Математическая биология и биоинформатика. 2014. Т. 9. № 2. С. 438-452.
URL: http://www.matbio.org/2014/Bukin_9_438.pdf

Индивидуально ориентированная модель для имитации популяционно-генетических процессов у видов, населяющих одномерный ареал

Букин Ю.С.[*1,2], Горбылев А.Л.[1]

[1]*Национальный исследовательский Иркутский государственный технический университет, Иркутск, 664074, Россия*
[2]*Федеральное государственное бюджетное учреждение науки Лимнологический институт СО РАН, Иркутск, 664033, Россия*

Аннотация. В работе предлагается имитационная модель, описывающая популяционно-генетические процессы у вида, населяющего одномерный (ленточный) ареал. В процессе имитационного моделирования учитываются: начальное распределение организмов по ареалу; подвижность особей; распределение плотности ресурсов на ареале; вероятность возникновения нуклеотидной замены в митохондриальном и ядерном генетическом маркере; продолжительность имитационного процесса в поколениях. Модель предназначена для «проигрывания» различных сценариев формирования генетического разнообразия. Полученные в ходе моделирования последовательности ДНК можно сравнить с последовательностями ДНК популяционных выборок организмов естественных популяций для проверки гипотез формирования генетического разнообразия. Программно реализованный алгоритм модели и исходный код можно получить по адресу: https://yadi.sk/d/80YAOsvYckiTr.

Ключевые слова: *имитационное моделирование, индивидуально ориентированная модель, генетика популяций, популяционные процессы, дрейф генов, потоки генов, мутационный процесс, диффузионные процессы.*

1. ВВЕДЕНИЕ

Выборки расшифрованных последовательностей ДНК генетических маркеров широко применяются для анализа популяционной структуры видов и оценки различных популяционных параметров. Подобные исследования дают обширную информацию об особенностях экологических и микроэволюционных процессов, происходящих в разных популяциях [1].

В настоящее время разработан достаточно широкий спектр теоретических моделей, описывающих механизмы формирования генетического разнообразия в пространственно неоднородных системах взаимодействующих популяций [2–4]. Рассмотрим некоторые виды популяционных моделей. Островная модель [1, 2] предполагает формирование генетического разнообразия в локальных популяциях за счет дрейфа генов в каждой популяции и потоков генов между популяциями, заданными матрицей скоростей миграции. Лестничная модель формирования генетического разнообразия предполагает наличие дрейфа генов в каждой популяции и последовательный обмен генами (потоки генов) между изолированными популяциями за счет промежуточных звеньев[5, 6]. Модель изоляции расстоянием описывает процесс формирования генетического разнообразия в пространственно однородной

[*]bukinyura@mail.ru

популяции, занимающей достаточно обширный ареал, такой, что перенос генетической информации из одной части ареала в другую занимает несколько поколений [1, 2, 7]. Модель изоляции расстоянием может быть приближена к островной модели с большим количеством локальных субпопуляций и соответствующей матрицей миграции, или к лестничной модели с большим количеством ступеней (субпопуляций).

Оценка генетического разнообразия, пространственной структуры и подразделенности популяций на основе расшифрованных последовательностей ДНК проводится с помощью разнообразных статистических методов [8–12]. Одним из наиболее используемых критериев подразделенности популяций и интенсивности потоков генов является критерий F_{st}, который можно вычислить на основе расшифрованных последовательностях ДНК [13]. С помощью F_{st} критерия можно определить генетическую разобщенность между группами и популяциями организмов, изолированными как географическими барьерами, так и расстоянием.

Для определения генетической подразделенности в группах организмов, изолированных расстоянием, предложен ряд теоретических программных методов [14-16]. Основная задача, которая ставится при проведении подобных исследований, – это оценка подвижности или эффективного расстояния миграции за одно поколение в вязких популяциях организмов. Под вязкими популяциями подразумеваются группы организмов либо с малой подвижностью, либо с достаточно обширным ареалом, способствующим тому, что перенос генетической информации из одной части ареала в другую требует нескольких поколений.

Обычно исследователь при проведении генетического анализа и установлении подразделенности исследуемого вида на локальные популяции и группы выдвигает какую-либо гипотезу, объясняющую механизм формирования наблюдаемого внутривидового генетического полиморфизма. Подобная гипотеза состоит из ряда предположений, касающихся 1) наличия географических барьеров в пределах ареала, замедляющих потоки генов; 2) определенной вязкости популяции или изоляции расстоянием; 3) предположения относительно эффективного размера локальных групп особей; 4) информацию о вероятности мутации в исследуемом генетическом маркере. Проверить предложенную гипотезу можно следующим способом: необходимо задать все известные параметры, начальные и граничные условия в теоретической модели, которая будет «проигрывать» сценарий формирования генетического разнообразия. На выходе подобная модель должна выдавать выборки последовательностей ДНК. Сравнивая значения популяционных параметров и критериев, рассчитанных на основе выборок расшифрованных последовательностей ДНК природных образцов и модельных данных, можно сделать выводы относительно вероятности принятия выдвинутой гипотезы.

Существует ряд компьютерных программ, позволяющих получать генетические данные, в том числе последовательности ДНК, соответствующие тому или иному сценарию (гипотезе) формирования внутривидового генетического разнообразия. Все эти программы можно разделить на две группы. Первая группа включает в себя программы, имитирующие популяционные процессы в панмиксных популяциях, обменивающихся генетическим материалом в течение ряда поколений. Вторая группа программ включает модели популяций, распределенных в пространстве с географическими барьерами и заданной подвижностью организмов.

Программы, симулирующие дрейф генов и потоки генов в панмиксных популяциях, основаны на принципах индивидуально ориентированного моделирования [17] и на коалесцентных методах [18, 19]. С помощью программ, описанных в работах [17–19], можно достаточно просто задать всевозможные сценарии формирования генетического разнообразия согласно островной и лестничной модели, а также комбинаций данных моделей. Преимуществом программ данной серии является достаточно простой интерфейс задания начальных параметров и популяционного сценария.

