УДК 004.9; 575.112

СЕГРЕГАЦИОННЫЕ МОДЕЛИ СЛОЖНЫХ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ И АНАЛИЗ СЦЕПЛЕНИЯ В РАСШИРЕННЫХ ДИАЛЛЕЛЬНЫХ СКРЕЩИВАНИЯХ ПАНЕЛИ РЕКОМБИНАНТНЫХ ИНБРЕДНЫХ ЛИНИЙ

© 2013 г. **М.С. Дьяков, А.В. Осадчук**

Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия,
e-mail: dkmike@gmail.com

Поступила в редакцию 15 августа 2013 г. Принята к публикации 5 сентября 2013 г.

Представлены классы сегрегационных моделей, описывающих характер наследования количественного признака, для расширенных нереципрокных диаллельных скрещиваний рекомбинантных инбредных линий. Отличительной особенностью данной работы являются использование многолокусного подхода и учет эпистатических взаимодействий групп локусов между собой. В построенных классах моделей производится поиск решений, которые с точностью до средовых шумов описывают экспериментальные данные. Далее выполняется анализ сцепления, т.е. определение положения модельных локусов найденных решений на генетической карте хромосом. Апробация процедуры поиска в пространстве моделей и подхода к анализу сцепления проводилась на реальных данных, где в качестве количественного признака выступала масса мозжечка лабораторных мышей. Дано краткое описание реализованного программного обеспечения.

Ключевые слова: генетический анализ, количественные признаки, многолокусный подход, эпистатические взаимодействия, анализ сцепления, регрессионный анализ, рекомбинантные инбредные линии, расширенные диаллельные скрещивания, информационные технологии анализа данных.

ВВЕДЕНИЕ

Диаллельные скрещивания представляют собой потомков первого поколения, полученных от скрещивания инбредных линий во всевозможных комбинациях. Диаллельный анализ позволяет получать достаточно точную оценку генотипических значений для каждой комбинации скрещиваний, поскольку в ней используются группы животных, одинаковые по своему генотипу. Это свойство делает его привлекательным подходом в физиологической генетике, так как физиологические признаки часто характеризуются большой средовой изменчивостью. Кроме того, в отличие от других систем генетического анализа, комплекс диаллельных скрещиваний обеспечивает однозначное установление всех генотипов диаллельной матрицы скрещиваний при условии, что известны генотипы инбредных линий. Эти два свойства диаллельных скрещиваний позволяют эффективно конструировать довольно сложные модели генетической детерминации исследуемых признаков с минимально необходимым для этих целей числом генетических локусов.

ПОСТРОЕНИЕ МОДЕЛИ

В данной работе в качестве экспериментального материала используются расширенные диаллельные скрещивания. Матрица расширенных диаллельных скрещиваний содержит: 1) рекомбинантные инбредные (РИ) линии; 2) потомков первого поколения скрещиваний панели РИ линий во всевозможных комбинациях; 3) линии-основатели панели РИ линий; 4) кроссы между

РИ линиями и их линиями-основателями; 5) гибриды первого поколения линий-основателей.

Для анализа характера наследования рассматриваемого количественного признака в расширенных диаллельных скрещиваниях используется метод построения статистических генетических моделей на основе множественного регрессионного анализа диаллельных матриц. Аутосомный локус рассматривается как фактор, имеющий 3 градации (уровня): 2 градации для гомозигот и 1 градацию для гетерозиготы. Генотипическое значение анализируемого признака в модели представлено как результат суммарного влияния некоторого числа вышеуказанных факторов, который описывается линейным регрессионным уравнением. Это уравнение является линейной комбинацией генетических эффектов данных локусов или факторов. Каждый аутосомный локус имеет два главных эффекта: аддитивный и доминантный. При взаимодействии локусов между собой в регрессионное уравнение вводятся соответствующие эффекты. При взаимодействии двух аутосомных локусов имеется 4 вида эффектов: гомо-гомозиготные, гомо-гетерозиготные, гетеро-гомозиготные, гетеро-гетерозиготные. Такого рода уравнения, выражающие генотипическое значение анализируемого признака как линейную регрессионную функцию от генотипа, впервые были введены ван дер Вином (Van der Veen, 1959) и описаны в классической монографии по биометрической генетике К. Мазера и Дж. Джинкса (1985). Эти уравнения использовались ими главным образом для описания компонент фенотипической изменчивости. В нашей работе эти уравнения используются для адекватного описания генотипической изменчивости в диаллельных скрещиваниях на основе минимального числа генетических локусов.

