WY}	0.922^1	0.884^1	0.893^1	0.960^1	0.856^1	0.742^2	0.873^1
η_{CO}	0.104^1	0.119^1	0.121^1	0.064^2	0.048^2	0.120^1	0.121^1
m_{WY-CO}	NA	0.887	0.906	NA	2.857	0.952	0.923
m_{CO-WY}	NA	NA	0.002	NA	0.176	0.0	0.0
τ_{SPLIT}	0.066	0.080	0.081	0.075	0.241	0.079	0.080

наоборот, что темпы миграций настолько малы, что их можно считать нулевыми, а значит модель с миграциями корректно отражает историю без миграций. Модели выведенные нашим методом также имеют пренебрежительно малые значения миграций и подтверждают этот факт.

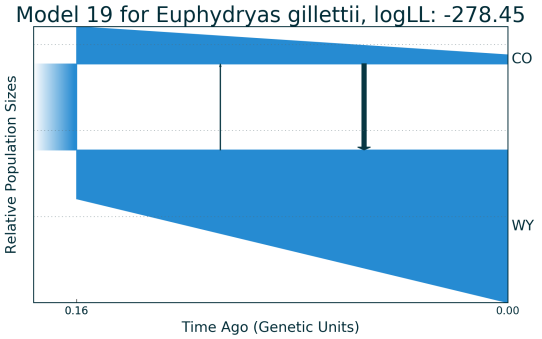
Более того, можно убедится, что полученная лучшая модель согласуется с действительной историей. Средняя численность популяции штата Колорадо (CO), за которой велось наблюдение, была оценена авторами статьи [9] как $N_{CO} = 34$ особи. Если отмасштабировать параметры лучшей модели так, чтобы средняя численность CO была равна этому значению ($x \cdot (\eta_{CO} + (1 - \eta_{WY0}))/2 = N_{CO}$), то получим 33.6 поколений ($t_{SPLIT} = 2 \cdot x \cdot \tau_{SPLIT}$) после разделения популяций, что соответствует реальным 33 поколениям (1977–2010 года).



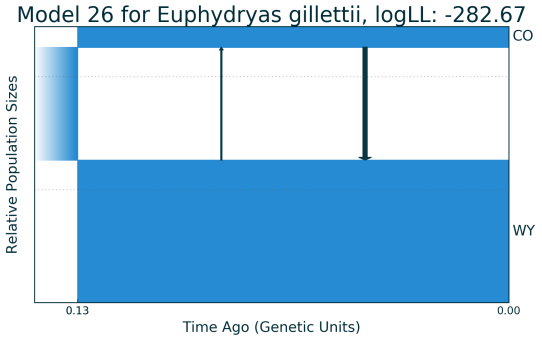
(а) Модель типа A: без миграций.



(б) Модель 1 типа B2: с миграциями.



(в) Модель 2 типа B2: с миграциями.



(г) Модель 3 типа B2: с миграциями.

Рисунок 11 – Альтернативные модели истории развития популяций бабочек *Euphydryas gillettii* из штата Колорадо (CO) и Вайоминг (WY), полученный нашим методом. Модели типа B2 на самом деле имеют близкие к нулю значения миграций.

Таблица 6 – Демографические модели для аллель-частотного спектра из всех SNV.
[1] Внезапное изменение численности
[2] Линейное изменение численности

	A [9]	A	B2 (1)	B2 (2)	B2 (3)
Log likelihood ($\partial a \partial i$)	-283.53	-278.67	$\mathbf{-267.49}$	-278.57	-282.74
Log likelihood (*moments*)	-283.60	-277.82	$\mathbf{-267.50}$	-278.45	-282.67
Значения параметров:					
η_{WY0}	NA	0.584	0.028	0.566	$(=\eta_{WY})$
η_{WY}	1.320^1	1.358^1	1.225^1	1.773^2	1.261^1
η_{CO}	0.173^1	0.089^2	0.074^2	0.105^2	0.183^1
m_{WY-CO}	NA	NA	1.244	0.801	0.316
m_{CO-WY}	NA	NA	0.171	0.056	0.091
τ_{SPLIT}	0.117	0.138	0.259	0.162	0.129
Время разделения в поколениях:					
t_{SPLIT}	45	37	**33**	40	47

3.3. Выводы по разделу «Экспериментальные исследования»

В данном разделе приведены результаты поиска демографических моделей по реальным данным. Для всех исследований рассматривается аллель-частотный спектр, для которого была подобрана с помощью существующих оптимизаций модель. Были рассмотрены три случая: два для популяций современных людей и один для популяции бабочек.