Компьютерные программы, имитирующие процессы в пространственно-распределенных популяциях, чаще всего основаны на принципах индивидуально-ориентированного моделирования [20–23]. Двумерное пространство в таких моделях может быть представлено сеткой [20–23], где в каждой ячейке заданы условия предельного количества и подвижности организмов, либо прямоугольником заданной ширины и высоты и непрерывными координатами организмов [21]. Данный класс программного обеспечения позволяет проигрывать множество сценариев популяционной дифференциации, в том числе с учетом изоляции расстоянием. Недостатком программ данного типа является громоздкость описания качеств двумерного пространства и достаточно большой набор индивидуальных характеристик организмов.

В природе существует большое количество видов с ленточным ареалом. Чаще всего это организмы, населяющие литоральные зоны морей и озер. Кроме этого, ленточные ареалы имеют организмы, обитающие в зонах высотной поясности или на границах стыка климатических зон и т. п. Обычно в подобных условиях поперечное пересечение ареала требует гораздо меньшего времени (меньше продолжительности жизни одного поколения), чем продольное пересечение ареала. С точки зрения накопления мутаций в ДНК поперечной составляющей пространства на таком ареале можно пренебречь. Генетическая дифференциация, связанная с пространственной изоляцией, будет приходиться только на продольную координату. Таким образом, достаточно узкий ленточный ареал распространения вида можно считать одномерным.

Имеется ряд работ, посвященных изучению генетической дифференциации видов с условно одномерными ареалами обитания. В частности, изучались механизмы формирования популяционно-генетического разнообразия у видов, населяющих береговую зону древних озер [24–29] и морей [30]. Многие виды, описанные в этих работах, имеют достаточно высокую степень генетической подразделенности, связанную с изоляцией расстоянием и географическими барьерами.

Вопросам формирования внутривидового и межвидового генетического разнообразия у популяций организмов, населяющих одномерный ареал, посвящен ряд теоретических работ [31–33]. Все вышеперечисленные исследования используют индивидуально-ориентированные модели, с помощью которых имитируется тот или иной сценарий формировании генетического разнообразия. В работе [33] делается попытка сравнения популяционных показателей, рассчитанных на основе выборок расшифрованных последовательностей ДНК природных организмов и компьютерных данных.

В свете вышесказанного можно сделать вывод о том, что компьютерная модель, имитирующая популяционные процессы у организмов, населяющих одномерный ареал с заданной структурой, и дающая на выходе последовательности ДНК (которые можно сравнить с природными данными) была бы интересна для многих исследователей. Одним из условий широкого использования подобной компьютерной программы должна быть простота интерфейса, задающего структуру ареала, начальное распределение организмов и их характеристики. Для этого необходимо четко определиться с алгоритмом модели, выделить основные факторы, участвующие в формировании генетического разнообразия, для минимизации количества параметров. Созданию подобной модели, реализованной в виде компьютерной программы, посвящена данная работа.

2. РАЗРАБОТКА И ОПИСАНИЕ МОДЕЛИ

2.1. Основное уравнение динамики популяции

В основу модели были заложены принципы описания конкурентного взаимодействия организмов, предложенные в работах [34, 35], и принципы построения

динамической модели с конкуренцией, предложенной в работах [36–40]. Для описания подвижности организмов был использован диффузионный принцип, предложенный в работе [33]. В результате обобщения всех вышеперечисленных идей было сформулировано следующее уравнение динамики популяции на одномерном ареале:

$$\frac{\partial N(x,T)}{\partial T} = rN(x,T)\left(1-\frac{\int_{x_{\min}}^{x_{\max}} C(x-y)N(y,T)dy}{K(x)}\right) + D\frac{\partial^2 N(x,T)}{\partial x^2}. \qquad (1)$$

В уравнении (1) функция $N(x,T)$ определяет число особей в точке пространства x в момент времени T (T – промежуток времени, в течение которого рассматривается существование популяции), r – скорость размножения организмов, D – коэффициент диффузии, функция $K(x)$ определяет максимальную плотность особей в точке пространства x, или другими словами, максимальное количество пищевых ресурсов доступных для организмов в точке x. Функция $C(x-y)$ определяет уровень конкуренции между удаленными в пространстве особями, эта функция должна обладать следующими свойствами: $C(x-y)=1$ при $|x-y|=0$, $C(x-y)$ убывает при увеличении модуля разности $|x-y|$.

На основе уравнения (1) была разработана индивидуально-ориентированная, пошаговая компьютерная модель с дискретным количеством организмов. Каждый организм в модели обладает набором собственных характеристик. Общая схема алгоритма определяется последовательностью идущих друг за другом жизненных циклов. Рассмотрим по порядку приближения модели, касающиеся уравнения (1).

2.2. Диффузионное приближение

Если вычленить диффузионную составляющую из уравнения (1), то полученная компонента приобретет следующий вид:

$$\frac{\partial N(x,t)}{\partial t} - D\frac{\partial^2 N(x,t)}{\partial x^2} = 0. \qquad (2)$$

В уравнении (2) $N(x,t)$ – количество особей в момент времени t в точке одномерного пространства с координатой x, D – коэффициент диффузии, t – промежуток времени, в течение которого рассматривается движение особей в популяции.

Следует заметить, что в глобальном смысле время T уравнения (1) и время t уравнения (2) при мгновенном рассмотрении системы является одним и тем же временем, текущим с одной и той же скоростью. Отличие заключается в том, что промежуток времени, в течение которого рассматриваться история популяции T намного превышает промежутку времени t, в течение которого рассматривается движение особей. Промежуток времени t ограничен временем жизни организмов.