В настоящей работе используется множественная регрессионная модель вида:

$$X_{mf} = \mu + E_1 + E_2 + ... + E_L + \varepsilon, \qquad (1)$$

где X_{mf} – генотипическое значение признака для *mf*-го кросса диаллельной матрицы (*m* обозначает номер отцовской рекомбинантной линии, *f* – материнской); μ – свободный член уравнения, который оказывается равным среднему значению признака по всевозможным кроссам; E_1 – вклад главных (аддитивных и доминантных) генетических эффектов; E_i – вклад эпистатических эффектов взаимодействия всевозможных групп локусов размера *i*.

Генотип каждой рекомбинантной линии представляется в виде вектора $\theta_i = (\theta_i^1, ..., \theta_i^L)$, $i = 1...S$, S – число рекомбинантных линий, L – количество локусов.

Каждая из компонент вектора может принимать значения 0 или 1, т. е. каждый локус представлен двумя аллелями. Для определенности полагаем, что значение 0 будет указывать на аллель, соответствующий гену, унаследованному от материнской гомозиготы, и наоборот значение 1 будет указывать на аллель, соответствующий гену отцовской линии. Таким образом, значения, соответствующие генотипам материнской и отцовской линий, будут выражаться как $\theta_{S+1} = (0, 0, ..., 0)$ и $\theta_{S+1} = (1, 1, ..., 1)$ соответственно.

Из имеющихся значений рассматриваемого множества из $(S + 2)^2$ кроссов, полученных от нереципрокного скрещивания рекомбинантных линий друг с другом и с линиями-основателями во всевозможных сочетаниях с добавлением генотипов линий основателей и их нереципрокного гибрида F1, составляется система линейных уравнений (всего не более $(S + 2)^2$ уравнений). Фиксированием значений генотипов для рекомбинантных линий однозначно определяются значения индикаторных переменных в уравнении (1).

Каждая индикаторная переменная в линейном уравнении (1) умножена на коэффициент, равный соответствующим аддитивному, доминантному или различного рода эпистатическим генетическим эффектам. Кроме того, каждая индикаторная переменная является целочисленной функцией от значений генотипа рекомбинантных линий и может принимать значения: 1, 0 или –1. $A(i)^1_{mf} = D(i)_{mf} = \theta^i_m + \theta^i_f - 1$, $A(i)^2_{mf} = H(i)_{mf} = (\theta^i_m - \theta^i_f)^2$ – индикаторные переменные перед коэффициентами, равными аддитивным и доминантным эффектам соответственно. $B(i,j)^{rs}_{mf} = A(i)^r_{mf} \cdot A(j)^s_{mf}$ – индикаторные переменные перед коэффициентами, равными вышеуказанным эффектам взаимодействия между парами локусов, входящими в *E*2: $B(i,j)^{11}_{mf}$ – перед гомо-гомозиготными эффектами; $B(i,j)^{12}_{mf}$ – перед гомо-гетерозиготными эффектами; $B(i,j)^{21}_{mf}$ – перед гетеро-гомозиготными эффектами; $B(i,j)^{22}_{mf}$ – перед гетеро-гетерозиготными эффектами.

Таким образом, вклад главных генетических эффектов и эпистатических эффектов взаимодействия локусов может быть выражен следующим образом:

$$E_1 = \sum_{r=1}^{2} \sum_{i=1}^{L} [A(i)_{mf}^{r} \cdot a(i)^{r}],$$

$$E_2 = \sum_{r=1}^{2} \sum_{s=1}^{2} \sum_{i=1}^{L} \sum_{j=i+1}^{L} [B(i,j)_{mf}^{rs} \cdot b(i,j)^{rs}],$$

где L – число локусов. Значения вклада генетических эффектов E_i, $2 < i \leq L$ вычисляются аналогично.

Методом множественной линейной взвешенной регрессии (Кобзарь, 2006) определяются такие значения генетических эффектов, которые минимизировали бы отклонения от полученных в эксперименте значений фенотипов, т. е. минимизировали бы ошибку. Для этого решается система линейных уравнений и рассчитывается коэффициент множественной детерминации R^2. В качестве весов w_{mf} берется количество особей, использованное для определения генотипических средних значений признака X_{mf}. Качество полученного решения проверяется с использованием критерия Фишера сравнением остаточной дисперсии, не учтенной множественной регрессией, со средовой вариансой. Если оценка адекватности регрессии больше некоторого уровня, считаем, что полученное решение описывает экспериментальный материал с точностью до средовой вариансы – отклонений, обусловленных случайными средовыми факторами. При этом коэффициент множественный детерминации R^2 указывает на долю межкроссной наследственно обусловленной изменчивости, объясняемую с помощью множественной регрессионной модели. Если выбранное решение не подходит, выбираются другие значения генотипов у рекомбинантных линий и вычислительная процедура повторяется. Если окажется, что ни один из вариантов не описывает адекватно экспериментальные данные, то необходимо выбрать более сложную модель (большее число взаимодействующих локусов L).