Для популяций людей были подобраны демографические модели с теми же ограничениями на параметры, что и в существующих, которые при этом оказались лучше по значению правдоподобия. Из чего можно сделать следующий вывод: наш метод сравним по эффективности с существующим алгоритмом поиска демографической модели с фиксированным числом параметров, однако при этом он не требует большого числа запусков и начальных значений параметров, которые практически всегда неизвестны.

Также во всех случаях были подобраны демографические модели без ограничений на параметры. Полученные модели имеют не только лучшее правдоподобие, но иногда и лучшее значение AIC, что говорит об отсутствии переобучения в связи с большим числом параметров, значения которых при этом близки к существующим значениям. Все это подтверждает эффективность метода для поиска демографической модели из аллель-частотного спектра.

Для популяции бабочек были получены несколько демографических моделей, все из них имеют значение правдоподобия лучше, чем у модели, построенной локальными оптимизациями. Они показывают альтернативные истории развития популяций, однако выявление их соответствия реальной истории было проведено только на основании времени разделения и требует дальнейших исследований специалистами в соответствующей области.

Заключение

Настоящая работа посвящена выводу демографической истории нескольких популяций из аллель-частотного спектра. Был разработан метод, основанный на генетическом алгоритме, который использует существующие решения — *∂a∂i* и *moments*, для симулирования аллель-частотного спектра из предложенной демографической модели. Метод был реализован в программном обеспечении GADMA, которое лежит в открытом доступе по ссылке `https://github.com/ctlab/GADMA`, поддерживает до трех популяций, а его эффективность была проверена на реальных данных. Были выведены новые, лучшие по значению правдоподобия демографические модели для популяций современных людей и бабочек *E. gillettii*. Более того, полученные модели согласуются с известной историей и другими исследованиями.

По результатам экспериментальных исследований на реальных данных выявлена эффективность метода при поиске большого числа параметров. Метод показывает себя более эффективным, чем существующие оптимизации, которые предлагают *∂a∂i* и *moments*, так как подбирает модели лучшие по правдоподобию. Также была продемонстрирована стабильность поиска, начинающегося с более простой структуры, чем сразу со сложной, что отражает рентабельность использования схемы с ее увеличением. Дополнительно было подтверждено превосходство *moments* над *∂a∂i* по времени работы, что было неоднозначно продемонстрировано на симулированных данных до этого. Таким образом GADMA является первым эффективным программным обеспечением, которое подбирает демографическую модель из аллель-частотного спектра, не требуя от пользователя ничего, кроме единственной информации — структуры, которая определяет то, насколько подробная модель требуется.

В ходе экспериментальных исследований для популяций людей были получены новые несимметричные темпы миграций и показан рост численности евразийской популяции после выхода ее из Африки. Для бабочек *E. gillettii* были выведены несколько альтернативных историй развития популяций, для которых можно проводить дальнейшие исследования для выявления биологически обоснованной.

В качестве дальнейших направлений развития работы можно отметить увеличение числа рассматриваемых популяций (*moments* поддерживает до 5), подбор значений отбора и разработку удобного интерфейса для различных экспертных данных, так как на данный момент можно ограничить только время разделения. Также можно попытаться улучшить предложенный метод, попробовать использовать различные модификации генетического алгоритма, например, подбирающие заведомо различные решения.

Список литературы

[1] Adams A. M., Hudson R. R. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms // Genetics. — 2004. — Vol. 168, no. 3. — P. 1699–1712.

[2] Akaike H. A new look at the statistical model identification // IEEE transactions on automatic control. — 1974. — Vol. 19, no. 6. — P. 716–723.

[3] Consortium 1000 Genomes Project. A global reference for human genetic variation // Nature. — 2015. — Vol. 526, no. 7571. — P. 68.

[4] Demographic history and rare allele sharing among human populations / S. Gravel, B. M. Henn, R. N. Gutenkunst et al. // Proceedings of the National Academy of Sciences. — 2011. — Vol. 108, no. 29. — P. 11983–11988.

[5] Ewens W. J. Mathematical population genetics 1: theoretical introduction. — Springer Science & Business Media, 2012. — Vol. 27.

[6] Excoffier L., Foll M. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios // Bioinformatics. — 2011. — Vol. 27, no. 9. — P. 1332–1334.