Уравнение (2) при начальном условии $N(x,0)=\delta(x)$ – точечный источник, выраженный дельта-функцией Дирака, и граничном условии $N(\infty,t)=0$ имеет решение [41]:

$$N(x,t) = \sqrt{\frac{1}{4\pi Dt}}\exp\left(-\frac{x^2}{4Dt}\right). \qquad (3)$$

Соотношение (3) представляет собой функцию нормального распределения. Согласно соотношению (3), средний квадрат удаления диффундирующей частицы $\bar{x}^2$ (в нашем случае организма) от первоначального положения за время t определяется следующим образом:

$$\bar{x}^2 = \int_{-\infty}^{\infty} x^2 N(x,t)\, dx = 2Dt\,. \qquad (4)$$

Таким образом, расстояние удаления организма от первоначальной точки определяется случайным образом по закону нормального распределения с параметром среднеквадратичного отклонения $\sigma = \sqrt{2Dt}$. Если за единичный временной промежуток в модели принять продолжительность жизни одного поколения, то среднее расстояние $\bar{x}(1)$, на которое организм переместится от начальной точки, определяется соотношением $\bar{x}(1) = \sqrt{2D}$. За любой другой промежуток времени *t*, в течение которого будет рассматриваться процесс движения организма, среднее расстояние удаления от первоначальной точки запишется как $\bar{x}(t) = \bar{x}(1)\sqrt{t}$. Следовательно, наиболее реальным биологическим параметром, определяющим подвижность организмов в индивидуально ориентированной модели, будет расстояние $\bar{x}(1)$, на которое организмы в среднем удалятся от своего первоначального положения за одно поколение. Расстояние удаления от первоначальной точки за время *t* определяется как случайное число, распределенное по нормальному закону со среднеквадратичным отклонением $\sigma = \bar{x}(1)\sqrt{t}$.

Другим следствием из диффузионного приближения при индивидуально ориентированном моделировании будет определение расстояния, на котором два репродуктивных партнера (самка и самец) могут встретиться для воспроизводства потомства. Очевидно, что в течение момента времени *t* репродуктивные партнеры могут встретиться на расстоянии, определяемом подвижностью организмов $\bar{x}(1)$. Расстояние выбора партнера будет определяться числом, распределенным по нормальному закону со среднеквадратичным отклонением $\sigma = \bar{x}(1)\sqrt{t}$.

Функция конкуренции $C(x-y)$ также определяется подвижностью организмов. Чем больше подвижность особей $\bar{x}(1)$, тем интенсивней будет конкуренция между организмами, находящимися в удаленных точках. Для индивидуально ориентированной модели был выбран следующий явный вид функции $C(x-y)$:

$$C(x-y) = \exp\left(-\frac{(x-y)^2}{a(\bar{x}(1)\sqrt{t})^2}\right). \qquad (5)$$

В формуле (5) параметр *a* можно подобрать таким образом, чтобы значение функции $C(x-y) = 0.01$ при $|x-y| = 3\bar{x}(1)\sqrt{t}$. В этом случае конкуренция между особями, удаленными друг от друга на расстояние большее, чем $3\bar{x}(1)\sqrt{t}$ будет ничтожно мала. Значение параметра *a* при заданных условиях будет приблизительно равно 1.954.

Таким образом, осуществляя предложенное в этом разделе диффузионное приближение для индивидуально-ориентированной модели, можно избавиться от коэффициента диффузии уравнения (1) и заменить его параметром $\bar{x}(1)$, определяющим подвижность особей.

2.3. Структура ареала и географические барьеры

В модели в качестве ареала задается единичный отрезок с координатами на оси 0 и 1. Границы отрезка являются отражающими: если какой-либо организм в процессе движения потенциально может пересечь границу ареала на известное расстояние, то он отражается обратно на это известное расстояние.

Диффузионное приближение в модели при единичной длине ареала можно использовать в том случае, если значение произведения $\bar{x}(1)\sqrt{t}$ соотношения (5) намного меньше единицы ($\bar{x}(1)\sqrt{t} \ll 1$). При этом соотношении параметра $\bar{x}(1)$ и промежутка времени движения организма t для полного переноса генетической информации из одного участка единичного ареала в другой будет требоваться смена нескольких поколений организмов. В этой ситуации мы действительно будем иметь дело с группами организмов вида, изолированных расстоянием. Изоляция расстоянием при этом будет играть существенную роль в формирование генетического разнообразия. При больших значениях $\bar{x}(1)\sqrt{t}$ изоляция расстоянием не будет играть существенной роли, и популяционная модель трансформируется в модель свободно скрашивающейся панмиксной популяции.

По биологическому смыслу рассматривать время движение особей t можно только в течение промежутка, меньшему или равному средней продолжительности жизни организмов. Если условная продолжительность поколения равна единице, то промежуток времени t в модели должен быть меньшим или равным 1 ($t \le 1$).

Структура ареала и географические барьеры в его пределах определяются видом функции $K(x)$. Для функции $K(x)$ необходимо определить максимально возможную плотность особей на единицу длины ареала Km. Затем задается кусочно-постоянная функция $K'(x)$, которая может принимать значения от 0 до 1 на единичном отрезке. Сама функция плотности особей определяется следующим соотношением: $K(x) = Km\ K'(x)$. Таким образом, одномерный ареал будет состоять из участков с благоприятными условиями для жизни особей, где значение $K(x)$ близко к Km и участков, где условия обитания неблагоприятны, где значение $K(x)$ мало или близко к нулю. Протяженные участки на единичном отрезке, длина которых больше чем $\bar{x}(1)$ и значение $K(x)$ мало, можно считать географическими барьерами, препятствующими свободному перемещению особей.

Для преобразования ареала природных популяций в единичный необходимо его длину принять за единицу. Отталкиваясь от этого, можно определить Km и задать функцию $K'(x)$, затем определить параметр $\bar{x}(1)$ как долю от общей длины ареала, на которую организмы в среднем перемещаются за одно поколение от начального положения.

2.4. Переход к дискретному количеству организмов и алгоритм модели

В модели предполагается существование дискретного количества организмов N, каждая особь обладает двумя характеристиками: положение в пространстве x_i и пол (самка или самец). Первоначальное количество организмов N, их координаты x_i и половая принадлежность задается при инициализации начальных условий. Для этого задается кусочно-постоянная функция $N(x, T = 0)$, имеющая целые положительные значения на отрезках $(0 < x_1 < x_2 < \ldots < x_n < \ldots < 1)$. На каждом отрезке случайным образом по равномерному закону распределяются координаты x_i, заданные для интервала количества особей. Пол особей определяется случайным образом с вероятностью 0.5 – самка либо самец (см. рис. 1).