Таким образом, при выборе оптимального решения мы приходим к перебору всех возможных вариантов генотипов рекомбинантных линий. Исходя из того, что L векторов $\theta^i = (\theta^i_1, \theta^i_2, ..., \theta^i_S)$, $i = 1 ... L$ должны быть различны и из того что от перестановки номеров локусов решение не изменяется, общий размер пространства решений вычисляется по следующей формуле:

$$N = \frac{2^S \cdot (2^S - 1) \cdot \ldots \cdot (2^S - L + 1)}{L!},$$

где S – число рекомбинантных линий, L – число локусов. Часто на практике в генетических экспериментах это число является очень большим, например, нами была исследована задача с $N \approx 1{,}88 \cdot 10^{14}$.

ПЕРЕБОР ПРОСТРАНСТВА РЕШЕНИЙ

Поскольку при данной постановке задачи не существует методов, которые бы гарантированно находили все адекватные решения из пространства моделей вида (1) быстрее, чем полный перебор вариантов, было применено улучшение алгоритма перебора, которое позволяет гораздо раньше находить адекватные решения. Данный алгоритм лучше всего можно описать как «рандомизированный поиск в ширину с приоритетом». Он основан на одновременном использовании двух методов: метода случайного поиска решений и метода поиска в ширину с приоритетом. Приоритеты определяются посредством ранжирования решений-кандидатов по качеству.

Метод поиска в ширину с приоритетом. Данный метод основывается на том принципе, что соседи более хорошего решения должны перебираться раньше соседей более плохого решения. Метод можно описать следующим образом.

Имеется контейнер («контейнер непросчитанных решений»), представляющий собой очередь с приоритетом, в нем содержатся непросчитанные решения. В качестве приоритета выступает коэффициент множественной детерминации. Каждый раз из контейнера извлекается наилучшее решение и просчитываются значения критерия для его соседей. Соседи, согласно их приоритетам, также кладутся в контейнер. Контейнер ограничен, и поэтому решения, занимающие положение ниже определенного («объем контейнера»), вытесняются. Рассмотренные решения убираются из контейнера непросчитанных решений поиска в ширину и помечаются как просчитанные.

Так как одно решение является соседом нескольких других, то для того чтобы избежать

повторного попадания решения в «контейнер непросчитанных решений», решение помечается как просчитанное после просмотра его соседей и добавляется в «контейнер просчитанных решений». Перед добавлением нового решения в «контейнер непросчитанных решений» проверяется, не просчитывалось ли данное решение ранее. В случае если решение рассматривалось, то оно игнорируется, иначе попадает в «контейнер непросчитанных решений».

Если в результате расчета критерия решение описывает экспериментальный материал с точностью до средовых шумов, то оно добавляется в результирующий набор. Первоначально контейнер заполняется либо указанными исследователем решениями, либо случайным образом.

Метод случайного поиска решений. Для вывода алгоритма из локальных областей необходимо введение случайных решений. В случае наличия большого числа хороших решений у системы поиска в ширину с приоритетом есть недостаток – при добавлении случайного решения в «контейнер непросчитанных решений» поиска очень вероятно, что это решение окажется в самом низу и будет рассмотрено в последнюю очередь или даже не будет рассмотрено вообще. Для решения этой проблемы введен «контейнер индивидуального поиска», в котором воспроизводится процесс поиска в ширину с приоритетом для решений, полученных из одного случайного решения. Все содержимое контейнера через несколько итераций добавляется в «контейнер непросчитанных решений», где некоторые удачные решения, возможно, будут в дальнейшем рассмотрены. Решения из «контейнера индивидуального поиска», которые были просчитаны, перемещаются сразу в «контейнер просчитанных решений».

Особенности программной реализации. На иллюстрации (рис. 1) изображена одна итерация алгоритма поиска решений в пространстве моделей на основе данных о расширенных диаллельных скрещиваниях.

1) В управляющем потоке из «контейнера непросчитанных решений» извлекается группа решений с наибольшим приоритетом. Эти решения помечаются как просчитанные и добавляются в «контейнер просчитанных решений».

2) Из окрестности извлеченных решений формируются решения-соседи.

3) При помощи набора переиспользуемых потоков (thread pool) в модуле расчета регрессии параллельно рассчитываются критерии адекватности для сформированных решений-соседей.

4) После расчета критериев в управляющем потоке новые решения добавляются в «контейнер непросчитанных решений» и происходит выбор решений, описывающих экспериментальные данные с точностью до средовых шумов.

