

РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ

На правах рукописи

УДК 575.11:591.15:631.520

Аульченко Юрий Сергеевич

**Развитие методов сегрегационного анализа: комплексные
признаки и сложные родословные**

Генетика – 03.00.15

**Диссертация на соискание ученой степени
кандидата биологических наук**

Научный руководитель
д.б.н., Аксенович Т. И.

Новосибирск — 2000

Оглавление

Список принятых обозначений	5
Введение	6
1 Комплексный сегрегационный анализ: обзор литературы	11
1.1 Модели комплексного сегрегационного анализа	13
1.2 Тестирование гипотез	23
1.3 Вычислительные алгоритмы	25
1.4 Заключение	31
2 Разработка моделей наследования	33
2.1 Модели наследования признаков в гибридных родословных	33
2.2 Модели наследования совместных признаков	41
2.3 Обсуждение	44
3 Разработка алгоритмов: анализ родословных с множественными пет-	
лями	48
3.1 Алгоритм подсчета функции правдоподобия родословной без петель . .	48
3.2 Свертка коротких инбредных петель	50
3.3 Точность генетического анализа при различных аппроксимациях функции правдоподобия	52
3.4 Обсуждение	56
4 Анализ наследования гетерохромии глаз у домашней свиньи	59
4.1 Обзор литературы и постановка задачи	59
4.2 Материалы и методы	62
4.3 Результаты	66

4.4	Обсуждение	69
5	Анализ наследования числа новорожденных у мускусной землеройки (<i>Suncus murinus</i>, Insectivora, Soricidae)	71
5.1	Обзор литературы и постановка задачи	71
5.2	Материалы и методы	74
5.3	Результаты	76
5.4	Обсуждение	81
6	Анализ наследования ген-цитоплазматической мужской стерильности в родословной сахарной свеклы	84
6.1	Обзор литературы и постановка задачи	84
6.2	Материалы и методы	88
6.3	Результаты	93
6.4	Обсуждение	96
	Выводы	99
	Список литературы	101
	Приложения	112
A	Пример подсчета функции правдоподобия	112
B	Описание пакета программ МАІА-1.0	115
C	Вычисление апостериорных вероятностей генотипов по гену-ингибитору масти	118
D	Формализация ген-цитоплазматического наследования	120

Список иллюстраций

1.1	Пороговая модификация: функция пенетрантности количественного признака и кумулятивная функция пенетрантности	21
1.2	Родословная, изображенная в виде ориентированного графа	26
1.3	Различные способы разрыва петель	29
2.1	Формирование совместного признака	42
3.1	Ядерная родословная	50
3.2	Пример части родословной	51
3.3	Частота ложного отвержения менделевского наследования	56
4.1	Полная и частичная гетерохромия радужной оболочки глаза	60
4.2	Родословная свиней, полученная в экспериментальном хозяйстве ИЦиГ СО РАН	63
5.1	Фрагмент гибридной родословной SK	74
6.1	Родословная семейства инбредных линий СОАН-104	89
6.2	Обобщенная модель контроля ген-ЦМС	91
A.1	Родословная с моногенной болезнью	113
B.1	Структура пакета программ MAIA-1.0	116

Список таблиц

3.1	Средние значения оценок параметров и стандартное отклонение среднего при различных способах подсчета функции правдоподобия	55
4.1	Наследование гетерохромии радужки у белых миниатюрных свиней . .	61
4.2	Результаты сегрегационного анализа гетерохромии. Проверка майор-генной гипотезы	67
4.3	Результаты сегрегационного анализа гетерохромии. Варианты майор-генной гипотезы	68
5.1	Средний размер помета в различных скрещиваниях	76
5.2	Результаты сегрегационного анализа размера помета. Проверка майор-генной гипотезы	79
5.3	Результаты сегрегационного анализа размера помета. Варианты майор-генной гипотезы	80
6.1	Результаты сегрегационного анализа признака "переход нормальной цитоплазмы в стерильную"	93
6.2	Пенетрантности генотипов по генам-восстановителям фертильности при различных значениях параметра ω_{FR}	94

