ГЕНЕТИКА, 2021, том 57, № 11, с. 1318–1329

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ

УДК 303.725.23:575.177:303.725.23

К АНАЛИЗУ СЛУЧАЙНЫХ ПРОЦЕССОВ ИЗОНИМИИ. II. ДИНАМИКА ДИВЕРГЕНЦИИ ПОПУЛЯЦИЙ

© 2021 г. В. П. Пасеков*

Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра "Информатика и управление" Российской академии наук, Москва, 119991 Россия

**e-mail: pass40@mail.ru*

Поступила в редакцию 03.02.2021 г.
После доработки 21.04.2021 г.
Принята к публикации 27.04.2021 г.

Рассматривается случайная динамика фамильного состава популяции конечного размера в дискретном времени при неперекрывающихся поколениях. Наследование фамилий предполагается патрилинейным. Динамика анализируется на малом промежутке эффективного времени $t/NE(t)$, где $NE(t)$ – средний гармонический эффективный размер популяции за t поколений. Так как при этом систематическими давлениями можно пренебречь, то в результате фамильная микроэволюция приближенно соответствует процессу случайного генного дрейфа, синхронно протекающему в той же самой популяции с вчетверо меньшей, чем для фамилий, интенсивностью. Подобно модели дрейфа генов фамильный состав очередного поколения τ является случайной выборкой с возвращением размера $Ne(\tau)/2$ из фамилий мужской составляющей родительской популяции, т. е. размера в 4 раза меньшего, чем выборка $2Ne(\tau)$ гамет при дрейфе генов ($Ne(\tau)$ – эффективный размер популяции в поколении τ). Изучается динамика вероятности случайной встречи индивидуума с однофамильцем и поведение вероятности встречи индивидуумов с разными фамилиями. Данные вероятности аналогичны гомозиготам и гетерозиготам соответственно при анализе генетической структуры. Приведены точные зависимости от времени для указанных вероятностей, дисперсий концентраций фамилий и фамильного аналога коэффициента инбридинга. Дана аппроксимация точных зависимостей более простыми при малой величине эффективного времени $t/NE(t)$, где дивергенция фамилий идет вчетверо быстрей, чем генная. Результаты не предполагают монофилии фамилий и описывают умозрительную теоретическую совокупность популяций-реплик, как бы прошедших заново микроэволюционную историю рассматриваемой популяции в тех же условиях. Использование малого по сравнению с размером популяции времени оправдано недавним возникновением большинства фамилий в России и тем, что прошедшее время в поколениях много меньше типичных размеров популяций. В реальных подразделенных популяциях процедуры оценивания коэффициента инбридинга по концентрациям фамилий не позволяют различить ситуации механической смеси субпопуляций или их общего происхождения.

Ключевые слова: теоретическая популяционная генетика, популяции с неперекрывающимися поколениями, коэффициент инбридинга, дрейф фамилий, случайный генный дрейф, математические модели, динамика дивергенции фамильного состава, асимптотика на малом эффективном времени.

DOI: 10.31857/S0016675821110114

Фамильный состав популяции отражает ее этническую принадлежность, происхождение, генеалогию (всем хотелось бы разузнать побольше о своих корнях) и он любопытен сам по себе [1, 2]. Известно, что между передачей потомкам фамилии и родительских генов имеется сходство. Для нас наибольший интерес представляет связь фамильной и генетической структур [3–8].

Всплеск публикаций в этом направлении появился после статьи [3], посвященной оценке коэффициента инбридинга популяции по частоте изонимных браков (см. критические моменты использования данных по изонимным бракам в [9]). Библиография подобных работ приведена в [10]. В первой части серии [11] и в настоящей работе для анализа фамильной структуры привлекаются модификации соответствующих простейших популяционно-генетических моделей популяции ограниченного размера. Подробный анализ этой связи мы начали в первой части [11], в основном посвященной описанию структуры в статике (в фиксированный момент времени). В настоящей статье основное внимание уделено динамике фамильной структуры.

Хотя пришедший из древней Греции термин "изонимия" означает равенство всех перед законом,

в последующем расширении этого понятия под изонимией стали понимать совокупность одинаковых названий. Например, изонимными считаются браки, в которых фамилия супруга совпадает с девичьей фамилией жены (браки между однофамильцами). Применительно к области популяционной генетики широкое распространение получил метод оценивания коэффициента инбридинга популяции по частоте изонимных браков, но использование изонимии не ограничено только этим подходом.

Очевидно, изонимию также характеризуют состав и количество однофамильцев в популяции. Пусть k – количество разных фамилий, а m_i обозначает количество индивидуумов с i-й фамилией в популяции размера N. Доля i-й фамилии $p_i \equiv m_i/N$ показывает ее распространенность, а совокупность таких долей дает распределение фамилий в популяции (фамильное состояние), причем

$$\sum_{i=1}^{k} m_i = N, \quad p_i \equiv m_i/N, \quad \sum_{i=1}^{k} p_i = 1.$$

Распространенность (долю, концентрацию) однофамильцев с учетом всех фамилий можно было бы подсчитать как ее среднее значение вида

$$\left(\sum_{i=1}^{k} p_i\right)\Big/k = \sum_{i=1}^{k}(m_i/N)/k = (N/N)/k = 1/k.$$

Однако такая характеристика явно мало информативна, так как тождественно равна $1/k$ при любом распределении однофамильцев. Ее недостаток в том, что не учитывается "вес" отдельных фамилий, т.е. их разная распространенность. Например, если при $k = 2$ в популяции размера 1000 индивидуумов было 998 однофамильцев по первой фамилии и 2 по второй, то данный показатель распространенности равен 1/2 и будет таким же при одинаковом (по 500) числе однофамильцев по каждой из двух фамилий.

Подойдем к построению характеристики распространенности однофамильцев иначе. Выберем из всей популяции случайным образом некоторого индивидуума. У него окажется i-я фамилия с вероятностью p_i, очевидно равной ее концентрации в популяции. Независимо от результата первого наблюдения снова наугад выберем индивидуума. Вероятность, что у него та же фамилия, равна p_i. Значит вероятность $p_{iso,i}$ наблюдения пары однофамильцев с i-й фамилией (их случайной встречи) находится как p_i^2 независимо от монофилии фамилий (от условия, согласно которому у однофамильцев с конкретной фамилией один и тот же патрилинейный родоначальник). При суммировании по i находим вероятность Pr_{iso} для индивидуума случайно встретить в популяции своего однофамильца с учетом всех фамилий как

$$Pr_{iso} \equiv \sum_i p_{iso,i} = \sum_i p_i^2, \quad p_{iso,i} \equiv p_i^2.$$

Заметим, что если при наблюдении пар однофамильцев исключать пары типа индивидуум сам с собой, то вероятность выбрать второй раз индивидуума с i-й фамилией равна не p_i, а $(m_i - 1)/N = p_i - 1/N$. Величиной $1/N$ можно пренебречь, когда N не слишком мало. Учет поправки $1/N$ дает несмещенную оценку (см., например, [12]) вероятности встречи двух однофамильцев. Суть появления различий в том, используем ли мы выборку с возвращением (например, выборку фамилий из родительской популяции, когда отцом у каждого потомка при каждом выборе может быть любой из N мужчин, не имеющих в этом отношении преимуществ друг перед другом) или выборку без возвращения (например, при выборе следующего индивидуума из оставшихся в популяции).