[7] Fisher R. A. On the Dominance Ratio // Proceedings of the royal society of Edinburgh. — 1922. — Vol. 42. — P. 321–341.

[8] Gao N., Yang N., Tang J. Ancestral genome inference using a genetic algorithm approach // PloS one. — 2013. — Vol. 8, no. 5. — P. e62156.

[9] Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population / R. C. McCoy, N. R. Garud, J. L. Kelley et al. // Molecular ecology. — 2014. — Vol. 23, no. 1. — P. 136–150.

[10] 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. — P. 1497–1502.

[11] Harris K., Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths // PLoS genetics. — 2013. — Vol. 9, no. 6. — P. e1003521.

[12] Hey J. On the number of New World founders: a population genetic portrait of the peopling of the Americas // PLoS biology. — 2005. — Vol. 3, no. 6. — P. e193.

[13] Holland J. H. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. — 1975.

[14] Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data / D. Garrigan, S. B Kingan, M. M. Pilkington et al. // Genetics. — 2007. — Vol. 177, no. 4. — P. 2195–2207.

[15] Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation / J. Jouganous, W. Long, A. P. Ragsdale, S. Gravel // Genetics. — 2017. — Vol. 206, no. 3. — P. 1549–1567. — http://www.genetics.org/content/206/3/1549.full.pdf.

[16] Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data / R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante // PLoS genetics. — 2009. — Vol. 5, no. 10. — P. e1000695.

[17] Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes / B. F. Voight, A. M. Adams, L. A. Frisse et al. // Proceedings of the National Academy of Sciences of the United States of America. — 2005. — Vol. 102, no. 51. — P. 18508–18513.

[18] Kimura M. Stochastic processes and distribution of gene frequencies under natural selection. // Cold Spring Harbor symposia on quantitative biology. — Vol. 20. — 1955. — P. 33–53.

[19] Kimura M. Diffusion models in population genetics // Journal of Applied Probability. — 1964. — Vol. 1, no. 2. — P. 177–232.

[20] Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations // Genetics. — 1969. — Vol. 61, no. 4. — P. 893–903.

[21] Kolmogoroff A. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung // Mathematische Annalen. — 1931. — Vol. 104, no. 1. — P. 415–458.

[22] Kuhner M. K. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters // Bioinformatics. — 2006. — Vol. 22, no. 6. — P. 768–770.

[23] Lohse K., Harrison R. J., Barton N. H. A general method for calculating likelihoods under the coalescent process // Genetics. — 2011. — Vol. 189, no. 3. — P. 977–987.

[24] Mellars P. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia // Science. — 2006. — Vol. 313, no. 5788. — P. 796–800.

[25] Mulligan C. J., Kitchen A., Miyamoto M. M. Updated three-stage model for the peopling of the Americas // PloS one. — 2008. — Vol. 3, no. 9. — P. e3199.

[26] Myers S., Fefferman C., Patterson N. Can one learn history from the allelic spectrum? // Theoretical population biology. — 2008. — Vol. 73, no. 3. — P. 342–348.

[27] Recent and ongoing selection in the human genome / R. Nielsen, I. Hellmann, M. Hubisz et al. // Nature Reviews Genetics. — 2007. — Vol. 8, no. 11. — P. 857.

[28] Sawyer S. A., Hartl D. L. Population genetics of polymorphism and divergence. // Genetics. — 1992. — Vol. 132, no. 4. — P. 1161–1176.

[29] Schiffels S., Durbin R. Inferring human population size and separation history from multiple genome sequences // Nature genetics. — 2014. — Vol. 46, no. 8. — P. 919.

[30] Schraiber J. G., Akey J. M. Methods and models for unravelling human evolutionary history // Nature Reviews Genetics. — 2015. — Vol. 16, no. 12. — P. 727.

[31] Schumer M., Steiglitz K. Adaptive step size random search // IEEE Transactions on Automatic Control. — 1968. — Vol. 13, no. 3. — P. 270–276.

[32] Schwarz G. Estimating the dimension of a model // The annals of statistics. — 1978. — Vol. 6, no. 2. — P. 461–464.

[33] 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.

[34] Sousa V., Hey J. Understanding the origin of species with genome-scale data: modelling gene flow // Nature Reviews Genetics. — 2013. — Vol. 14, no. 6. — P. 404.

[35] Watterson G. A. On the number of segregating sites in genetical models without recombination // Theoretical population biology. — 1975. — Vol. 7, no. 2. — P. 256–276.