После инициализации начальных условий один за другим следует определенное количество репродуктивных циклов. В каждом репродуктивном цикле рождается N потомков (размер популяции удваивается). Если предположить, что плотность особей на каком-либо участке ареала ограничена (см. ур. (1)), то в следующее поколение из удвоенного количества особей в среднем перейдет только половина. Для популяции, динамика которой описывается логистическим уравнением со случайной смертностью

особей, средняя продолжительность жизни зависит от скорости размножения. При скорости размножения $r = 1$ средняя продолжительность жизни составит две временные единицы. В популяциях с перекрывающимися поколениями средняя продолжительность поколения равна средней репродуктивной продолжительности жизни особей. Это означает, что в среднем за две временных единицы в модели происходит полная смена поколений. Если в качестве меры времени выбрать количество поколений, прошедших с момента старта моделирования, то один репродуктивный цикл по времени будет продолжаться 0.5 поколения. Соответственно, за два репродуктивных цикла сменится одно поколение, если взять другую скорость размножения r, отличную от единицы, то это приведет только к изменению средней продолжительности жизни особей и изменению количества репродуктивных циклов, затрачиваемых на смену одного поколения. Таким образом, в уравнении (1) можно избавиться от параметра «скорость размножения» r, заменив его числом поколений, в течение которых происходит имитационное моделирование. Средняя продолжительность репродуктивной жизни, оцененная по возрастным характеристикам организмов, является более доступным параметром, чем скорость размножения.

В диффузионном приближении движение организмов будет рассматриваться течении промежутка времени t от 0 до 0.5 соответственно. Следовательно, в формуле (5), среднем расстоянии, пройденном организмом за время $t = 0.5$ и среднем расстоянии для выбора партнера для размножения вместо $\sqrt{t}$ будет фигурировать $\sqrt{0.5}$.

На каждом последующем репродуктивном цикле в момент времени $T + 0.5$, $N(T + 0.5) = N(T) + N(T)$, где первое слагаемое – это особи, пришедшие из предыдущего цикла, а второе слагаемое – это вновь рожденные особи. Перед воспроизводством потомства в количестве $N(T)$ штук самки и самцы образуют репродуктивные пары. Выбор партнера для размножения происходит согласно следствию из диффузионного приближения. Репродуктивные пары в модели образуются с помощью следующего алгоритма: из всего количества организмов выбираются случайные самка и самец, находится расстояние $x_r = |x_i - x_j|$ между ними, генерируется случайное число по закону нормального распределения с параметром среднего значения x_r и среднеквадратичным отклонением $\bar{x}(1)\sqrt{0.5}$, если сгенерированное число по модулю меньше чем x_r, то выбранные организмы образуют репродуктивную пару. Выбор партнеров продолжается до тех пор, пока пару себе не найдут все самки, удаленные от других самцов на расстояние меньшее, чем $2\bar{x}(1)\sqrt{0.5}$ (расстояние 2σ для нормального распределения). Каждая выбранная самка со своим репродуктивным партнером запоминается как репродуктивная пара (в модели формируется массив репродуктивных пар). Из массива репродуктивных пар выбирается случайная пара с возвратом, которая дает одного потомка. Процедура выполняется до тех пор, пока общее количество потомков не достигнет $N(T)$. Таким образом, каждая самка, для которой был найден репродуктивный партнер, дает случайное количество потомков так, чтобы суммарное количество рожденных организмов составило $N(T)$.

Самки, для которых репродуктивный партнер не был найден, потомства не дают. Все вновь рожденные особи получают координату матери.

После воспроизведения потомства все $N(T + 0.5)$ особи в модели меняют свои координаты (передвигаются). Каждая особь получает новую координату x_i', которая представляет собой случайное число, сгенерированное по нормальному закону со средним значением x_i – предыдущая координата и среднеквадратичным отклонением $\bar{x}(1)\sqrt{0.5}$.

Следующим событием, происходящим на репродуктивном шаге, является случайная элиминация особей в результате конкуренции. Для каждой особи на основании уравнения (1) и формулы (5) рассчитывается вероятность смерти P_i на данном репродуктивном шаге по формуле:

$$P_i = \frac{\sum_{j=1}^{N(T+0.5)} C(x_i - x_j)}{K(x_i)}. \quad (6)$$

Данная формула получается путем перехода от интегрирования по пространству в правой части уравнения (1) к суммированию по особям.

После расчета P_i для каждого организма генерируется случайное число h, распределенное по равномерному закону в диапазоне от 0 до 1, если $P_i \geq h$, то вероятность смерти для организма реализовалась, и он элиминируется из популяции. При реализации элиминации организмов подсчитывается общее количество умерших особей $N'(T+0.5)$, и в следующий репродуктивный шаг переходит $N(T+0.5) - N'(T+0.5)$ особей. Общая блок-схема реализации алгоритма индивидуально-ориентированной имитационной модели представлена на рисунке 1.

2.5. Реализация и наследование генетической информации

Каждый организм в модели «обладает» тремя последовательностями ДНК протяженностью в 700 букв: A, T, G или C. Первая последовательность ДНК наследуется по материнской линии (модель передачи митохондриальной ДНК), то есть передается от матери к потомству. Вероятность мутации для этой последовательности задается как *Pm* на поколение. Вторые две последовательности представляют собой гомологичные локусы диплоидного организма. При наследовании ядерной ДНК потомок в качестве одного своего локуса получает один из двух случных локусов отца и в качестве другого своего локуса один из двух случайных локусов матери. В обоих локусах с вероятностью *Pn* может произойти мутация. Как в митохондриальном, так и в ядерном локусах мутация представляет собой процесс замены случайно выбранной позиции на одну из четырех случайно выбранных букв. При инициализации начальных условий все организмы получают в качестве начальной одинаковую случайно сгенерированную последовательность ДНК.

2.6. Сохранение и визуализация результата

В конце работы имитационного алгоритма координаты организмов и их последовательности ДНК сохраняются на диск в виде текстового файла. В другом текстовом файле, имеющим структуру csv формата, сохраняется информация об изменении координат организмов в процессе моделирования. Эту информацию можно использовать для восстановления картины миграции и изменения плотности организмов в пределах ареала с момента старта и до окончания имитационного процесса. Для визуализации данных о расселении и изменении плотности организмов внутри ареала был разработан скрипт на языке программирования R. Описание программы, скрипт для визуализации результата, исполняемый файл программы, и исходный код на языке программирования C++ можно скачать по ссылке (https://yadi.sk/d/80YAOsvYckiTr).