Дополнительной к вероятности случайно наблюдать пару однофамильцев является вероятность случайной встречи индивидуумов с разными фамилиями, происходящих от заведомо разных родоначальников при патрилинейном наследовании фамилии и при отсутствии у них изменений ("мутаций"). Вероятность наблюдения индивидуума с i-й фамилией равна p_i, а вероятность его встречи с носителем другой фамилии $1 - p_i$, так что вероятность (Hs_i) наблюдения двух индивидуумов с i-й и иной фамилиями с учетом порядка равна $p_i(1 - p_i)$. Когда упорядоченность индивидуумов при встрече не имеет значения, такая вероятность равна $2p_i(1 - p_i)$ и складывается из $p_i(1 - p_i)$ и $(1 - p_i)p_i$, а при учете всех фамилий равна $\sum_i 2p_i(1 - p_i)$. Данную сумму можно интерпретировать как вероятность (Hs) встречи двух потомков разных родоначальников.

Согласно [11, 13] на малом по сравнению с размером популяции промежутке времени доминирующим фактором динамики является случайный дрейф. Именно на влиянии случайного дрейфа на указанном промежутке мы сосредоточимся далее. При этом будем использовать предпосылки стандартных широко использующихся популяционно-генетических моделей с неперекрывающимися поколениями. Так, наиболее известный вариант модели случайного генного дрейфа представлен моделью Райта–Фишера диплоидной популяции размера N со случайным скрещиванием и неперекрывающимися поколениями. Мы будем использовать модель с неперекрывающимися поколениями и дискретным временем для популяций человека, где присутствует перекрывание. Хотя очевидно, что такое применение противоречиво, но, например, при обработке данных по популяциям человека широко используется также выведенный для модели с неперекрывающимися поколениями закон Харди–Вайнберга. Его многократное употребление обычно подтверждает согласие закона с дан-

ными. Аналогично для анализа более сложных ситуаций в популяционной генетике используются как модели с дискретным, так и с непрерывным временем. Результаты их применения обычно не противоречат друг другу и могут рассматриваться как соответствующие аппроксимации.

При использовании далее модели со случайным скрещиванием Райта–Фишера генетическое состояние популяции в отношении аутосомного локуса с множественными аллелями описывается вектором концентраций аллелей, которые равноценны в том смысле, что у них нет преимуществ друг перед другом в передаче потомкам. В следующем поколении генетический состав популяции с численностью N представляет собой случайную выборку с возвращением размера $2N$ из пула гамет родительского поколения (по две гаметы на одного потомка).

При отсутствии давления систематических факторов, неслучайно изменяющего генетическую структуру, эта выборка извлекается из совокупности аллелей, в свою очередь полученных в предыдущем поколении как выборка аллелей из своего родительского поколения. Значит динамика состояния популяции представляет собой последовательность вложенных выборок, и в каждый момент времени нет систематического тренда в распространении одних и вымирании других аллелей.

Что касается модели динамики фамильной структуры, то фамилии равноценны в отношении передачи потомкам, предполагается их патрилинейное наследование и изменения (мутации) фамилий отсутствуют. После заключения брака жена принимает фамилию мужа. В результате фамильная структура женщин (матерей для следующего поколения) просто дублирует структуру мужчин (отцов) и не несет дополнительной информации. Более того, даже назначение женщинам любых фамилий не скажется на фамилиях следующего поколения, передающихся только от мужчин. На практике использование этой модели возможно и при отклонениях от данных предположений, но достаточно малых, чтобы ими было допустимо пренебречь. В итоге траектория состава фамилий в популяции в ряду поколений представляет собой последовательность составов вложенных выборок из фамилий мужчин соответствующих родительских поколений с объемами $N(\tau)/2$, и на нее не влияют фамилии женщин.

Однако отметим, что информация по *девичьим* фамилиям женщин может быть полезна, например, при использовании данных по изонимным бракам. Если дополнительно потребовать, чтобы фамилии комбинировались в брачных парах независимо, то вероятность изонимного брака совпадает с вероятностью случайной встречи двух однофамильцев.

ДИНАМИКА ФАМИЛЬНОГО СОСТОЯНИЯ ПОПУЛЯЦИИ

Состав фамилий в следующем поколении является случайной выборкой с возвращением из совокупности фамилий мужчин родительского поколения. Динамика фамильного состава популяции представляет собой последовательность вложенных выборок с возвращением из совокупности фамилий в соответствующих родительских поколениях. Каждая выборка определяет фамильный состав популяции на очередном шагу, и при равном соотношении полов ее размер совпадает с размером $N(\tau)/2$ мужской части популяции с общей численностью $N(\tau)$ в следующем поколении τ. Таким образом, динамика фамильной структуры формально совпадает с процессом генного дрейфа, синхронно идущего в той же самой популяции, но *при вчетверо меньших размерах выборок* ($N(\tau)/2$ для дрейфа фамилий и $2N(\tau)$ для дрейфа генов). При этом роль множественных аллелей локуса играют фамилии индивидуумов.

Уточним информацию о фамильном составе выборки. Он представляет собой результат $N(\tau)/2$ испытаний, а если рассматривать отдельную фамилию, то результат биномиальных испытаний. У нас “испытанию” соответствует рождение потомка (мужского пола, так как женщин не рассматриваем). У потомка будет i-я фамилия с вероятностью p_i, где p_i – ее концентрация в поколении родителей, или какая-либо другая (с вероятностью $1 - p_i$). По свойствам биномиальных испытаний с вероятностью успеха p_i у концентрации успехов x_i в выборке размера K будут следующие характеристики:

$$E\{x_i\} = E\{x_i|p_i\} = p_i,$$
$$V(x_i) \equiv E\left\{(x_i - E\{x_i\})^2\right\} = p_i(1 - p_i)/K. \quad (1)$$

Напомним, что $E\{.\}$ является символом операции получения математического ожидания случайной величины, заключенной в фигурные скобки (его можно понимать как среднее значение при неограниченном увеличении размера выборки), а $V(.)$ обозначает дисперсию случайной величины в круглых скобках.

Из (1) следует, что случайный дрейф как последовательность вложенных выборок не имеет тенденций в динамике концентраций “успехов”, оставляя их в среднем неизменными (т.е. у нас концентрации фамилий в среднем остаются равными начальным значениям). Однако при неизменности в среднем возможные концентрации фамилий как бы расплываются вокруг начальных значений. Поэтому обратимся к свойствам динамики фамильного дрейфа.

Результат 1. *Пусть динамика вектора* $\mathbf{x}(t)$ *концентраций фамилий* $\{x_i(t)\}$ *в популяции постоянного эффективного размера* N *рассматривается как по-*

следовательность вложенных случайных выборок с возвращением размера $N/2$ из фамилий родителей. На каждом шагу вероятности выбора фамилий равны их концентрациям среди родителей. Обозначим начальные концентрации фамилий через $\mathbf{p} = \{p_i\} = \{x_i(0)\}$ *и положим, что можно пренебречь их мутациями и миграциями.*

Тогда независимо от наличия или отсутствия монофилии фамилий динамика фамильного состава $\{x_i(t)\}$ *популяции характеризуется следующим образом*:

1. *В наугад выбранной популяции-реплике из умозрительной совокупности реплик (как бы повторяющих микроэволюцию рассматриваемой популяции в тех же условиях и с теоретически возможными фамильными состояниями) вероятность* $E\{Hs(\mathbf{x}(t))\}$ *случайной встречи двух потомков разных родоначальников и вероятность* $E\{Hs_i(\mathbf{x}(t))\}$, $i = 1, 2, \ldots$ *встречи потомков с i-й фамилией и иной убывают по поколениям* $t = 1, 2, \ldots$ *с одним и тем же темпом* $2/N$ *своей величины за шаг (поколение) независимо от предположения о монофилии фамилий. При этом*

$$E\{Hs_i(\mathbf{x}(t))\} \equiv E\{x_i(t)(1-x_i(t))\} = \\ = p_i(1-p_i)(1-2/N)^t \equiv Hs_i(\mathbf{p})(1-2/N)^t \underset{t\to\infty}{\to} 0,$$

$$E\{Hs(\mathbf{x}(t))\} \equiv E\left\{1-\sum_{i=1}^{k} x_i^2(t)\right\} = \\ = Hs(\mathbf{p})(1-2/N)^t \underset{t\to\infty}{\to} 0. \qquad (2)$$