[36] Wright S. Evolution in Mendelian populations // Genetics. — 1931. — Vol. 16, no. 2. — P. 97–159.

[37] Zwickl D. J. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion : Ph. D. thesis / D. J. Zwickl. — 2006.

[38] The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations / G. T. Marth, E. Czabarka, J. Murvai, S. T. Sherry // Genetics. — 2004. — Vol. 166, no. 1. — P. 351–372.

[39] The joint allele-frequency spectrum in closely related species / H. Chen, R. E. Green, S. Pääbo, M. Slatkin // Genetics. — 2007. — Vol. 177, no. 1. — P. 387–398.

[40] A reference genome for common bean and genome-wide analysis of dual domestications / J. Schmutz, P. E. McClean, S. Mamidi et al. // Nature genetics. — 2014. — Vol. 46, no. 7. — P. 707.

A. Пример представления демографической модели в ∂a∂i

Так как *moments* имеет тот же интерфейс, что и ∂a∂i, продемонстрируем только ∂a∂i. Для примера возьмем следующую демографическую модель: была одна общая предковая популяция размера $N_A = N_{ref}$. Затем она разделилась в пропорции s и $(1 - s)$ на две, которые затем экспоненциально менялись до размеров N_1, N_2 и между ними существовала несимметричная миграция размера m_{12} и m_{21}. Код соответсвующей модели будет следующим:

```
def Isolation_and_migration(params, ns, pts):
    s,nu1,nu2,T,m12,m21 = params

    xx = Numerics.default_grid(pts)

    phi = PhiManip.phi_1D(xx)
    phi = PhiManip.phi_1D_to_2D(xx, phi)

    nu1_func = lambda t: s * (nu1/s)**(t/T)
    nu2_func = lambda t: (1−s) * (nu2/(1−s))**(t/T)
    phi = Integration.two_pops(phi, xx, T, nu1_func, nu2_func,
                               m12=m12, m21=m21)

    fs = Spectrum.from_phi(phi, ns, (xx,xx))
    return fs
```

Заметим, что в параметрах функции нет N_A, так как оно является размером референсной популяции N_{ref} и, по умолчанию, равно 1. Аллель-частотный спектр имеет линейную зависимость от N_{ref}, а значит для получения спектра модели с определенным значением N_A можно просто умножить АЧС на соответсвующий коэффициент.

B. Глоссарий

Аллель

вариант гена, локуса хромосомы.

Аллель-частотный спектр N популяций

совместное распределение частот аллелей у популяций.

Аллопатрическое видообразование

способ видообразования, который возникает, когда биологические популяции одного и того же вида становятся изолированными друг от друга.

Гамета

репродуктивные клетки, имеющие гаплоидный (одинарный) набор хромосом и участвующие, в частности, в половом размножении.

Гаплотип

совокупность аллелей на локусах одной хромосомы, обычно наследуемых вместе.

Генетическое расстояние

мера генетического различия между видами, подвидами, или популяциями одного вида. Малое генетическое расстояние означает генетическое сходство.

Гетерозигота

диплоидный организм или клетка, несущий разные аллели в гомологичных хромосомах (Aa).

Гомоплазия

это параллельная или конвергентная эволюция, приводящая к сходству признаков, не опираясь на общее происхождение. Например, крылья насекомых, летучих мышей и птиц, имеют одинаковые функции и подобны по структуре, но имеют независимое эволюционное происхождение.

Дрейф генов

явление ненаправленного изменения частот аллельных вариантов генов в популяции, обусловленное случайными статистическими причинами.

Естественный отбор

эволюционный процесс, в результате действия которого в популяции увеличивается число особей, обладающих максимальной приспособленностью (наиболее благоприятными признаками), в то время, как количество особей с неблагоприятными признаками уменьшается.

Мутация

постоянное изменение нуклеотидной последовательности генома организма.

Мутагенез

процесс возникновения мутаций.

Популяционная генетика

часть генетики и эволюционной биологии, которая занимается генетическими различиями внутри и между популяциями.

Поток генов

перенос аллелей генов из одной популяции в другую.

Рекомбинация

процесс обмена генетическим материалом путем разрыва и соединения разных молекул (ДНК, иногда РНК у вирусов).

Эволюция

естественный процесс развития живой природы, сопровождающийся изменением генетического состава популяций, формированием адаптаций, видообразованием и вымиранием видов, преобразованием экосистем и биосферы в целом.