Список принятых обозначений

$|\mathcal{A}|$ — мощность множества \mathcal{A} . В случае дискретного множества равна числу его элементов

$Pr(\mathcal{B})$ — вероятность события \mathcal{B}

$Pr(\mathcal{B}|\mathcal{C})$ — вероятность события \mathcal{B} при условии \mathcal{C}

$\varphi_x(\mu, \sigma^2)$ — плотность нормального распределения в точке x . Центр распределения находится в точке μ , а дисперсия равна σ^2

$\phi_x(\mu, \sigma^2)$ — функция нормального распределения в точке x

\mathcal{S} — структура родословной

\vec{x} — вектор фенотипов членов родословной

x_i — значение признака (фенотип) i -й особи

\vec{g} — вектор генотипов членов родословной

g_i — значение генотипа i -й особи

\mathcal{N} — остаточная наследственная компонента

e — средовая модификация

$\vec{\theta}$ — вектор параметров

$p(A_i)$ — частота аллеля A_i в популяции

$\tau(A_i|g)$ — переходная вероятность — вероятность того, что особь с генотипом g передаст потомку аллель A_i

Глава 1

Комплексный сегрегационный анализ: обзор литературы

Сегрегационным анализом называется статистический метод, позволяющий ответить на вопрос, объясняется ли вариация исследуемого признака полиморфизмом одного или нескольких менделевских генов. Если да, то сколько этих генов и каковы их характеристики — число аллелей, их частоты и эффекты, тип взаимодействия и пр. Материалом для анализа служат данные о распределении признака у групп родственных особей (Elston, 1981; Jarvik, 1998).

Существует две основных группы методов сегрегационного анализа.

Методы первой группы нацелены на анализ данных, полученных в специально проведенном гибридологическом эксперименте, спланированном таким образом, чтобы максимально полно охарактеризовать модель наследования признака. Эти методы требуют, чтобы распределение генотипов скрещиваемых особей было известным. Потомства, полученные при скрещивании таких родителей, объединяются, образуя большие выборки — F_1 , F_2 и т.д. Распределение признака в этих выборках сравнивается с ожидаемым при справедливости рассматриваемой гипотезы. Если распределения значимо не отличаются, гипотеза принимается, иначе вырабатывается новая гипотеза и т.д. (Серебровский, 1970; Тихомирова, 1990; Lynch, Walsh, 1998). Основные достоинства методов первой группы — высокая разрешающая способность и относительная простота применяемой при тестировании гипотез статистической методологии. Недостатком является высокая стоимость и трудоемкость проведения самого гибридологического эксперимента.

Методы второй группы не требуют специального гибридологического эксперимента и позволяют извлекать генетическую информацию из практически любых данных о распределении признака в ряду поколений, например, из родословных человека или генеалогий, полученных при разведении животных. Однако, при таком подходе генотипы скрещивающихся особей, как правило, не известны и проведение сегрегационного анализа требует разработки сложнейшего методологического и вычислительного аппарата (Elston, 1981; Ott, 1991; Lynch, Walsh, 1998; Sham, 1998). Характерными представителями второй группы являются методы тестирования эффектов отдельных генов по данным о популяционном распределении признака (Tan, Chung, 1972; Elston et al., 1974; Hoeschele, 1988), или распределении признака у фиксированных групп родственников (Fain 1978; Karlin et al., 1979; Sokal, Rohlf, 1995). Из второй группы методов наиболее мощным является комплексный сегрегационный анализ (КСА, Lynch, Walsh, 1998; Jarvik, 1998; Sham, 1998), который позволяет максимально полно использовать имеющуюся информацию и проводить анализ по родословным любой структуры.