2. *В пределе с течением времени популяция будет состоять только из однофамильцев и, более того, из потомков лишь одного родоначальника.*

Доказательство. **1.** Очевидно, потомки с разными фамилиями заведомо происходят от разных родоначальников независимо от монофилии фамилий. Сначала найдем значение $E\{x_i(t)(1-x_i(t))\}$ в первом поколении при отсутствии мутаций и миграций фамилий. Согласно (1) при $x_i(0) = p_i$ ожидаемое значение $E\{x_i(t)\}$ в первом поколении равно p_i и

$$E\{Hs_i(\mathbf{x}(1))\} \equiv E\{x_i(1)(1-x_i(1))\} = \\ = E\{x_i(1)\} - E\{x_i^2(1)\} = \\ = E\{x_i(1)\} - \left(Vs(x_i(1)) + (E\{x_i(1)\})^2\right) = \\ = p_i - Vs(x_i(1)) - p_i^2 = p_i(1-p_i) - Vs(x_i(1)),$$

так как для любой случайной величины x

$$V(x) \equiv E\{(x-E\{x\})^2\} = E\{x^2\} - (E\{x\})^2, \\ E\{x^2\} = (E\{x\})^2 + V(x). \qquad (3)$$

Подставим в полученное выражение для $E\{Hs_i(\mathbf{x}(1))\}$ значение дисперсии $Vs(x_i(1))$, соответствующее выборочному дрейфу фамилий, т.е. по (1) равное $p_i(1-p_i)\Big/\frac{N}{2}$:

$$E\{x_i(1)(1-x_i(1))\} = \\ = p_i(1-p_i)(1-V(x_i(1)))/(p_i(1-p_i)) = \\ = p_i(1-p_i)(1-2/N), \quad i = 1, 2, \ldots .$$

Здесь фамилия может быть произвольной, а ожидаемое значение $E\{Hs_i(x(1))\}$ вероятности $Hs_i(x(1))$ случайной встречи потомков с разными фамилиями уменьшилось за поколение (шаг) по сравнению с предыдущей величиной $Hs_i(p)$ в $1-2/N$ раз, какими бы ни были концентрация фамилий $x(0) \equiv p$ и сами фамилии. Поэтому еще за одно поколение при любом из полученных случайных значений концентраций фамилий $\{x_i(2)\}$ вероятность $E\{Hs_i(x(2))\}$ встречи двух потомков разных родоначальников на втором шагу также уменьшится в $1-2/N$ раз, откуда за два первых поколения уменьшение будет в $(1-2/N)^2$ раз.

Таким образом, за каждое поколение вероятность $Hs(\mathbf{x}(t))$ наблюдения пары потомков разных родоначальников с учетом всех фамилий уменьшается в $1-2/N$ раз, а через t шагов (поколений) в последовательности случайных выборок из соответствующих родительских популяций вплоть до родоначальной имеем

$$E\{Hs(\mathbf{x}(t))\} = Hs(\mathbf{p})(1-2/N)^t \to 0 \text{ при } t \to \infty.$$

Если эффективная численность изменяется и в поколении τ равна $Ne(\tau)$, то

$$E\{Hs(\mathbf{x}(t))\} = Hs(\mathbf{p})\prod_{\tau=1}^{t}(1-2/Ne(\tau)).$$

2. Когда среди конечного количества возможных вариантов реализации фамильного состава популяции среднее значение для неотрицательной вероятности $x_i(1-x_i)$ наблюдения двух потомков разных родоначальников равно нулю, то и на каждой реализации $x_i(1-x_i)$ будет равно нулю, что означает случайную утерю фамилий. В итоге вся популяция будет состоять из однофамильцев, в том числе и при росте размера популяции до некоторого предела, а также при отсутствии монофилии фамилий. В конечном итоге все индивидуумы будут потомками одного родоначальника. ◄

Следствие 2. *Динамика характеристик фамильной структуры популяции в рамках предположений предыдущего результата описывается следующими зависимостями от времени*:

$$Vs(x_i(t)) = p_i(1-p_i)(1-(1-2/N)^t) = \\ = Hs_i(\mathbf{p}) - E\{Hs_i(\mathbf{x}(t))\} \underset{t\to\infty}{\to} p_i(1-p_i) = Hs_i(\mathbf{p}),$$

$$E\left\{x_i^2(t)\right\} = p_i - p_i(1-p_i)(1-2/N)^t = \\ = p_i - E\{Hs_i(\mathbf{x}(t))\} \underset{t\to\infty}{\to} p_i, \quad i = 1, 2, \ldots k, \qquad (4)$$
$$Fs(\mathbf{x}(t)) = 1 - (1-2/N)^t \underset{t\to\infty}{\to} 1.$$

Здесь $Vs(x_i(t))$ – дисперсия концентрации x_i у i-й фамилии в теоретической совокупности популяций-реплик, $E\left\{x_i^2(t)\right\}$ – вероятность случайной встречи двух индивидуумов с i-й фамилией в популяции, случайно выбранной из теоретической совокупности, $Fs(\mathbf{x}(t))$ – фамильный аналог коэффициента инбридинга популяции в поколении t (в момент времени t).

Доказательство. 1. Для отыскания $Vs(x_i)$ воспользуемся выражением $Vs(x_i)$ через $E\{x_i(1-x_i)\}$ как при выводе (3) и (2):

$$Vs(x_i(t)) = p_i(1-p_i) - E\{x_i(t)(1-x_i(t))\} = \\ = p_i(1-p_i) - p_i(1-p_i)(1-2/N)^t = \\ = p_i(1-p_i)\left(1-(1-2/N)^t\right) \to \\ \to p_i(1-p_i) = Hs_i(\mathbf{p}) \text{ при } t \to \infty.$$

2. Вероятность случайной встречи однофамильцев с i-й фамилией при условии ее концентрации $x_i(t)$ в фиксированной популяции равна $x_i^2(t)$. При случайном выборе популяции-реплики вероятность такой встречи согласно (3) находится как $Vs(x_i) + (E\{x_i\})^2$, где подстановка $Vs(x_i) = p_i(1-p_i) - p_i(1-p_i)(1-2/N)^t$ из п. 1 дает

$$E\left\{x_i^2(t)\right\} = Vs(x_i) + (E\{x_i\})^2 = \\ = p_i - p_i(1-p_i)(1-2/N)^t = p_i - E\{Hs_i(\mathbf{x}(t))\} \underset{t\to\infty}{\to} p_i.$$

3. Из п. 1 следует, что

$$E\{x_i(1-x_i)\} = p_i(1-p_i) - Vs(x_i) = \\ = p_i(1-p_i)(1 - Vs(x_i)/(p_i(1-p_i))).$$

Правую часть данного выражения можно записать как $p_i(1-p_i)(1-Fs)$, где коэффициент $Fs \equiv Vs(x_i)/(p_i(1-p_i))$ является аналогом случайного коэффициента инбридинга популяции F (точнее F_{ST}) для данных по фамилиям. Поскольку $Fs(x_i(t)) \equiv Fs(t) = Vs(x_i(t))/(p_i(1-p_i))$, то подстановка сюда найденного выражения для $Vs(x_i)$ дает $Fs(t) = 1 - (1-2/N)^t$ и $Fs(t) \to 1$ при $t \to \infty$. ◄

В более реальной ситуации с изменениями размера популяции вместо $(1-2/N)^t$ будет $\prod_{\tau=1}^{t}(1-2/Ne(\tau)) \to 0$ при $t \to \infty$, если у $Ne(\tau)$ существует конечный предел или $Ne(\tau)$ ограничено сверху константой. Отметим, что формулы (4) являются точными для модели случайного дрейфа, а далее займемся их аппроксимациями.