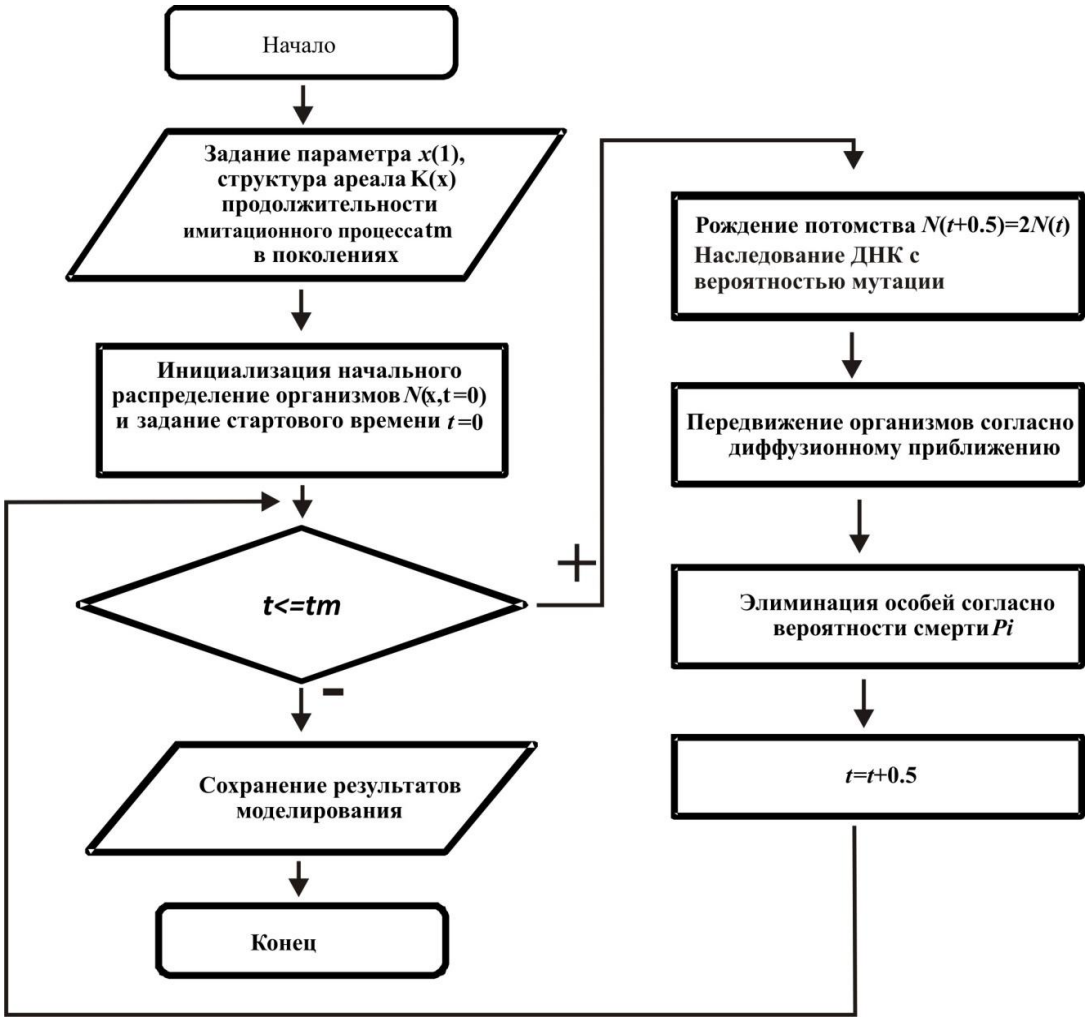

Рис. 1. Общая блок-схема алгоритма имитационного моделирования.

3. ДЕМОНСТРАЦИЯ ПРИМЕРА МОДЕЛИРОВАНИЯ

В качестве примера моделирования популяционного сценария была выбрана структура ареала, предполагающая лестничную модель миграции с включением некоторой степени изоляции расстоянием. Ареал состоял их четырех протяженных отрезков от ноля до 0.23, от 0.27 до 0.48, от 0.52 до 0,73 и от 0.77 до 1, в пределах которых значение плотности особей $K(x)$ равнялось 600. Участки ареала, в пределах от 0.23 до 0.27, от 0.48 до 0.52 и от 0.73 до 0.77, представляли собой неабсолютные географические барьеры с плотностью особей $K(x)$ равной 120. Структура ареала приведена на рисунке 2. Подвижность организмов, определяемая параметром $\bar{x}(1)$, равнялась 0.02. При данном значении $\bar{x}(1)$ на полный перенос генетического материала от одной границы ареала до другой, при отсутствии географических барьеров, в среднем должно затрачиваться 50 поколений. Общее время имитации популяционного процесса составило T = 1000 поколений. Вероятность мутации для митохондриального маркера была Pm = 0.01 и для ядерного маркера Pn = 0.005. В начальный момент времени было задано 40 организмов, равномерно распределенных на отрезке от 0.9 до 1.

В результате имитационного моделирования, проведенного при вышеуказанных значениях параметров, была получена картина расселения организмов, изображенная на рисунке 2. Из рисунка видно, что после старта моделирования организмы начали постепенно расселяться по ареалу. К моменту времени, равному приблизительно 100 поколений с момента старта, весь ареал был заселен. При столкновении организмов с

географическими барьерами процесс заселения ареала замедлялся. Особенно это заметно при заселении предпоследнего отрезка с координатами от 0.27 до 0.48. Наличие географических барьеров привело к тому, что заселение ареала произошло не за расчетные 50 поколений, как следовало ожидать, исходя из значения параметра подвижности, а за больший промежуток времени (100 поколений).

В конце имитационного моделирования были сохранены последовательности ДНК ядерного и митохондриального маркеров организмов. Для сохранения данных было выделено 4 группы организмов из участков ареала со значениями $K(x)$ равным 600. Из каждой группы для анализа были взяты последовательности 30 организмов. Выделенные группы подпадают под понятия популяции вида. Таким образом, мы имеем дело с 4 популяциями организмов, разделенных географическими барьерами по лестничной структуре и изоляцией расстоянием.