Статистической основой комплексного сегрегационного анализа является метод максимума правдоподобия (Кендалл, Стюарт, 1973). Он заключается в следующем: формулируется генетико-математическая модель, на базе которой записывается аналитическое выражение для вероятности эмпирических данных (натуральный логарифм которой называется функцией правдоподобия). Это выражение является функцией от набора параметров, описывающих выбранную модель. Затем находятся такие значения параметров, которые максимизируют функцию правдоподобия, т.е. наилучшим образом описывают имеющиеся данные. В результате этого получается набор оценок параметров и максимальное значение функции правдоподобия. Эти характеристики позволяют исследователю тестировать различные генетические гипотезы, получая информацию о генетическом контроле анализируемого признака (Elston, Stewart, 1971; Elston, 1981; Ott, 1991; Lynch, Walsh, 1998; Sham, 1998).

Таким образом, метод комплексного сегрегационного анализа требует разработки моделей, позволяющих определить функцию правдоподобия, и создания специальных генетико-статистических критериев, отвечающих на вопросы, поставленные в ходе генетического анализа.

1.1 Модели комплексного сегрегационного анализа

Теоретические основы комплексного сегрегационного анализа были описаны в серии работ Эльстона с сотр. (Elston, Stewart, 1971; Elston, Yelverton, 1975; Lange, Elston, 1975; Elston et al., 1975; Elston et al., 1978; Elston, 1981). Ими была разработана генетико-математическая модель, применяемая в настоящее время для сегрегационного анализа признаков по выборкам родословных. Формализация этой модели основывается на определении трех вероятностных правил (распределений). Два из них описывают генетическую основу полиморфизма признака в популяции: это распределение, описывающее возможный полиморфизм генов, и правило построения фенотипа на основе генотипа. Третье определяет правило, согласно которому полиморфизм признака передается из поколения в поколение (Elston, 1981).

Конкретный вид этой модели зависит от типа исследуемого признака. Однако даже при описании одних и тех же моделей разные исследователи часто используют разную терминологию. Для того, чтобы описать существующие модели наследования и показать связь между ними, необходимо ввести общую терминологию и единый способ формализации.

Целью комплексного сегрегационного анализа является проверка возможности участия в контроле признака некоторого гипотетического менделевского гена (генов) g . Традиционно, этот ген называют "майоргеном" (Elston, Stewart, 1971; Elston, 1981). Наследование и полиморфизм признака может быть обусловлен не только передачей различных аллелей майоргена, но также рядом других "остаточных" наследственных причин \mathcal{N} , например, другими ядерными или цитоплазматическими генами (Morton, MacLean, 1974; Bonney, 1984; Ginsburg, Livshits, 1999). Кроме того, считается, что полиморфизм признака обусловлен множественными случайными средовыми модификациями, e . Как правило, предполагается, что они действуют на признак не направленно, независимо и аддитивно¹. В рамках этой модели значение признака x равно сумме эффектов генетических и средовых факторов:

$$x = \mu(g, \mathcal{N}) + e. \quad (1.1)$$

¹Существует ряд других факторов, которые могут детерминировать полиморфизм признака в популяции — общесемейная среда, поведенческие корреляции и пр. Кроме того, возможно взаимодействие генотипа и среды. В этой работе мы не рассматриваем такие ситуации. Тем не менее, они не требуют принципиального изменения описываемой нами модели.

Материалом для комплексного сегрегационного анализа служит набор фенотипов всех членов родословной, представленный в виде вектора² \vec{x} . Кроме того, для построения функции правдоподобия необходима информация о структуре родословной, \mathcal{S} , которая определяет направление передачи генов³. В множестве особей O , составляющих родословную, выделяется два подмножества — множество родоначальников P и множество особей, которые в родословной являются потомками, F . Принципиальная разница между двумя этими множествами заключается в том, что вероятности генотипов родоначальников определены априорным популяционным распределением генотипов, в то время как генотипы потомков зависят от генотипов родоначальников.

Наблюдаемый вектор фенотипов членов родословной — \vec{x} — детерминирован ненаблюдаемыми векторами майоргенотипов \vec{g} , остаточных наследственных факторов $\vec{\mathcal{N}}$ и вектором случайных средовых воздействий \vec{e} . Модель комплексного сегрегационного анализа задает распределение этих ненаблюдаемых характеристик у всех членов родословной и соответствие этих характеристик вектору фенотипов \vec{x} . Модель определена вектором генетических параметров, $\vec{\theta}$, в который могут входить число генов, участвующих в контроле признака, число и частоты аллелей, эффекты генотипов и т.д.