КОЭФФИЦИЕНТ ИНБРИДИНГА ПОПУЛЯЦИИ И СВОЙСТВА ДРЕЙФА ФАМИЛИЙ ПРИ МАЛОМ ЭФФЕКТИВНОМ ВРЕМЕНИ ДИВЕРГЕНЦИИ

Во многих исследованиях внимание фокусируется на микроэволюции, и желательно найти удобную и несложную аппроксимацию процессов динамики на небольших периодах эффективного времени $2t/N$. В отношении дрейфа фамилий это особенно актуально, так как подавляющее количество фамилий появилось в России сравнительно недавно после первой Всероссийской переписи населения в 1897 г., хотя они встречались еще в новгородских летописях. Отметим, что небольшой период должен быть таковым лишь по сравнению с размером популяции, достигающим сотен, тысяч и более. Так как мы время измеряем в поколениях, то для человека это уже тысячи и десятки тысяч лет.

Рассмотрим асимптотику дивергенции состояния одной популяции от начального значения в терминах Hs (или H) и Fs (F) на малом промежутке эффективного времени в поколениях. В стационарных условиях среды зависимость от времени t это зависимость от $\mathbf{x}(t)$, например $F(t) = F(\mathbf{x}(t))$. В популяционной генетике вероятности Hs соответствует концентрация H гетерозигот в случайно скрещивающейся популяции, и по формуле С. Райта

$$E\{H(\mathbf{x}(t))\} = H(\mathbf{p})(1 - F(\mathbf{x}(t))),$$
$$\text{откуда } F(\mathbf{x}(t)) = \left(H(\mathbf{p}) - E\{H(\mathbf{x}(t))\}\right)/H(\mathbf{p}),$$

где $H(\mathbf{p})$ – начальное значение концентрации гетерозигот, $E\{H(\mathbf{x}(t))\}$ – ожидаемая концентрация гетерозигот в теоретической совокупности популяций-реплик с возможными генетическими состояниями $\mathbf{x}(t)$ при заданных условиях существования, $F(\mathbf{x}(t))$ – коэффициент инбридинга популяции (см., скажем, [12]).

Аналог коэффициента инбридинга F в случае фамилий мы обозначили ранее как Fs. Случайный генный дрейф отличается от фамильного дрейфа лишь тем, что для первого изменение состояния при смене поколений происходит за счет выборки размера $2N$ для аллелей, а для фамилий размер выборки $N/2$ вчетверо меньше. Значит, подстановка нужного значения $2N$ или $N/2$ в соответствующие формулы дает характеристики дрейфа генов или дрейфа фамилий. К сожалению, по данным обследования популяций, относящихся к одному моменту времени, входящие в эти формулы начальные значения $H(\mathbf{p}) = H(0)$ и $Hs(0)$ неизвестны.

Для дрейфа фамилий в случае постоянной по поколениям численности популяции N зависимость $Hs(\mathbf{x}(t))$ для фамилий согласно (2) довольно проста: $E\{Hs(\mathbf{x}(t))\} = Hs(\mathbf{p})(1-2/N)^t$, а для гетеро-

зигот $E\{H(\mathbf{x}(t))\} = H(\mathbf{p})(1 - 1/2N)^t$. На малом интервале эффективного времени эти зависимости еще более упрощаются:

$$(1 - 2/N)^t = e^{t\ln(1-2/N)} \sim e^{-2t/N},$$
$$(1 - 1/2N)^t \sim e^{-t/2N} \text{ при } 1/N \ll 1.$$

Отсюда следует, например, что в популяционной генетике приближенно

$$1 - F(t) = e^{-t/2N},$$
$$F(t) = 1 - e^{-t/2N} \sim t/2N, \quad t/2N \ll 1.$$

Уточним полученные выражения для ситуации неслучайного изменения численности N во времени и нарушения некоторых предпосылок модели случайного дрейфа. В таком случае размер популяции в каждом поколении τ следует заменить на эффективный размер $Ne(\tau)$) (см., например, [14, 15]), а при учете t поколений заменить N на средний гармонический эффективный размер популяции $\tilde{N}e(t)$ за весь период дивергенции t (см. нижеследующее замечание). При этом асимптотически на малом эффективном времени $t/2\tilde{N}e(t)$ зависимости рассматриваемых характеристик от $t/2\tilde{N}e(t)$ оказываются линейными.

Замечание 3. *Пусть в каждом поколении* $\tau = 1, 2, \ldots$ *фамильное состояние популяции с неперекрывающимися поколениями формируется как случайная выборка с возвращением размера* $\tilde{N}e(t)/2$ *из совокупности фамилий родителей, а вероятности выбора фамилий равны их концентрациям среди родителей.*

Тогда в t-м поколении при $\tilde{N}e(t)/t \to \infty$ ($t/\tilde{N}e(t) \to 0$) *асимптотически*

$$E\{Hs(\mathbf{x}(t))\} \sim Hs(\mathbf{x}(0))(1 - 2t/\tilde{N}e(t)) = Hs(\mathbf{x}(0))(1 - Fs(\mathbf{x}(t))),$$

$$Fs(\mathbf{x}(t)) \sim 2t/\tilde{N}e(t), \quad \tilde{N}e(t) \equiv t\Big/\sum_{\tau=1}^{t}(1/Ne(\tau)),$$

$$F(\mathbf{x}(t)) \sim 1/4 \cdot Fs(\mathbf{x}(t)) \sim t/2\tilde{N}e(t). \qquad (5)$$

Здесь $\tilde{N}e(t)$ *– средний гармонический эффективный размер популяции за t поколений дивергенции, F – случайный коэффициент инбридинга.*

Доказательство. В случае фамилий величина $E\{Hs_i(\mathbf{x}(t))\} \equiv E\{x_i(t)(1 - x_i(t))\}$ уменьшается за поколение τ в $1 - 2/Ne(\tau)$ раз согласно ранее доказанному. Значит через 2 поколения $E\{Hs_i(\mathbf{x}(2))\} = Hs(\mathbf{x}(0))(1 - 2/Ne(1))(1 - 2/Ne(2))$, а через t поколений

$$E\{x_i(t)(1 - x_i(t))\} = p_i(1 - p_i)\prod_{\tau=1}^{t}(1 - 2/Ne(\tau)) =$$
$$= p_i(1 - p_i)e^{\sum_{\tau=1}^{t}\ln(1-2/Ne(\tau))} \sim p_i(1 - p_i)e^{-2\sum_{\tau=1}^{t}1/Ne(\tau))} \approx$$
$$\approx p_i(1 - p_i)e^{-2t/\tilde{N}e(t)} \sim p_i(1 - p_i)(1 - 2t/\tilde{N}e(t)),$$
$$t/\tilde{N}e(t) \ll 1,$$

поскольку $\ln(1 - 2/Ne(\tau)) \sim -2/Ne(\tau)$, $\sum_{\tau=1}^{t}1/Ne(\tau)) \approx t/\tilde{N}e(t)$ при малом значении $t/\tilde{N}e(t)$. В результате суммирования $E\{Hs_i(\mathbf{x}(t))\}$ по i при $t/\tilde{N}e(t) \ll 1$ получаем

$$E\{Hs(\mathbf{x}(t))\} \sim Hs(\mathbf{p})(1 - 2t/\tilde{N}e(t)) = Hs(\mathbf{p})(1 - Fs(t)),$$
$$Fs(t) \sim 2t/\tilde{N}e(t), \quad t/\tilde{N}e(t) \ll 1 \ (F,\ Fs \ll 1).$$

Очевидно, что при увеличении $\tilde{N}e(t)$ в 4 раза согласно размеру выборки гамет мы получим соответствующие результаты для генного дрейфа, и $F = 1/4 \cdot Fs = t/2\tilde{N}e(t)$. ◄

Таким образом, *коэффициент инбридинга F приблизительно в 4 раза меньше его фамильного аналога Fs*. Однако это простое соотношение становится неверным при большом количестве поколений t, отделяющих рассматриваемую популяцию от родоначальной (когда коэффициент инбридинга нельзя считать малым). Например, и тот и другой показатели в пределе с течением времени теоретически станут равны единице и не будут отличаться в 4 раза.