Таблица 1. Значения характеристик внутригруппового полиморфизма

Название группы	Средняя внутригрупповая доля замен, оцененная по митохондриальному маркеру D_{vm}	Средняя внутригрупповая доля замен, оцененная по ядерному маркеру D_{vn}
g1	0.00124	0.02847
g2	0.00440	0.01975
g3	0.00182	0.04583
g4	0.02047	0.03464

Для каждой группы рассчитывался коэффициентов внутригруппового генетического разнообразия, исходя из митохондриальных *Dvm* и ядерных *Dvn* генетических маркеров. В качестве меры внутригруппового разнообразия использовалась средняя доля замен при попарном сравнении последовательностей ДНК (табл. 1). Расчеты производились с помощью программы MEGA 5.1 [42, 43].

Коэффициент генетического разнообразия, рассчитанный для ядерных маркеров D_{vn} в среднем оказался больше, чем коэффициент генетического разнообразия, рассчитанный для митохондриальных маркеров D_{vm} (см. табл. 1), несмотря на то, что заданная вероятность мутации в митохондриальном маркере была больше, чем в ядерном ($Pm = 0.01$, $Pn = 0.005$). Объяснить данный факт можно тем, что коэффициенты генетического разнообразия прямо пропорциональны произведению вероятности мутации и эффективному размеру популяции ($D_{vm} \approx NePm$, $D_{vn} \approx NePn$) [1, 2]. Эффективный размер популяции для митохондриальных маркеров в четыре раза меньше, чем для ядерных маркеров, так как митохондриальные маркеры находятся в организме только в одной копии и передаются только по материнской линии. Поэтому, при вдвое большей вероятности мутации, генетическое разнообразие в митохондриальных маркерах будет меньше, чем разнообразие в ядерных маркерах в заданном модельном сценарии.

На следующем этапе для каждой пары групп был подсчитан коэффициент межгруппового генетического разнообразия (см. табл. 2). Полученные расчеты показывают, что средняя доля замен между последовательностями разных групп как по ядерному, так и по митохондриальному маркеру в среднем в 2.4 раза превышает расчетные коэффициенты внутригруппового разнообразия. Это свидетельствует о том, что географические барьеры в пределах ареала действительно оказываются изолирующим фактором между выделенными группами.

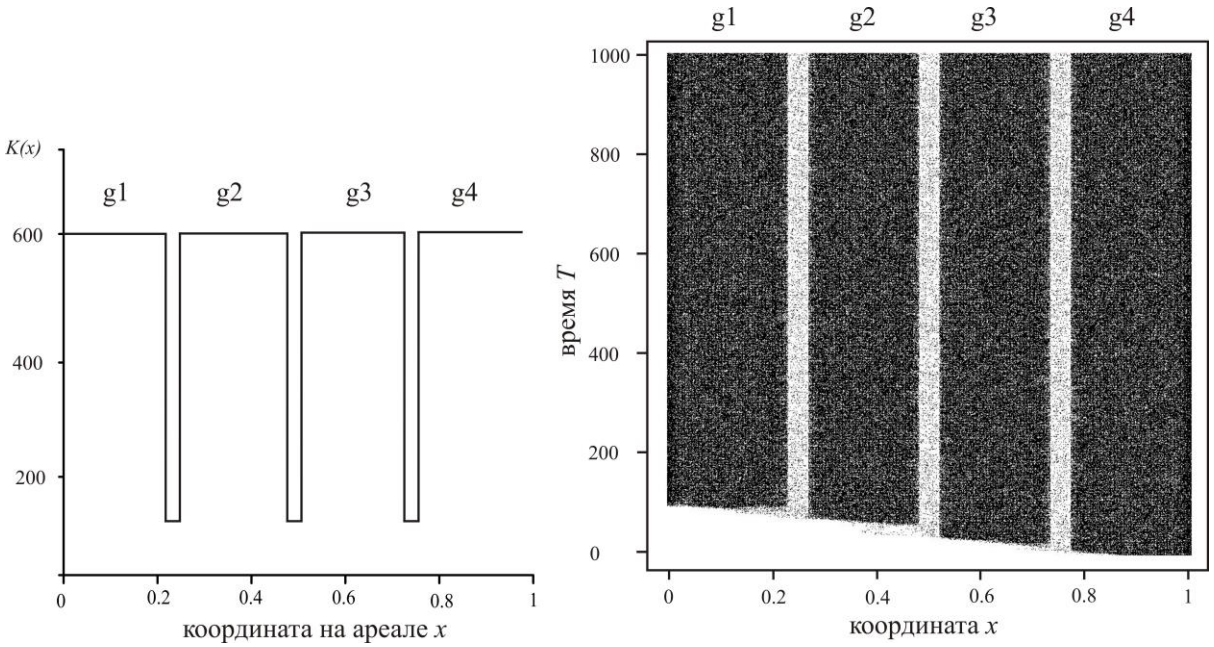

Рис. 2. Структура ареала, заданная функцией $K(x)$ и картина расселения организмов по ареалу в процессе моделирования, g1, g2, g3 и g4 – выделенные для дальнейшего анализа группы организмов.

По соотношению, предложенному в работе [13], рассчитаны межгрупповые показатели *Fst* критерия. Результаты приведены в таблице 2. Значение *Fst* показывает степень изоляции между популяциями: чем ближе значение *Fst* к единице, тем меньше поток генов между популяциями. Данные таблицы 2 говорят о том, что значения *Fst* критерия рассчитанного, на основе митохондриальных маркеров больше, чем *Fst*, рассчитанный на основе ядерных маркеров. Это объясняется тем, что особи, мигрируя из одной группы в другую, переносят с собой две последовательности ядерной ДНК и только самки переносят одну последовательность митохондриальной ДНК. Следовательно, при одном и том же количестве переходящих особей расчетное значение *Fst* на основе митохондриальных маркеров покажет большую степень изоляции.