АНАЛИЗ СЦЕПЛЕНИЯ МОДЕЛЬНЫХ ЛОКУСОВ С МАРКЕРАМИ ХРОМОСОМ

После формулирования гипотез о характере наследования локуса по рекомбинантным линиям и поиска адекватных решений на основе этих гипотез актуальной становится задача точного определения положения локуса в исследуемом геноме. Произведенные в последние годы исследования предоставили такую возможность, обеспечив экспериментатора идеальными хромосомными микросателлитными маркерами, плотно картирующими весь геном. Микросателлитные маркеры имеют значительно более высокий уровень полиморфизма, чем ранее используемые для этой цели мутантные аллели генов и полиморфизм по генам ферментов, и их применение позволяет с большой степенью достоверности найти хромосомную локализацию модельного локуса. Характеристика степени близости модельного локуса к определенному месту хромосомы носит название сцепленности и определяется через сходство характеров распределения аллелей модельного и фланкирующих его микросателлитных локусов, а также расстояния между маркерными микросателлитными локусами на генетической карте рекомбинантных линий.

Таким образом, построенная сегрегационная модель будет не только адекватно описывать межкроссную генотипическую изменчивость, но ее модельные локусы будут сцеплены с некоторыми картированными микросателлитными маркерами. Это позволит произвести отсев несцепленных адекватных решений-кандидатов.

Трехлокусное сцепление. Одномерный случай трехлокусного сцепления был описан П. Ньюманом (Neumann, 1991). Рассмотрим тестовый локус *C* и пару сцепленных маркерных

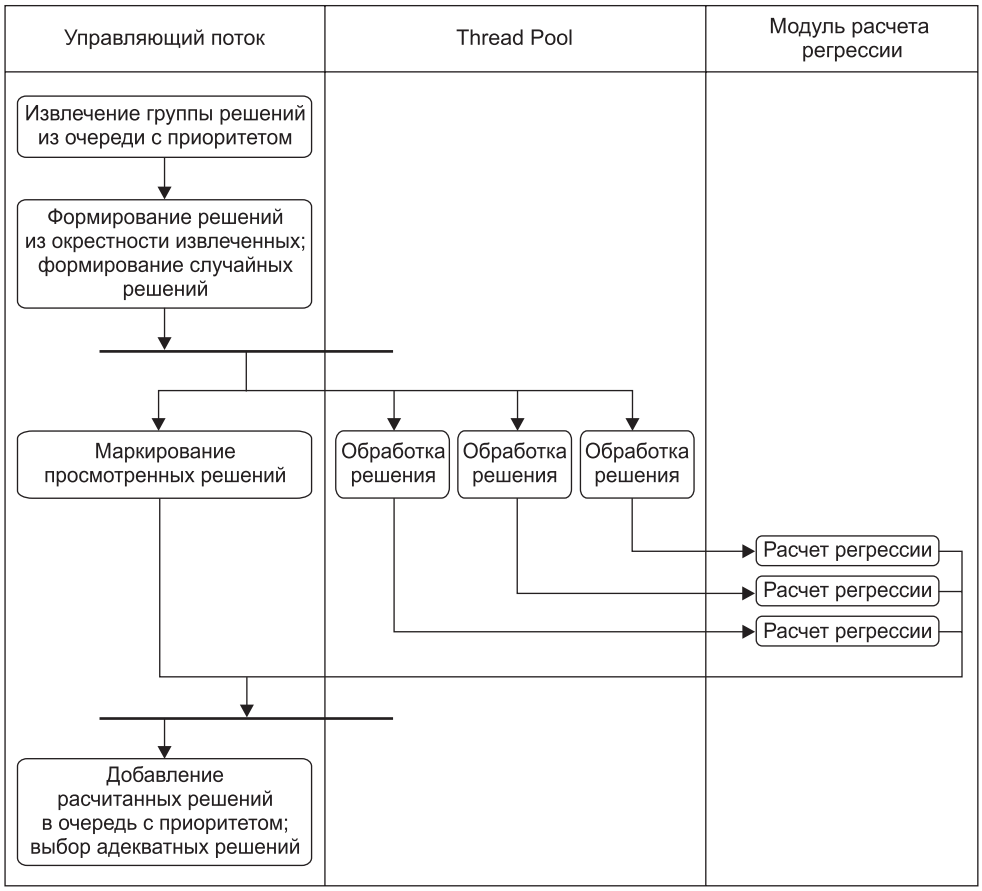

Рис. 1. Диаграмма деятельности (activity diagram). Представлена одна итерация алгоритма поиска решений в пространстве моделей.

локусов A и B (рис. 2). Количество различий (рекомбинаций) в генотипах локусов A и C в S рекомбинантных линиях обозначим как a. Аналогично определим b как количество рекомбинаций между B и C и c как количество различий в генотипах между локусами A и B. Количество двойных рекомбинаций d (число линий, которые имеют различия между локусами A и C и B и C одновременно) можно определить как $d = (a + b - c)/2$. n – разность $(S - c)$.