Выражение $2t/\tilde{N}e(t)$ показывает, что параметр $\tilde{N}e$ играет роль масштабирования времени. Например, возрастание его в 10 раз равносильно уменьшению времени в 10 раз и такому же уменьшению скорости его течения. Подобный результат верен для любой фамилии. Поэтому можно говорить, что *при относительно малом количестве поколений процесс дрейфа фамилий отличается от генного дрейфа тем, что для первого из них время течет приближенно в 4 раза быстрее.*

Ремарка 4. Выражение $2t/\tilde{N}e(t)$ для процесса дрейфа фамилий мы назвали *эффективным временем.* Мы видим, что оно входит в асимптотические формулы в качестве одного неразделимого параметра. Так как асимптотически $Fs(t) \sim 2t/\tilde{N}e(t)$ согласно (5), то *фамильный коэффициент инбридинга совпадает с эффективным временем*. Он монотонно увеличивается вместе с привычным временем в силу роста инбредности в последовательности поколений. Поэтому можно сказать, что приведенные *асимптотические формулы верны при малом коэффициенте инбридинга* (что типично для популяций человека).

Подчеркнем еще раз, что сказанное справедливо на относительно малом промежутке времени. Когда численность популяции изменяется по поколениям и допускаются некоторые отклонения от идеализированных предпосылок модели случайного дрейфа, то течение процесса дрейфа лучше характеризуется эффективным временем, в котором в качестве численности популяции N фигурирует ее средний гармонический эффективный размер $\tilde{N}e(t)$, учитывающий особенности репродукции и структуры популяции, а также динамику ее численности.

Следствие 5. *Пусть фамильное состояние популяции с неперекрывающимися поколениями формируется как случайная выборка с возвращением размера* $Ne(\tau)/2$ *из фамилий родителей в каждом поколении* $\tau = 1, 2, \ldots$, *и вероятности выбора фамилий равны их концентрациям среди родителей.*

Тогда асимптотически при $\tilde{N}e(t)/t \to \infty$ ($t/\tilde{N}e(t) \to 0$) *дисперсия концентраций фамилий* $Vs(x_i(t))$, $i = 1, 2, \ldots$ *линейно зависит от эффективного времени как*

$$Vs(x_i(t)) \sim p_i(1-p_i) \times 2t/\tilde{N}e(t) = Hs_i(\mathbf{x}(\mathbf{p})) \times 2t/\tilde{N}e(t) = Hs_i(\mathbf{x}(\mathbf{p})) \times Fs(t), \quad (6)$$

$$\tilde{N}e(t) \equiv t \Big/ \sum_{\tau=1}^{t} 1/Ne(\tau).$$

Здесь $Ne(\tau)$ *– дисперсионный* (*см., например,* [14]) *эффективный размер популяции в поколении* τ, $\tilde{N}e(t)$ *– средняя гармоническая эффективная численность популяции за t поколений дивергенции.*

Доказательство. Дисперсия $Vs(x_i(t))$ концентрации i-й фамилии характеризует дивергенцию популяций-реплик в теоретической совокупности мыслимых вариантов фамильного состава популяции. Согласно (4) и (5) она представима в виде

$$Vs(x_i(t)) = Hs_i(\mathbf{p}) - E\{Hs_i(\mathbf{x}(t))\} \sim Hs_i(\mathbf{p}) - Hs_i(\mathbf{x}(\mathbf{p}))(1 - 2t/\tilde{N}e(t)) = Hs_i(\mathbf{x}(\mathbf{p})) \times 2t/\tilde{N}e(t). \blacktriangleleft$$

Следствие 6. *Пусть в рамках предыдущего следствия размеры выборок фамилий при каждой смене поколений* τ *достаточно велики для аппроксимации распределений выборочных отклонений* $\{\delta_i(\tau)\}$ *концентрации i-й фамилии нормальными.*

Тогда в t-м поколении распределение этой концентрации $x_i(t)$ *приближенно будет нормальным со средним значением* p_i *и дисперсией* $p_i(1-p_i) \times 2t/\tilde{N}e(t)$.

Доказательство. Значение концентрации $x_i(t)$ у i-й фамилии в t-м поколении представляет собой при $\tilde{N}e(t)/t \to \infty$ сумму приближенно нормально распределенных некоррелирующих выборочных отклонений $\delta_i(\tau)$ с нулевыми средними плюс фиксированная начальная концентрация p_i. Поэтому аппроксимацией распределения $x_i(t)$ также будет нормальное распределение со средним значением $p_i = x_i(0)$ и найденной дисперсией $p_i(1 - p_i) \times 2t/\tilde{N}e(t)$. ◄

К сожалению, здесь по-прежнему фигурирует начальная концентрация p_i, которую нельзя найти по данным о текущем состоянии популяции (как, впрочем, и $\tilde{N}e(t)$).

ОПИСАНИЕ НЕЗАВИСИМОЙ ДИВЕРГЕНЦИИ ДВУХ ПОПУЛЯЦИЙ

До сих пор мы рассматривали дивергенцию фамильного состояния одной популяции от начального состояния. В практическом отношении, возможно, более интересно исследование дивергенции друг от друга нескольких популяций, имеющих общее происхождение. Примером могут служить популяции, состояния которых соответствуют одному и тому же моменту времени (часто моменту обследования). При общем происхождении дивергенция состояний таких популяций напоминает дивергенцию в теоретической совокупности популяций-реплик. Ее основные черты отражают начальный этап микроэволюции популяций, когда доминирующим фактором является случайный дрейф.

Рассмотрим какую-либо пару популяций, возникшую при разделении t поколений тому назад участка родословного древа на две ветви, существующие далее изолированно. Некоторые свойства фамильной дивергенции одной популяции от начального состояния мы уже рассмотрели. Фамильное состояние другой популяции аналогично претерпевает независимые изменения в результате процесса случайного дрейфа. В итоге наблюдается фамильная дивергенция популяций друг от друга. Независимо от степени дивергенции (ее длительности и размеров популяций) имеют место следующие соотношения для концентраций $x^{(1)}(t)$ и $x^{(2)}(t)$ какой-либо фамилии в первой и второй популяциях соответственно в поколении t при одинаковой у них начальной концентрации p:

$$E\{x^{(1)}(t)\} = E\{x^{(2)}(t)\} = p,$$

$$E\{x^{(1)}(t)\,x^{(2)}(t)\} = E\{x^{(1)}(t)\} \times E\{x^{(2)}(t)\} = p^2,$$

$$E\{x^{(1)}(t) - x^{(2)}(t)\} = E\{x^{(1)}(t)\} - E\{x^{(2)}(t)\} = 0.$$

Данные соотношения (кроме свойства произведения) верны для любого множества популяций с общим происхождением *без предположения о независимости дивергенции*. Таким образом, приведенные показатели не дают информации о течении процесса дивергенции (не зависят от времени). Эту зависимость характеризует, например, средний квадрат расстояния (дисперсия) между концентрациями фамилии $x^{(1)}(t)$ и $x^{(2)}(t)$,