Таблица 2. Значения характеристик межгруппового полиморфизма

Название группы №1	Название группы №1	Средняя межгрупповая доля замен по митохондриальному маркеру *Dbm*	Средняя межгрупповая доля замен по ядерному маркеру *Dbm*	Значение *Fst* критерия, рассчитанного на основе митохондриального маркера *Fstm*	Значение *Fst* критерия, рассчитанного на основе ядерного маркера *Fstn*
g1	g2	0.0219	0.0701	0.871	0.656
g1	g3	0.0224	0.0834	0.931	0.554
g1	g4	0.0282	0.0816	0.967	0.613
g2	g3	0.0127	0.0762	0.755	0.569
g2	g4	0.0192	0.0779	0.869	0.651
g3	g4	0.0096	0.0561	0.873	0.283

Все полученные при моделировании результаты в принципе согласуются с общими теоретическими и экспериментальными представлениями популяционной генетики.

4. РЕКОМЕНДАЦИИ ПО ВЫБОРУ ЗНАЧЕНИЯ ПАРАМЕТРОВ ДЛЯ МОДЕЛИРОВАНИЯ

Основной проблемой, возникающей при имитационном индивидуально-ориентированном моделировании, является большая продолжительность расчетов по времени. Применение программного распараллеливания самого алгоритма не приведет к существенному ускорению расчетов, потому что потребуется передача большого количества информации между потоками. В результате этого рост скорости вычисления будет существенно меньшим, чем рост числа используемых процессоров в параллельной системе. Поэтому расчеты по имитационным моделям с такими параметрами, как число особей и число поколений, соответствующими естественным популяциям, становятся практически невозможными. Модель при таких значениях параметров потребует описания миллионов организмов, существующих сотни тысяч и миллионы поколений. Однако, выход из этой ситуации есть. Уже говорилось о том, что показатель внутрипопуляционного генетического полиморфизма Dv вычисляется как произведение эффективного размера популяции Ne и вероятность мутации в исследуемом маркере за поколение P [1,2], то есть $Dv \approx NeP$. Для представленной модели Ne равен количеству особей и прямо пропорционален максимальной плотности организмов в пределах ареала Km. Для того чтобы получить популяцию с тем же показателем генетического разнообразия, что и природная популяция, можно увеличить вероятность мутации P (для модели увеличить Pm и Pn) и, соответственно, уменьшить Ne (для модели уменьшить Km). Варьируя значения параметров Km, Pm и Pn, можно получить приемлемые для расчетов численности организмов.

Вопрос о количестве поколений, в течение которых рассматривается «модельная» популяция, связанный с продолжительностью имитационного моделирования, разрешается с помощью следующих предположений. Если для природной популяции или исследуемых видов известно время происхождения, которое может быть рассчитано методами молекулярной филогении, то можно определить долю замен, накопленную маркером ДНК с момента образования исследуемой группы организмов до наших дней. Такую же долю замен должен накопить маркер в имитационной модели от своего первоначального состояния. Если мы знаем вероятность мутаций Pm и Pn, скорректированную для задания Km (см. предыдущий абзац), то по расчетной доле замен для естественной популяции можно рассчитать количество поколений, в течение которых должны существовать «модельные» популяции по формулам:

$$tm = \frac{700sm}{Pn} \quad \text{или} \quad tm = \frac{700sp}{Pm} \ . \tag{7}$$

В этих формулах tm – это продолжительность имитационного моделирования в поколениях, 700 – длина нуклеотидной последовательности в модели, sm – доля замен, накопленная митохондриальным маркером природной популяции с момента расхождения популяций, sp – доля замен, накопленная ядерным маркером с момента расхождения популяций, Pm и Pn – вероятности мутации в модели на последовательность за поколение.

После определения максимальной плотности заселения организмами ареала Km, вероятностей мутации Pm и Pn и времени моделирования tm, можно определить подвижность организмов. Если известно время существования T исследуемого вида или групп популяций исследуемого вида в годах, можно оценить в годах продолжительность жизни модельного поколения Tm как $Tm = T / tm$. После этого можно рассчитать количество поколений n естественной популяции организмов, укладывающихся в промежуток времени Tm по формуле $n = Tm / tp$, где tp – время в годах средней продолжительности жизни поколения у исследуемого вида организмов. Если известна величина xp – среднее расстояние, на которое перемещается организм от

первоначальной точки за одно поколение в природной популяции, то для имитационной модели $\bar{x}(1) = xp\sqrt{n}$. Число *n*, полученное на основании расчетов, может быть дробным.

Работа выполнена при финансовой поддержке программы подготовки магистров по специальности «Биоинформатика» НИ ИрГТУ и темы бюджетного финансирования VI.61.1.3.

СПИСОК ЛИТЕРАТУРЫ

1. Алтухов Ю.П. *Генетические процессы в популяциях*. М.: ИКЦ Академкнига, 2003. 432 С.
2. Etheridge A. *Some Mathematical Models from Population Genetics*. Springer, 2011. 119 p.
3. Wright S. Evolution in Mendelian populations. *Genetics*. 1931. V. 16. P. 97–159.
4. Wright S. The genetical structure of population. *Ann. Eugenics*. 1951. V. 15. P. 323–354.
5. Kimura M. “Stepping stone” model of population. *Ann. Rep. Nat. Ist. Genet. Mishima*. 1953. V. 3. P. 63–65.
6. Kimura M., Weiss G.H. The stepping stone model of population structure and decrease of genetic correlation with distance. *Genetics.* 1964. V. 49. P. 561–567.
7. Wright S. Isolation by distance. *Genetics*. 1943. V. 28. P. 114–138.
8. Maruyama T. Effective number of alleles in a subdivided population. *Theor. Popul. Biol*. 1970. V. 1. P. 273–306.
9. Slatkin M. Inbreeding coefficients and coalescence times. *Genet. Res*. 1991. V. 58. P. 167–175.
10. Strobeck C. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics*. 1987. V. 117. P. 149–153.
11. Guillot G. On the inference of spatial structure from population genetics data. *Bioinformatics*. V. 25 P. 1796–180.
12. Wilkins J.F., Wakeley J. The Coalescent in a continuous, finite, linear population. *Genetics*. 2002. V. 161. P. 873–888.
13. Hudson R.R., Slatkin M., Maddison W.P. Estimation of levels of gene flow from DNA sequence data. *Genetics*. 1992. V. 132. P. 583–589.
14. Bocquet-Appel J.P., Bacro J.N. Isolation by distance, trend surface analysis, and spatial autocorrelation. *Human Biology.* 1993. V. 65. P. 11–27.
15. Slatkin M, Maddison W.P. Detecting isolation by distance using phylogenies of genes. *Genetics*. 1990. V. 126. P. 249–260.
16. Jensen J.L., Bohonak A.J., Kelley S.T. Isolation by distance, web service. *BMC Genetics*. 2005. V. 6. P. 13.
17. Carvajal-Rodríguez A. GENOMEPOP: A program to simulate genomes in populations. *BMC Bioinformatics*. 2008. V. 9.
18. Laval G., Excoffier L. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*. 2004. V. 20. P. 2485–2487.
19. Anderson C.N.K., Ramakrishnan U., Chan Y.L, Hadly E.A. Serial SimCoal: A population genetics model for data from multiple populations and points in time. *Bioinformatics*. 2005. V. 21. P. 1733–1734.
20. Landguth E.L., Cushman, S.A. CDPOP: A spatially-explicit cost distance population genetics program. *Molecular Ecology Resources.* 2010. V. 10. P. 156–161.
21. Strand A.E, Niehaus J.M. KERNELPOP, a spatially explicit population genetic simulation engine. *Mol. Ecol. Notes*. 2007. V. 7. P. 969–973.