Для формализации модели достаточно определить три распределения вероятности:

1. *Априорное совместное распределение генотипов в популяции* $Pr_{\mathcal{S}, \vec{\theta}}(\vec{g}_P, \vec{\mathcal{N}}_P)$.

Как правило, предполагается, что генотипы родоначальников независимы. Тогда

$$Pr_{\mathcal{S}, \vec{\theta}}(\vec{g}_P, \vec{\mathcal{N}}_P) = \prod_{j \in P} Pr_{\vec{\theta}}(g_j, \mathcal{N}_j),$$

где $Pr_{\vec{\theta}}(g_j, \mathcal{N}_j)$ — одинаковое для всех родоначальников популяционное распределение генотипов.

²Вектор фенотипов определяется следующим образом: все члены родословной нумеруются таким образом, чтобы каждой особи соответствовал свой уникальный номер от 1 до N , где N — число особей в родословной. Фенотип i -й особи определяет i -ю координату вектора \vec{x} .

³Структура родословной определена матрицей коинцидентности ориентированного графа, в котором вершины соответствуют особям, а дуги — родственным связям между ними. Подробная формализация изложена в разделе 1.3

Для майоргенной компоненты параметрами, описывающими такое распределение, являются популяционные частоты аллелей майоргена⁴.

2. *Совместное распределение генотипов потомков при условии генотипов родоначальников* $Pr_{\mathcal{S}, \vec{\theta}}(\vec{g}_F, \vec{\mathcal{N}}_F | \vec{g}_P, \vec{\mathcal{N}}_P)$. Считается, что вероятность того, что некоторая особь наследует генный набор (g, \mathcal{N}) , зависит только от наследственных наборов ее родителей: отца — (g_m, \mathcal{N}_m) и матери — (g_f, \mathcal{N}_f) . Поэтому совместная вероятность вектора генотипов потомков при условии вектора генотипов родителей равна

$$Pr_{\mathcal{S}, \vec{\theta}}(\vec{g}_F, \vec{\mathcal{N}}_F | g_P, \mathcal{N}_P) = \prod_{j \in F} Pr_{\vec{\theta}}(g_j, \mathcal{N}_j | g_{m_j}, \mathcal{N}_{m_j}, g_{f_j}, \mathcal{N}_{f_j}),$$

где $Pr_{\vec{\theta}}(g_j, \mathcal{N}_j | g_{m_j}, \mathcal{N}_{m_j}, g_{f_j}, \mathcal{N}_{f_j})$ — одинаковое для всех потомков правило наследования.

Майоргенная часть распределения $Pr_{\vec{\theta}}(g_j, \mathcal{N}_j | g_{m_j}, \mathcal{N}_{m_j}, g_{f_j}, \mathcal{N}_{f_j})$ параметризуется с помощью переходных вероятностей⁵ $\tau(A_i | g)$ — вероятность того, что особь с генотипом g передаст потомку аллель A_i .

3. *Совместное распределение вероятностей фенотипов членов родословной при условии их генотипов* $Pr_{\vec{\theta}}(\vec{x} | \vec{g}, \vec{\mathcal{N}})$. При справедливости предположения о том, что фенотип особи строится на основе генотипа только этой особи (формула (1.1)), совместная вероятность вектора фенотипов равна произведению индивидуальных вероятностей реализации фенотипа:

$$Pr_{\vec{\theta}}(\vec{x} | \vec{g}, \vec{\mathcal{N}}) = \prod_{i=1, N} Pr_{\vec{\theta}}(x_i | g_i, \mathcal{N}_i),$$

⁴Стандартно предполагается, что для майоргена выполняется закон Харди-Вайнберга, и, следовательно, распределение частот генотипов $Pr(g)$ в популяции определяется частотами аллелей $p(A_i)$:

$$Pr(g = A_i A_j) = \begin{cases} p(A_i) \cdot p(A_j) + p(A_j) \cdot p(A_i) & , \text{ если } i \neq j \\ p(A_i) \cdot p(A_j) & , \text{ иначе.} \end{cases}$$

В дальнейшем мы будем пользоваться этим предположением.