Вероятность сцепления $P(L)$ представляет собой сумму вероятностей возникновения каждого из трех альтернативных порядков расположения генов: $P(L) = P(CAB) + P(ACB) + P(ABC)$, где C – тестовый локус, A и B – сцепленные маркерные локусы. Таким образом, вероятность (на основе данных о распределении аллелей) того, что локус C сцеплен с A и B, может быть получена из байесовских выражений (2):

$$P(L|a, b) = \frac{[P(CAB)P(a, b|CAB) + P(ACB)P(a, b|ACB) + P(ABC)P(a, b|ABC)]}{P(a, b)} \quad (2)$$

$$P(a, b) = P(CAB)P(a, b|CAB) + P(ACB)P(a, b|ACB) + P(ABC)P(a, b|ABC) + P(\overline{L})P(a, b|\overline{L}).$$

Хромосома, содержащая сцепленные маркерные локусы A и B, имеет длину l (выраженную в морганидах) и состоит из трех частей. Длину сегмента между локусами A и B обозначим как m_{AB}, расстояние от начала хромосомы до локуса A примем за m_A, а расстояние от локуса B до конца хромосомы – за m_B (рис. 2). Таким образом, $l = m_A + m_{AB} + m_B$.

Априорная вероятность того, что тестовый локус C сцеплен с сегментом AB и расположен слева от маркерного локуса A, пропорциональна m_A и равна $P(CAB) = m_A/T$, где

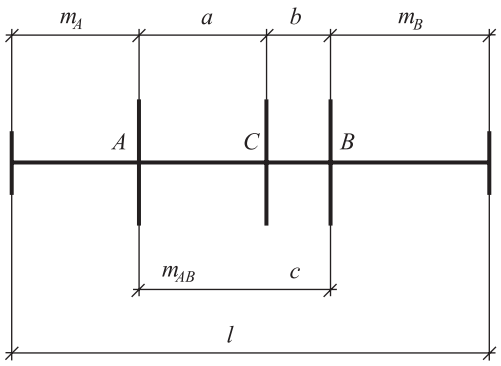

Рис. 2. Схематическое изображение хромосомы.

A и *B* – маркерные локусы, *C* – тестовый локус; *a*, *b* и *c* – количество соответствующих рекомбинаций; m_A, m_B, m_{AB} и *l* – расстояния, выраженные в морганидах.

T – длина всего генома (например, для мыши *T* = 16 морганид). Аналогично определяются априорные вероятности того, что тестовый локус расположен между маркерными локусами или справа от локуса *B*: $P(ACB) = m_{AB}/T$, $P(ABC) = m_B/T$.

Априорная вероятность сцепления тестового локуса с маркерными равна $P(L) = l/T$. Соответственно, вероятность того, что тестовый локус не сцеплен с маркерными, составляет $P(\overline{L}) = 1 - P(L) = (T - l)/T$.

Вероятность того, что тестовый локус *C* имеет *a* и *b* рекомбинаций со сцепленными маркерами *A* и *B* соответственно в наборе из *S* рекомбинантных инбредных линий и не сцеплен с ними, равна (полиномиальное распределение):

$$P(a, b|\overline{L}) = \frac{C \cdot R^c(m_{AB}) \cdot (1 - R(m_{AB}))^n}{2^S},$$

где $C = \dfrac{S!}{[d!(n-d)!(a-d)!(b-d)!]}$,

$R(x) = \dfrac{4 \cdot r(x)}{[1 + 6 \cdot r(x)]}$ – ожидаемая доля различий в распределении аллелей между двумя локусами, которые находятся на расстоянии *x* морганид, в панели РИ линий (Haldane, Waddington, 1931). $r(x) = \dfrac{0{,}5 \cdot (e^{2x} - e^{-2x})}{(e^{2x} + e^{-2x})}$ – картирующая функция Косамби (Kosambi, 1943). $R(m_{AB})$ можно считать равным рекомбинантному соотношению *c*/*S*.