22. Strand A.E. Metasim 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Mol. Ecol. Notes*. V. 2002. V. 2. P. 373–376.
23. Balloux F. EASYPOP (version 1.7): a computer program for population genetics simulations**.** *J. Hered***.** 2001. V. 92. P. 301–302.
24. Гоманенко Г.В., Камалтынов Р.М., Кузьменкова Ж.В., Беренос К., Щербаков Д.Ю. Популяционная структура байкальского бокоплава *Gmelinoides fasciatus* (Stebbing). *Генетика*. 2005. Т. 41. № 8. С. 1108–1114.
25. Перетолчина Т.Е., Букин Т.Я., Ситникова Т.Я., Щербаков Д.Ю. Генетическая дифференциация эндемичного байкальского вида *Baicalia carinata* (Mollusca: Caenogastropoda). *Генетика*. 2007. Т. 43. № 12. С. 1–9.
26. Mashiko K., Kamaltynov R.M., Sherbakov D.Yu., Mori-no H. Genetic separation of gammarid (*Eulimnogammarus cyaneus*) populations by localized topographic changes in ancient Lake Baikal. *Archive Hydrobiology*. 1997. V. 139. № 3. P. 379–387.
27. Mashiko K., Kamaltynov R.M., Morino H., Sherbakov D.Yu. Genetic differentiation among gammarid (*Eulimnogammarus cyaneus*) populations in Lake Baikal, East Siberia. *Archive Hydrobiology*. 2000. V. 148. № 2. P. 249–261.
28. Nevado B., Mautner S., Sturmbauer C., Verheyen E. Water-level fluctuations and metapopulation dynamics as drivers of genetic diversity in populations of three Tanganyikan cichlid fish species. *Mol Ecol*. 2013. V. 22. № 15. P. 3933–3948.
29. Duftner N., Sefc K.M., Koblmüller S., Nevado B., Verheyen E., Phiri H., Sturmbauer C. Distinct population structure in a phenotypically homogeneous rock-dwelling cichlid fish from Lake Tanganyika. *Mol Ecol*. 2006. V. 15. № 9. P. 2381–2395.
30. Kelly R.P, Palumbi S.R. Genetic Structure Among 50 Species of the northeastern pacific rocky intertidal community**.** *PLoS ONE*. 2010. V. 5. P. e8594.
31. Семовский С.В., Букин Ю.С., Щербаков Д.Ю. Видообразование в одномерной популяции: адаптивная динамика и нейтральная эволюция. *Исследовано в России*. 2002. Т. 125. С. 1385–1396, URL: http://zhurnal.ape.relarn.ru/articles/2002/125.pdf, http://lin.irk.ru/pdf/5847.pdf.
32. Semovski S.V., Bukin Yu.S., Sherbakov D.Yu. Speciation and neutral molecular evolution in one-dimensional closed population. *Int. J. of Modern Physics C*. 2003. V. 14. № 7. P. 973–983.
33. Bukin Ju.S., Pudovkina T.A., Sherbakov D.Ju., Sitnikova T.Ya. Genetic flows in a structured one-dimensional population: simulation and real data on Baikalian polychaetes *M. Godlewskii*. *In Silico Biology*. 2007. V. 7. № 3. P. 277–284.
34. Свирежев Ю.М., Логофет Д.О. *Устойчивость биологических сообществ*. М.: Наука, 1987. 353 с.
35. Doebeli M., Dieckmann U. Speciation along environmental gradients. *Nature*. 2003. V. 421. P. 259–264.
36. Doebeli M., Dieckmann, U. Evolutionary branching and sympatric speciation caused by different types of ecological interactions. *Am. Nat*. 2000. V. 156. P. 77–101.
37. Kondrashov A.S. Multilocus model of sympatric speciation III. Computer simulations. *Theor. Pop. Biol.* 1986. V. 29. P. 1–15.
38. Dieckmann U., Doebeli M., J. Metz J.A.J., Tautz D. *Adaptive Speciation*. Cambridge University Press, 2004. 445 p.
39. Semovski S.V., Verheyen E., Sherbakov D.Y. Simulating the evolution of neutrally evolving sequences in a population under environmental changes. *Ecological Modelling*. 2004. V. 176. P. 99–107.
40. Семовский С.В., Букин Ю.С. Щербаков Д.Ю. Модели симпатрического видообразования в изменяющихся условиях среды. *Сибирский экологический журнал*. 2004. Т. 5. С. 621–627.
41. Тихонов А.Н., Самарский А.А. *Уравнения математической физики* (*5-е изд.*). М.: Наука, 1977. 735 с.

42. Tamura K., Peterson D., Peterson N., Stecher G, Nei M., Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods**.** *Molecular Biology and Evolution*. 2011. V. 28. P. 2731–2739.
43. Kumar S., Stecher G., Peterson D., Tamura K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics*. 2012. V. 28. P. 2685–2686.

Материал поступил в редакцию 08.07.2014, опубликован 04.12.2014.