⁵Через эти параметры распределение, описывающее процесс наследования майоргена, задается как

$$Pr(g = A_i A_j | g_m, g_f) = \begin{cases} \tau(A_i | g_m) \cdot \tau(A_j | g_f) + \tau(A_j | g_m) \cdot \tau(A_i | g_f) & , \text{ если } i \neq j \\ \tau(A_i | g_m) \cdot \tau(A_j | g_f) & , \text{ иначе.} \end{cases}$$

где $Pr_{\vec{\theta}}(x_i | g_i, \mathcal{N}_i)$ — вероятностное правило реализации фенотипа, называемое *функцией пенетрантности*. Эта функция одинакова для всех членов родословной, если выполняются сформулированные ранее предположения о независимости средовых и генетических эффектов.

Если эффекты майоргена, остаточных наследственных и средовых воздействий известны, формула 1.1 позволяет однозначно определить значение признака. Однако средовая модификация e известна лишь с точностью до распределения вероятностей. В этом случае, значение признака при фиксированных генотипах тоже известно с точностью до распределения вероятности, которое и является функцией пенетрантности. Конкретный вид этого распределения определен типом признака, а также набором предположений о характере средовой модификации признака⁶.

В этих терминах функция правдоподобия родословной записывается по формуле полной вероятности (см., например, Феллер, 1984)

$$L(\vec{x}_A | \mathcal{S}, \vec{\theta}) = \log_e \left(\sum_{g^1} \int_{\mathcal{N}_1} \dots \sum_{g^N} \int_{\mathcal{N}_N} \left[\prod_{i \in A} Pr_{\vec{\theta}}(x_i | g_i, \mathcal{N}_i) \cdot \prod_{j \in P} Pr_{\vec{\theta}}(g_j, \mathcal{N}_j) \cdot \prod_{k \in F} Pr_{\vec{\theta}}(g_k, \mathcal{N}_k | g_{k_m}, \mathcal{N}_{k_m}, g_{k_f}, \mathcal{N}_{k_f}) \right] d\mathcal{N}_1 \dots d\mathcal{N}_N \right) \quad (1.2)$$

В том случае, если компонента \mathcal{N} дискретна, интегрирование можно заменить на суммирование.

Значение функция правдоподобия определено стандартными параметрами, описывающими распределение майоргена в родословной ($p(A_i)$ и $\tau(A_i | g)$), а также дополнительными параметрами, конкретный набор которых зависит от рассматриваемой модели наследования и типа признака (количественный или качественный). Один из простейших примеров подсчета функции правдоподобия приведен в приложении А.

⁶Например, если рассматривается количественный признак и предполагается действие множественных независимых средовых факторов, средовая модификация распределена по нормальному закону.

Модели сегрегационного анализа количественных признаков

Количественными называются признаки, определенные на оси действительных чисел (рост, вес и т.д.).

При анализе количественных признаков часто предполагается, что остаточная наследуемая компонента \mathcal{N} — это полиген G , определенный как большое число генов малого эффекта (Fisher, 1918). Они наследуются и действуют на признак независимо как от майоргена, так и друг от друга (Morton, MacLean, 1974). Такие предположения позволяют сформулировать модель, называемую смешанной. В рамках смешанной модели, значение признака у некоторой особи определяется как сумма эффектов майоргенотипа, полигенотипа и среды: $x_i = m(g_i) + G_i + e$ (Morton, MacLean, 1974; Lange et al., 1976; Hasstedt, 1982, 1991). Без потери общности можно считать, что математическое ожидание вклада любых двух компонент равно нулю (см., например, Amos, 1994). Обычно считается, что это средовая и полигенная компоненты; тогда $m(g)$ имеет смысл среднего значения признака у особей с фиксированным майоргенотипом g .