Остальные три условные вероятности определяются как

$$P(a, b|CAB) = \left(\frac{C}{m_A}\right) K(CAB),$$

$$P(a, b|ACB) = \left(\frac{C}{m_{AB}}\right) K(ACB),$$

$$P(a, b|ABC) = \left(\frac{C}{m_B}\right) K(ABC), \text{ где}$$

$$K(CAB) = \int_0^{m_A} R^c(m_{AB}) \cdot \left(1 - R(m_{AB})\right)^n \cdot R^a(x) \cdot \left(1 - R(x)\right)^{S-a} dx,$$

$$K(ACB) = \int_0^{m_{AB}} R^b(m_{AB} - x) \cdot \left(1 - R(m_{AB} - x)\right)^{S-b} \cdot R^a(x) \cdot \left(1 - R(x)\right)^{S-a} dx,$$

$$K(ABC) = \int_0^{m_B} R^c(m_{AB}) \cdot \left(1 - R(m_{AB})\right)^n \cdot R^b(x) \cdot \left(1 - R(x)\right)^{S-b} dx.$$

В многомерном случае при определении вероятности сцепления группы локусов можно считать, что сцепление по каждому локусу в отдельности происходит независимо. Поэтому вероятность в многомерном случае выражается следующим образом: $P(L^k|a^k, b^k) = \prod_{i=1}^{k} P(L_i|a_i, b_i)$, где *k* – размерность задачи (количество локусов), а $P(L_i \mid a_i, b_i)$ вычисляются для одномерного случая описанным выше способом.

Поиск всех точек сцепления. Для каждого адекватного решения, найденного на первом этапе, определяются все возможные точки сцепления на генетической карте хромосом. Используемый метод основывается на поиске с возвратом (backtracking).

Допустим, что для первых *i* из *k* локусов уже установлен возможный вариант сцепления. Для *i* + 1 локуса подбирается возможное положение, и алгоритм рекурсивно повторяется для *i* + 2 локуса и т. д. Данная процедура продолжается до тех пор, пока не найдено возможное положение для всех *k* локусов. Если итоговая *k*-мерная точка удовлетворяет многомерному критерию сцепленности, то она добавляется в результирующий список.

На каждом этапе алгоритма для ускорения поиска применяются два отсечения с использованием оценки снизу на суммарное количество

двойных рекомбинаций и оценки сверху на многомерную вероятность сцепления. Первое отсечение основывается на том, что необходимый уровень многомерного сцепления не может быть достигнут при суммарном количестве двойных рекомбинаций больше определенного. Оценка сверху на вероятность многомерного сцепления также позволяет не рассматривать поддеревья поиска, в листьях которых пороговое значение критерия сцепленности заведомо не может достигаться.

Так как вероятность одномерного сцепления при фиксированном количестве рекомбинантных инбредных линий зависит только от трех параметров (количества соответствующих рекомбинаций a, b и c, описанных выше) и не зависит от внутренней структуры локусов или микросателлитных маркеров, полученное соотношение (2) можно вычислять только один раз для каждой комбинации параметров. Далее эти значения можно хранить в виде таблицы и при необходимости получать их мгновенно.

Применение хеширования. Метод поиска всех точек сцепления (без учета отсечений), описанный выше, обладает временно́й сложностью $O(M^L)$, где M – количество маркеров на генетической карте хромосом, а L – количество локусов. Таким образом, сложность алгоритма растет экспоненциально в зависимости от числа локусов. Однако, если исследователем выбран уровень многомерного сцепления, при котором суммарное количество двойных рекомбинаций не может превышать 0 (т. е. равняется 0), то перед исполнением алгоритма поиска всех точек сцепления можно ввести дополнительную проверку, которая позволяет избавиться от лишних расчетов при отсутствии сцепления модельных локусов с маркерами на генетической карте хромосом.

Рассмотрим пару фланкирующих маркеров, между которыми может потенциально располагаться модельный локус. При условии, что количество двойных рекомбинаций в этой тройке локусов равняется нулю, можно по паре фланкирующих маркеров сформировать все возможные модельные локусы, которые удовлетворяют данному ограничению. Количество таких модельных локусов равняется 2^c, где c – число рекомбинаций между фланкирующими маркерами. Время формирования этих локусов пропорционально их количеству.

Далее для определения возможности сцепления конкретного модельного локуса с фланкирующими маркерами необходимо проверить, находится ли этот модельный локус среди сформированной группы локусов для данных фланкирующих маркеров. Для того чтобы делать это эффективно, для каждого локуса из группы рассчитывается полиномиальный хеш (3) и полученные значения добавляются в хеш-таблицу. Для модельного локуса также вычисляется значение хеш-функции (3) и проверяется, находится или нет полученное значение в хеш-таблице. Если значение присутствует, то существует возможность того, что модельный локус может быть сцеплен с парой фланкирующих маркеров. Если полученное значение хеш-функции отсутствует в таблице, значит количество двойных рекомбинаций в данной тройке больше нуля, следовательно, необходимый уровень многомерного сцепления не может быть достигнут. Таким образом, данный модельный локус может быть сразу исключен из рассмотрения.