который *при независимых изменениях* $x^{(1)}(t)$ и $x^{(2)}(t)$ равен сумме их дисперсий и согласно (5)–(6) при малом эффективном времени дивергенции находится как:

$$E\left\{\left(x^{(1)}(t)-x^{(2)}(t)\right)^2\right\}=Vs\left(x^{(1)}(t)\right)+Vs\left(x^{(2)}(t)\right)\sim$$
$$\sim p(1-p)\left(2t/\tilde{N}e_1(t)+2t/\tilde{N}e_2(t)\right)=$$
$$=p(1-p)\times 4t/\tilde{N}e(t)=p(1-p)\times$$
$$\times\left(Fs^{(1)}(t)+Fs^{(2)}(t)\right)=p(1-p)\times 2\bar{F}s(t).$$

Здесь $\tilde{N}e(t)=2/(1/\tilde{N}e_1(t)+1/\tilde{N}e_2(t))$ – среднее гармоническое значение для пары $\tilde{N}e_1(t)$ и $\tilde{N}e_2(t)$, $\bar{F}s(t)$ – среднее арифметическое значение коэффициентов фамильного инбридинга $Fs^{(1)}(t)$ и $Fs^{(2)}(t)$.

ОБСУЖДЕНИЕ

Отметим, что в приведенном анализе не принималась во внимание возможная инбредность общего предка в точке разделения ветвей, ведущих к рассматриваемой паре популяций. Поэтому, вообще говоря, в полученной выше формуле фигурируют не коэффициенты фамильного инбридинга популяций, а скорее их приращения за период t поколений от момента разветвления.

Дивергенция по отдельной фамилии между двумя популяциями на родословном древе (средний квадрат расстояния между концентрациями фамилий, т.е. дисперсия возможных различий концентраций) пропорциональна времени дивергенции или среднему арифметическому значению коэффициентов фамильного инбридинга популяций. С другой стороны, дивергенция обратно пропорциональна средней гармонической численности $\tilde{N}e(t)=2/(1/\tilde{N}e_1(t)+1/\tilde{N}e_2(t))$ популяций. Полученные зависимости подтверждают интуитивно ожидаемый качественный характер связей (их направления), но даны количественно.

К сожалению, здесь фигурирует обычно неизвестное начальное состояние p у концентраций фамилий, влияния которого желательно избежать. В точках ветвления, играющих роль ближайшего общего предка соответствующих пар, концентрации фамилии также отличаются от значений в корне родословного древа. Поэтому коэффициенты $p(1-p)$ в формуле для ожидаемого квадрата расстояния принимают случайные значения, свои для каждого узла ветвления, и у конкретного родословного древа их трудно учесть.

Коснемся проблем, возникающих при анализе данных, получаемых в результате обследования генетической структуры реальных популяций. В реальной подразделенной популяции, рассматриваемой как теоретическая совокупность, вместо начального состояния (математического ожидания) обычно используют среднее значение для распределения концентраций аллелей по субпопуляциям. При этом получают оценку коэффициента инбридинга в качестве *статистической корреляции* гомологичных генов объединяющихся гамет как в случае субпопуляций с общим происхождением, так и в случае произвольной группы субпопуляций, у которых нет идентичных по происхождению генов. Найденная *оценка не позволяет выявить, какой из этих случаев имеет место*, но статистическая корреляция сама по себе имеет важное значение. Она влияет на концентраций генотипов и тем самым результаты естественного и искусственного отбора, уровень наследственной отягощенности и др.

Обратим внимание на специфические черты процесса случайного выборочного дрейфа. Ожидаемым значением концентрации фамилии (аллеля) в новом поколении будет прежнее значение, т.е. *у случайного дрейфа нет преимущественного направления*. Поэтому величину дивергенции за поколение (“скорость” ненаправленной эволюции) можно измерять, скажем, дисперсией, а не просто средним отклонением, которое при ненаправленной эволюции равно нулю. При одинаковых прочих условиях выборочная дисперсия $x(t)$ при смене поколений, как характеристика скорости ненаправленной дивергенции, обратно пропорциональна размеру популяции, а также определяется текущим значением $x(t)$. Тем самым темп дивергенции для концентрации аллеля зависит от значения $x(t)$, что затрудняет интерпретацию величины наблюдаемых различий между популяциями с общим происхождением. Поэтому желательно использовать подходы, стабилизирующие темп дивергенции.

Настоящая статья не содержит каких-либо исследований с использованием в качестве объекта животных.

Настоящая статья не содержит каких-либо исследований с участием в качестве объекта людей.

СПИСОК ЛИТЕРАТУРЫ

1. *Бужилова А.П.* География русских фамилий // Восточные славяне. Антропология и этническая история. М.: Научный мир, 1999. С. 135–152.
2. *Балановская Е.В., Романов А.Г., Балановский О.П.* Однофамильцы или родственники. Подходы к изучению связи между гаплогруппами Y-хромосомы и фамилиями // Мол. биология. 2011. Т. 45. № 3. С. 473–485.
3. *Crow J.F., Mange A.P.* Measurement of inbreeding from the frequency of marriages between persons of the same surname // Soc. Biol. 1982. V. 29. № 1/2. P. 101–105.
4. *Lasker W.G.* Surnames and Genetic Structure. Cambridge: Cambr. Univ. Press, 2005. 148 p.
5. *Ревазов А.А., Парадеева Г.М., Русакова Г.И.* Пригодность русских фамилий в качестве квазигене-

тического маркера // Генетика. 1986. Т. 22. № 4. С. 699–703.
6. *Tarskaia L., El'chinova G., Scapoli C. et al.* Surnames in Siberia: A study of the population of Yakutia through isonymy // Am. J. Phys. Anthropol. 2009. V. 138. P. 190–198. https://doi.org/10.1002/ajpa.20918
7. *Сорокина И.Н., Чурносов М.И., Балтуцкая И.В., и др.* Антропогенетическое изучение населения центральной России. М.: Издательство РАМН, 2014. 336 с.
8. *Lasker G.W.* A coefficient of relationship by isonymy: A method fore stimating the genetic relationship between populations // Hum. Biol. 1977. V. 49. № 3. P. 489–493.
9. *Rogers A.R.* Doubts about isonymy // Hum. Biol. 1991. V. 63. № 5. P. 663–668.
10. *Сорокина И.Н., Рудых Н.А., Крикун Е.Н., Сокорев С.Н.* Применение фамилий в популяционно-генетических исследованиях (на примере зарубежных популяций) // Науч. ведомости БелГУ. Сер. Медицина. Фармация. 2016. № 19(240). Вып. 35. С. 5–10.
11. *Пасеков В.П.* К анализу случайных процессов изонимии. I. Структура изонимии // Генетика. 2021. Т. 57. № 10. С. 1194–1204.
12. *Вейр Б.* Анализ генетических данных: дискретные генетические признаки. М.: Мир, 1995. 400 с.
13. *Малютов М.Б., Пасеков В.П.* Об одной статистической задаче популяционной генетики // Теория вероятностей и ее применения. 1971. Т. 16. Вып. 3. С. 579–581.
14. *Хедрик Ф.* Генетика популяций. М.: Техносфера, 2003. 592 с.
15. *Ли Ч.* Введение в популяционную генетику. М.: Мир, 1978. 555 с.