В рамках этих предположений распределения вероятностей майоргенной и остаточной наследуемой компонент независимы, следовательно их совместное распределение в популяции — $Pr(g, G)$ — описывается произведением

$$Pr_{\vec{\theta}}(g) \cdot Pr_{\vec{\theta}}(G),$$

а у потомков — произведением

$$Pr_{\vec{\theta}}(g_k | g_{k_m}, g_{k_f}) \cdot Pr_{\vec{\theta}}(G_k | G_{k_m}, G_{k_f}).$$

При стандартном для комплексного сегрегационного анализа предположении о том, что в популяции выполняется закон Харди-Вайнберга, популяционное распределение частот майоргенотипов $Pr(g)$ определяется набором частот аллелей, $p(A_I)$. Переходные вероятности $\tau(A_i | g)$ задают правило наследования $Pr(g | g_m, g_f)$ и, следовательно, распределение вероятности генотипов у потомков.

Считается, что в популяции полигенотип распределен нормально с центром в нуле и дисперсией σ_G^2 , так что вероятность встретить полигенотип G может считаться равной плотности $N(0, \sigma_G^2)$ в точке G :

$$Pr_{\vec{\theta}}(G) = \varphi_G(0, \sigma_G^2).$$

Сегрегация аллелей локусов полигена определяет то, что при фиксированных полигенотипах родителей полигенотип потомка распределен по нормальному закону с центром, равным полусумме полигенотипов родителей и дисперсией $\sigma_G^2/2$ (Lange et al., 1976):

$$Pr_{\vec{\theta}}(G | G_m, G_f) = \varphi_G \left(\frac{G_m + G_f}{2}, \sigma_G^2/2 \right).$$

Стандартно предполагается, что случайных средовых эффектов бесконечно много и они действуют на признак аддитивно. Поэтому при фиксированном майор- и полигенотипе признак распределен нормально с некоторой дисперсией σ_e^2 , обусловленной средовыми воздействиями (Morton, MacLean, 1974). Функция пенетрантности равна

$$Pr_{\vec{\theta}}(x_i | g_i, G_i) = \varphi_{x_i}(m(g_i) + G_i, \sigma_e^2).$$

Таким образом, при смешанной модели функция правдоподобия родословной определяется популяционными частотами аллелей майоргена, переходными вероятностями, майоргенными средними $m(g)$, а также полигенной и средовой дисперсиями:

$$\vec{\theta} = \{p(A_i), m(g), \tau(A_i|g), \sigma_G^2, \sigma_e^2\}.$$

Смешанная модель наследования хорошо зарекомендовала себя при анализе большого количества признаков (Borecki et al., 1986; Clementi et al., 1997; Ruiz et al., 1997). Эта модель в различных вариантах была реализована в нескольких пакетах программ для сегрегационного анализа признаков (POINTER, PAP, SALP).

Подсчет функции правдоподобия в рамках смешанной модели требует одновременного суммирования по всем возможным майоргенотипам и интегрирования по полигенотипам. Для проведения такого суммирования или интегрирования в отдельности разработаны быстрые эффективные алгоритмы (Elston, Stewart, 1971; Cannings et al., 1978; Hasstedt, 1982, 1991; Lange, 1997, см. также раздел 1.3, стр. 25). Однако одновременное суммирование и интегрирование часто требует настолько больших компьютерных ресурсов, что анализ становится невозможным. Поэтому часто применяются упрощенные варианты смешанной модели. Эти варианты получены в рамках ряда дополнительных предположений, касающихся относительного вклада майор- и полигенотипа в определение полиморфизма признака в популяции (Elston et al., 1992; Sham, 1998), который можно охарактеризовать отношением: $\sigma_g^2/(\sigma_g^2 + \sigma_G^2)$, где σ_g^2 — та часть популяционной дисперсии признака, которая вызванная сегрегацией аллелей майоргена⁷. Если это отношение близко к 1, смешанная модель переходит в модель

⁷ $\sigma_g^2 = \sum_{i=1,|g|} Pr_{\vec{\theta}}(g_i) \cdot (E[m(g)] - m(g_i))^2$

майоргенную, если $k = 0$ — в полигенную.