Использованная нами полиномиальная хеш-функция выглядит следующим образом:

$$h(\theta) = \sum_{i=1}^{S} (\theta_i \cdot 2^{i-1}) \bmod p, \qquad (3)$$

где θ – распределение аллелей, S – число рекомбинантных линий, p – размер хеш-таблицы.

Для реализации данного отсечения необходимо один раз выполнить предрасчет для всех пар фланкирующих маркеров. Асимптотическая оценка сложности предрасчета равняется $O(M \cdot 2^{c_{max}})$, где M – количество маркеров на генетической карте хромосом, а c_{max} – максимальное число рекомбинаций для пар фланкирующих маркеров. Дополнительная проверка для модельных локусов перед поиском всех точек сцепления при этом будет занимать $O(L)$ времени, где L – число локусов.

На практике эта оптимизация в применении к данным, описанным в разделе «Апробация программного пакета», позволила увеличить производительность в 30 раз. Следует также отметить, что данный подход можно обобщить на случай, если выбранный уровень многомерного сцепления может быть достигнут и при суммарном количестве двойных рекомбинаций больше нуля. Однако это повлечет увеличение времени предрасчета и уменьшение эффек-

тивности отсечения, так как при хешировании возрастет количество коллизий.

АПРОБАЦИЯ ПРОГРАММНОГО ПАКЕТА

Тестирование системы проводилось на реальных данных, где в качестве сложного количественного признака выступала масса мозжечка у мышей. Была использована матрица расширенных диаллельных скрещиваний РИ линий панели СХВ. Количество рекомбинантных инбредных линий S равнялось 13. При использовании одно-, двух- и трехлокусных моделей не было выявлено ни одного решения, адекватно описывающего экспериментальные данные с точностью до средовых шумов (рис. 3). Причем для одно- и двухлокусной модели было просмотрено все пространство допустимых решений. При количестве взаимодействующих локусов L, равном 4, было выявлено приблизительно 1,2 млн адекватных решений, из них сцепленными оказались только 22, учитывая, что все пространство решений содержит $N \approx 1{,}88 \cdot 10^{14}$ кандидатов.

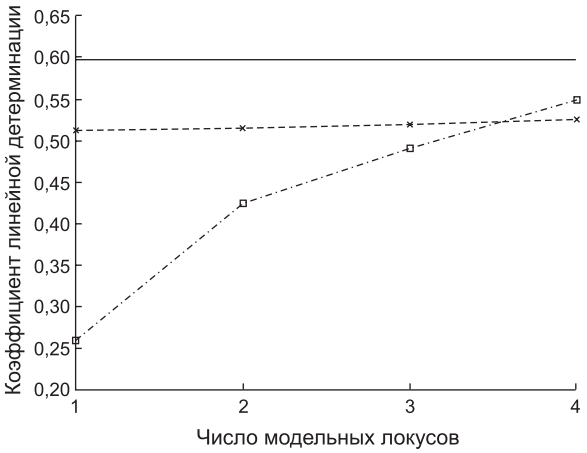

Рис. 3. Описательная способность различных многолокусных моделей.

Сплошной линией отмечен максимально возможный теоретический уровень описания (при помощи регрессионного анализа) экспериментальных данных в расширенных диаллельных скрещиваниях РИ линий СХВ. Он обусловлен различиями в фенотипе особей одного изогенного кросса, т. е. влиянием среды. Штриховой линией отмечены минимально необходимые уровни коэффициента линейной детерминации R^2 для каждой модели при описании данных на уровне $\alpha = 0{,}05$. Штрихпунктирной линией отмечены максимально достижимые коэффициенты R^2 для соответствующих моделей.

Следует отметить, что бóльшая часть найденных решений, описывающих экспериментальные данные с точностью до средовых шумов, выявляется разработанным алгоритмом на начальных итерациях (рис. 4). За динамикой поиска удобно проследить, используя логарифмическую шкалу (рис. 5).

Размещение программного пакета предполагается на современных персональных компьютерах и ноутбуках, поэтому система разрабатывалась с учетом многоядерности процессорных

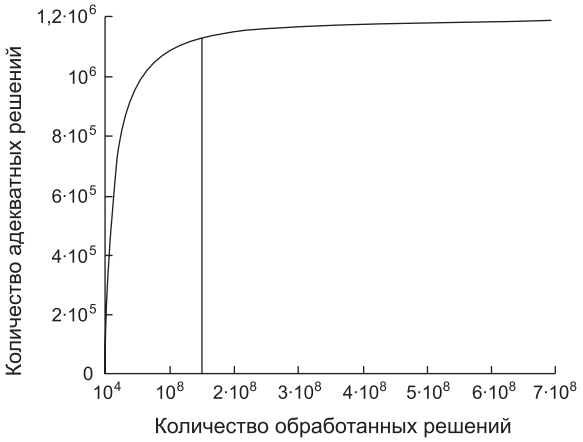

Рис. 4. Количество адекватных решений в зависимости от количества обработанных для поиска решений в пространстве моделей.