ПРИЛОЖЕНИЕ

СВОДКА РЕЗУЛЬТАТОВ ДЛЯ ПОПУЛЯЦИОННО-ГЕНЕТИЧЕСКОГО АНАЛИЗА

Чтобы было удобно видеть сходство и различия основных черт фамильной и генетической картин состояния популяции, просто перечислим в одном месте разбросанные по настоящей статье (иногда неявные) результаты для популяционно-генетической модели элементарной диплоидной популяции с неперекрывающимися поколениями, ограниченной численностью и случайным скрещиванием. Пусть анализируется генетический состав диплоидной популяции по одному аутосомному локусу с k аллелями. Тогда генетическое состояние (состав) популяции можно описывать вектором $\mathbf{x} = \{x_i\}$ концентраций аллелей. Динамика состояния в общем случае обусловлена систематическими давлениями и случайным дрейфом генов. *При малом количестве поколений микроэволюции t по сравнению со средней гармонической численностью $\tilde{N}e(t)$ популяции для этого периода времени мы можем пренебречь давлением систематических факторов на генетические состояния популяции и решающим будет влияние случайного генного дрейфа.* В результате дрейфа происходит дивергенция как состояния популяции от первичного, так и популяций с общим происхождением друг от друга. Данный этап является начальным для микроэволюции исходно одинаковых родственных популяций.

Описание дивергенции в статике

В фиксированный момент времени картину дивергенции можно охарактеризовать следующими показателями. Пусть p_i начальная концентрация i-го аллеля в популяции со свободным скрещиванием, подверженной случайному генному дрейфу. Тогда через некоторый промежуток времени ожидаемая концентрация гетерозигот $E\{H_i(\mathbf{x}(t))\}$ с i-м аллелем в популяции, *случайно выбранной из умозрительной совокупности популяций-реплик* (как бы повторивших микроэволюцию рассматриваемой) с теоретически возможными состояниями, находится как

$$E\{H_i(\mathbf{x}(t))\} \equiv E\{(x_i(t)(1-x_i(t))\} = \\ = p_i(1-p_i) - V(x_i(t)) = p_i(1-p_i)(1-F(t)).$$

Здесь $E\{.\}$ символ математического ожидания (операции получения среднего значения) величины в фигурных скобках, $V(.)$ дисперсия случайной величины, стоящей в скобках (у нас дисперсия концентрации в репликах), F – коэффициент инбридинга популяции, $p_i = x_i(0)$. В то же время при условии, что концентрация у i-го аллеля в конкретной популяции этой совокупности равна x_i, доля рассматриваемых гетерозигот в ней с учетом порядка аллелей по закону Харди–Вайнберга будет теоретически равна $x_i(1-x_i)$.

Ожидаемые концентрации гетерозигот всех типов $H(\mathbf{x})$ и гомозигот в наугад выбранной популяции из умозрительной совокупности находятся как

$$E\{H(\mathbf{x})\} \equiv E\left\{1 - \sum_{i=1}^{k} x_i^2\right\} = H(\mathbf{p})(1-F),$$

$$F = \sum_{i=1}^{k} V(x_i)/H(\mathbf{p}).$$

$$E\{x_i^2\} = p_i^2 + Fp_i(1-p_i) = \\ = p_i^2 + V(x_i) = p_i^2 + FH_i(\mathbf{p}), \quad i = 1, 2, \ldots, k,$$

$$E\left\{\sum_{i=1}^{k} x_i^2\right\} = \sum_{i=1}^{k} p_i^2 + V(\mathbf{x}) = \sum_{i=1}^{k} p_i^2 + FH(\mathbf{p}) =$$
$$= 1 - H(\mathbf{p}) + FH(\mathbf{p}), \quad V(\mathbf{x}) \equiv \sum_{i=1}^{k} V(x_i).$$

Здесь $V(x_i)$ – дисперсия распределения значений $\{x_i\}$, $i = 1, 2, \ldots k$ концентраций i-го аллеля в популяциях-репликах. Коэффициент инбридинга популяции F выражается через дисперсию $V(x_i)$ концентрации отдельного аллеля x_i, а также с учетом концентраций всех аллелей в теоретической совокупности популяций-реплик как

$$F = V(x_i)/(p_i(1-p_i)), \quad i = 1, 2, \ldots, k,$$
$$F = (H_i(\mathbf{p}) - E\{H_i(\mathbf{x})\})/H(\mathbf{p}),$$
$$F = V(\mathbf{x})\Big/\left(1 - \sum_{i=1}^{k} p_i^2\right) =$$
$$= V(\mathbf{x})/H(\mathbf{p}), \quad V(\mathbf{x}) \equiv \sum_{i=1}^{k} V(x_i).$$

Таким же образом дивергенцию реальных субпопуляций в подразделенной популяции можно описать как в терминах дисперсии распределения концентрации отдельных аллелей, так и в терминах случайного коэффициента инбридинга F.

Приведенные формулы характеризуют инбредную популяцию. *Формально она является как бы подразделенной* – состоит либо из умозрительных популяций-реплик, либо из реальных субпопуляций. В последнем случае приведенные формулы в равной степени приложимы как к совокупности популяций с общим происхождением, так и к произвольному механическому набору популяций со случайным скрещиванием. При этом подразумевается, что математические ожидания и дисперсии концентраций относятся к рассматриваемой подразделенной популяции, играющей роль теоретической совокупности. Для реальной группы субпопуляций математические ожидания и дисперсии вычисляются стандартным образом как средние значения и средние квадраты отклонений от этих средних. В результате в подразделенной популяции как едином целом, несмотря на выполнение в отдельных субпопуляциях соотношений Харди–Вайнберга, эти соотношения нарушаются (эффект Валунда). В ней будет наблюдаться дефицит гетерозигот по сравнению с ожидаемым при всеобщей панмиксии. Дефициту, вызванному различиями (дисперсией) концентраций аллелей между субпопуляциями, соответствует некоторое значение случайного коэффициента инбридинга F (точнее, F_{ST}).

Как для механической смеси популяций, так и для группы популяций с общим происхождением этот коэффициент инбридинга численно равен статистической корреляции между гомологичными генами объединяющихся в генотипах гамет. Получение ее оценки *не позволяет определить с какой из ситуаций сталкивается исследователь*. Лишь в случае общего происхождения популяций при их независимой и одинаковой микроэволюционной истории статистическая корреляция прямо связана с вероятностью идентичности по происхождению пары аллелей аутосомного локуса диплоидного генотипа. В общем случае родственные популяции не являются независимыми, и процедура оценивания коэффициента инбридинга как идентичности по происхождению гомологичных аллелей должна учитывать характер происхождения популяций и условия их микроэволюции.

Отметим, что статистическая корреляция между гомологичными генами объединяющихся гамет важна, так как с ее помощью определяются концентрации генотипов и тем самым, например, изменение генетической структуры в результате отбора.