Вертикальным отрезком обозначены 95 % всех найденных адекватных решений.

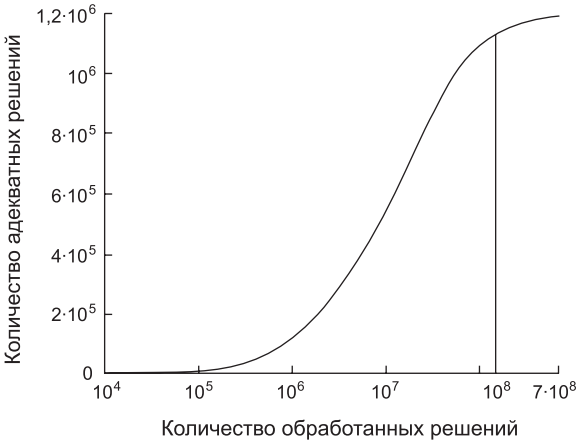

Рис. 5. Количество адекватных решений в зависимости от количества обработанных для поиска решений в пространстве моделей.

Ось абсцисс изображена в логарифмической шкале по основанию 10. Вертикальным отрезком обозначены 95 % всех найденных адекватных решений.

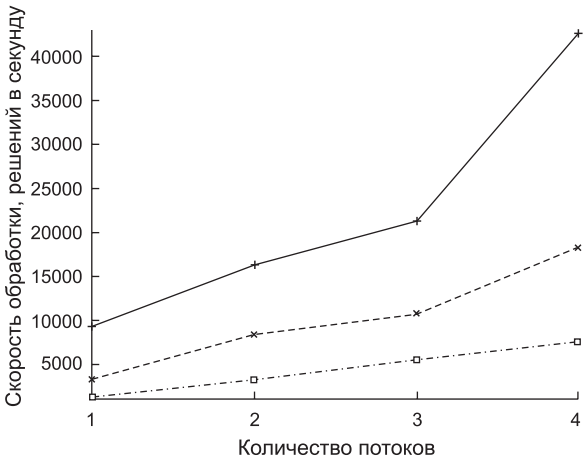

Рис. 6. Зависимость скорости счета от количества используемых потоков для разных моделей.

Сплошной линией обозначена двухлокусная модель, штриховой – трехлокусная, штрихпунктирной – четырехлокусная.

архитектур. За счет распараллеливания вычислений удалось добиться линейного повышения производительности в зависимости от количества физических ядер процессора. На графике (рис. 6) представлены данные о скорости счета с использованием компьютера на базе четырехъядерного процессора Intel Core i7 в зависимости от количества используемых потоков.

ЛИТЕРАТУРА

Кобзарь А.И. Прикладная математическая статистика. М.: ФИЗМАТЛИТ, 2006. 816 с.

Мазер К., Джинкс Дж. Биометрическая генетика. М.: Мир, 1985. 464 с.

Haldane J.B.S., Waddington C.H. Inbreeding and linkage // Genetics. 1931. V. 16. No. 4. P. 357–374.

Kosambi D.D. The estimation of map distance from recombination values // Annual. Eugenics. 1943. V. 12. No. 1. P. 172–175.

Neumann P.E. Three-locus linkage analysis using recombinant inbred strains and bayes' theorem // Genetics. 1991. V. 128. No. 3. P. 631–638.

Van der Veen J.H. Test of non-allelic interaction and linkage for quantitative characters in generations derived from two diploid pure lines // Genetics. 1959. V. 30. No. 3. P. 201–232.

SEGREGATION MODELS OF COMPLEX QUANTITATIVE TRAITS AND LINKAGE ANALYSIS IN EXTENDED RECOMBINANT INBRED CROSSES

M.S. Diakov, A.V. Osadchuk

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, e-mail: dkmike@gmail.com

Summary

This paper introduces classes of segregation models, which describe the inheritance of a quantitative trait, in extended diallelic recombinant inbred crosses. The distinctive feature of the method is usage of the multiple-QTL approach and incorporation of epistatic interactions of loci groups. Solutions that would describe experimental data set to an accuracy of environmental variance are sought in the constructed classes of models. Then linkage analysis is performed: model loci positions of the found solutions are mapped on chromosomes. The search procedure and linkage analysis have been tested with real data on cerebellum weight in laboratory mice as a quantitative trait. The developed software is briefly described.

Key words: segregation analysis, complex quantitative traits, multiple-QTL approach, linkage analysis, regression analysis, epistatic interactions, recombinant inbred strains, extended diallelic crosses, information technologies of data analysis.