Динамика основных характеристик дивергенции

Напомним, что t – время в поколениях, $\tilde{N}e(t) \equiv t\Big/\sum_{\tau=1}^{t}(1/Ne(\tau))$ – средний гармонический эффективный размер популяции за t поколений, $Ne(\tau)$ – эффективный размер популяции в поколении τ. При достаточно малой величине *эффективного времени* $t/\tilde{N}e(t)$ модель микроэволюции аппроксимируется процессом случайного генного дрейфа. Данный процесс приводит к случайному отклонению состояния популяции от начального, росту ее коэффициента инбридинга и увеличению дивергенции возможных состояний (дивергенции популяций с общим происхождением). При этом выполняются следующие зависимости от времени основных характеристик дивергенции (точные в рамках модели)

$$E\{H_i(\mathbf{x}(t))\} \equiv E\{x_i(t)(1-x_i(t))\} =$$
$$= p_i(1-p_i)\prod_{\tau=1}^{t}(1 - 1/2Ne(\tau)) \underset{t\to\infty}{\to} 0, \quad i = 1, 2, \ldots k,$$

$$E\{H(\mathbf{x}(t))\} \equiv E\left\{1 - \sum_{i=1}^{k} x_i^2(t)\right\} =$$
$$= H(\mathbf{p})\prod_{\tau=1}^{t}(1 - 1/2\,Ne(\tau)) \underset{t\to\infty}{\to} 0.$$

$$E\left\{x_i^2(t)\right\} = p_i - p_i(1-p_i)\prod_{\tau=1}^{t}(1-1/2\,Ne(\tau)) \underset{t\to\infty}{\to} p_i,$$

$$V(x_i(t)) = p_i(1-p_i)\times$$

$$\times\left(1-\prod_{\tau=1}^{t}(1-1/2\,Ne(\tau))\right)\underset{t\to\infty}{\to} p_i(1-p_i),$$

$$i = 1, 2, \ldots k,$$

$$V(\mathbf{x}(t)) \equiv \sum_{i=1}^{k} V(x_i(t)) =$$

$$= H(\mathbf{p})\left(1-\prod_{\tau=1}^{t}(1-1/2\,Ne(\tau))\right)\underset{t\to\infty}{\to} H(\mathbf{p}),$$

$$Fs(\mathbf{x}(t)) = 1-\prod_{\tau=1}^{t}\left(1-1/2\,Ne(\tau)\right)\underset{t\to\infty}{\to} 1.$$

Здесь пределы существуют при постоянной (N) и растущей $Ne(\tau)$ численности популяции (когда, например, она ограничена сверху константой). Формулы будут проще (и более привычны) при постоянной численности N. Тогда $\prod_{\tau=1}^{t}(1-1/2Ne(\tau)) = (1-1/2N)^t$.

Аппроксимация результатов при малом эффективном времени дивергенции

На относительно малом этапе дивергенции (при малом $t/\tilde{N}e(t)$) данные зависимости еще упрощаются до линейных по эффективному времени $t/2\tilde{N}e(t)$:

$$E\{H(\mathbf{x}(t))\} \sim H(\mathbf{p})\left(1-t/2\,\tilde{N}e(t)\right) =$$

$$= H(\mathbf{p})(1-F(\mathbf{x}(t))),\quad F(\mathbf{x}(t)) \sim t/2\,\tilde{N}e(t),$$

$$V(x_i(t)) \sim p_i(1-p_i)\times t/2\,\tilde{N}e(t),$$

$$V(\mathbf{x}(t)) \sim \left(1-\sum_{i=1}^{k} p_i^2\right)\times t/2\,\tilde{N}e(t) =$$

$$= H(\mathbf{p})\times t/2\,\tilde{N}e(t) = H(\mathbf{p})F(\mathbf{x}(t)).$$

Когда численность популяции постоянна, скажем, равна N, то $\tilde{N}e(t)$ заменяется на N.

Так как асимптотически $F(t) \sim t/2\tilde{N}e(t)$, то коэффициент инбридинга совпадает с эффективным временем и монотонно увеличивается вместе с временем t в поколениях. Поэтому можно сказать, что *эффективное время измеряется величиной коэффициента инбридинга и наоборот, т.е. данная асимптотика верна при малых коэффициентах инбридинга*, присущих популяциям человека.

Случайный процесс выборочного дрейфа остается таковым в отношении произвольных подгрупп аллелей, в частности, когда одна группа состоит из единственного аллеля, скажем, с концентрацией x, а в другой все остальные. Концентрация последней группы равна $1-x$, ее можно отбросить как зависимую переменную и сфокусироваться на изучении частного случая динамики концентраций аллелей по отдельности. При этом возможно выразить коэффициент инбридинга $F(t)$ через ожидаемую концентрацию $E\{x(t)(1-x(t))\}$ гетерозигот (с учетом порядка аллелей) в случайно выбранной популяции из теоретической совокупности популяций-реплик с одной и той же демографической историей:

$$F(t) = \left(p(1-p) - E\{x(t)(1-x(t))\}\right)/p(1-p) =$$

$$= \left(H(p) - E\{H(x(t))\}\right)/H(p),\quad p = x(0).$$

Здесь $E\{x(t)(1-x(t))\}$ означает усреднение $x(t)(1-x(t))$ по возможным значениям $x(t)$ в популяциях-репликах теоретической совокупности. Отметим, что последнее выражение для $F(t)$ верно и при учете всех аллелей.

Подчеркнем, что приведенная формула *не позволяет оценить инбридинг по типичным данным только о текущих концентрациях гетерозигот* $x(t)(1-x(t))$ в изучаемой популяции. Она относится к *ожидаемой* концентрации гетерозигот в наугад выбранной популяции из *теоретического множества популяций-реплик*. Возможные генетические составы реплик случайно различаются – при повторении в одних и тех же условиях микроэволюционной истории изучаемой популяции произойдет дивергенция реплик друг от друга в силу выборочной природы случайного дрейфа с присущей ей выборочными ошибками.

Кроме того, найденное выражение для $F(t)$ *зависит от неизвестного начального состояния p обследуемой популяции* (и ее умозрительных реплик), которое нельзя найти по текущему состоянию.

Случайный коэффициент инбридинга можно также представить как

$$F(t) = V(x(t))/p(1-p) = V(x(t))/H(x(p)).$$

Здесь V межпопуляционная дисперсия концентраций аллелей *в теоретическом множестве реплик* (дисперсия возможных результатов реализаций процесса дрейфа в одних и тех же условиях с одинаковыми начальными состояниями). Данный подход также требует сведений о неизвестном начальном значении $x(0) = p$.

To Analysis of Random Processes of Isonymy. II. Dynamics of Population Divergence

V. P. Passekov*

Dorodnitsyn Computing Center of the Federal Research Center "Informatics and Management" of the Russian Academy of Sciences, Moscow, 119991 Russia

**e-mail: pass40@mail.ru*

Random dynamics of the surname composition of a population of finite size with non-overlapping generations is considered in discrete time. Inheritance of surnames is assumed to be patrilineal. Dynamics is analyzed on a small $t/NE(t)$ effective time interval, where the $NE(t)$ is an average harmonic effective population size over t generations. Since here systematic pressures can be neglected, as a result, surname microevolution is approximated by the process of random gene drift, synchronously proceeding in the same population with four times less intensity than for surnames. Like the genetic drift model, the surname composition of the next generation is a random sample with replacement of the size $Ne(\tau)/2$ from the surnames of the male component of the parent population, that is, four times smaller than the sample of $2Ne(\tau)$ gametes under the genetic drift ($Ne(\tau)$ effective population size in generation τ). Dynamics of probability of random encounter of individual with namesake and the behavior of encounter probability for individuals with different surnames are studied. These probabilities are similar to homozygotes and heterozygotes, respectively, in the analysis of genetic structure. Exact time dependencies for those probabilities, for variances of concentrations of surnames and the surname analogue of inbreeding coefficient are given. An approximation of the exact dependencies by simpler ones is given over a small effective time $t/NE(t)$, where the surname divergence is four times faster than the genetic divergence. The results do not suggest surname monophilia and describe a speculative theoretical set of replica populations, as if having repeated the microevolutionary history of the population in question under the same conditions. The use of a small time compared to the size of the population is justified by the recent emergence of most surnames in Russia and the fact that the past time in generations is much smaller than the typical size of populations. In real subdivided populations, procedures for estimating the inbreeding coefficient using surname concentrations do not allow us to distinguish between situations of a mechanical mixture of subpopulations or their common origin.

Keywords: theoretical population genetics, populations with non-overlapping generations, inbreeding coefficient, random surname drift, random genetic drift, mathematical models, dynamics of surname composition divergence, asymptotics over small effective time.