Майоргенная модель

Майоргенная модель предполагает, что разнообразие признака в популяции обеспечивается, в основном, майоргенной компонентой, т.е. $\sigma_g^2 \gg \sigma_G^2$. Это предположение позволяет исключить интегрирование по полигенотипу. Вместо точного значения полигенотипа особи используется его математическое ожидание. Дисперсия признака при фиксированном майоргенотипе определяется суммой средовой и полигенной дисперсией.

Майоргенная модель описывается параметрами

$$\vec{\theta} = \{p(A_i), m(g), \tau(A_i|g), (\sigma_e^2 + \sigma_G^2)\}.$$

Майоргенные модели внесли большой вклад в развитие генетики: именно с помощью различных модификаций этих моделей были найдены, а в дальнейшем картированы и клонированы гены многих менделевских болезней (Williams, Anderson, 1984; Newman et al., 1988; Hall et al., 1990; Wooster et al., 1994).

Полигенная модель

Полигенная модель, напротив, предполагает, что вариация признака в популяции детерминирована, в основном, полигенотипом: $\sigma_G^2 \gg \sigma_g^2$. Это предположение позволяет исключить из формулы 1.2 суммирование по майоргенотипам. Вместо точного значения майоргенной модификации можно взять ее математическое ожидание⁸ μ . Дисперсия, вызванная сегрегацией аллелей гена g , частично включается в дисперсию признака, объясняемую сегрегацией полигена, а частично — в средовую дисперсию.

Полигенная модель описывается параметрами (Lange, 1976)

$$\vec{\theta} = \{\mu, \sigma_G^2, \sigma_e^2\}.$$

При анализе данных с помощью полигенной модели наибольший интерес представляет оценка наследуемости, определенная как относительный вклад генетической компоненты в вариацию признака в популяции. Полигенная модель является одной из старейших моделей наследования количественного признака, она внесла большой

⁸ $\mu = E[m(g)] = \sum_{i=1,|g|} Pr_{\vec{\theta}}(g_i) \cdot m(g_i).$

вклад в развитие генетики количественных признаков (Thoday, 1961; Mather, Jinks, 1982; Falconer, 1989).

В сегрегационном анализе, целью которого является нахождение и описание менделевского гена, полигенная модель имеет вспомогательное значение: на основе этой модели формулируются альтернативные гипотезы о возможном контроле признака (см. раздел 1.2).

Модели сегрегационного анализа бинарных признаков

Бинарные признаки являются простейшим случаем качественных признаков. Бинарные признаки имеют два возможных фенотипа, например, больной/здоровый, есть/нет, или, обобщая, 0 либо 1. При анализе бинарных признаков используют либо специально разработанные модели, либо модификации моделей наследования количественных признаков.

Одной из наиболее простых специальных моделей является модель майоргенного контроля бинарного признака, предполагающая, что полиморфизм признака в родословной обусловлен, в основном, сегрегацией аллелей гена большого эффекта, g . Популяционное распределение и наследование майоргена описывается стандартно.

Функцией пенетрантности является распределение вероятностей $Pr(x|g)$. Это распределение описывается параметрами w_g , определенными как вероятность реализации фенотипа 1 у особи с генотипом g . Понятно, что вероятность фенотипа 0 при условии того же генотипа равна $(1 - w_g)$. Параметры w_g называются пенетрантностями генотипов (Инге-Вечтомов, 1989).

Функция правдоподобия родословной определяется параметрами

$$\vec{\theta} = \{p(A_i), w_g, \tau(A_i|g)\}.$$

Пример подсчета функции правдоподобия при такой модели приведен в приложении А.

При анализе наследования бинарных признаков част феномен "неполной пенетрантности", когда майоргенотип особи определяет ее фенотип неоднозначно ($0 < w_g < 1$). Неполная пенетрантность может возникнуть, если в контроле признака, кроме гена основного эффекта, принимают участие другие гены или признак подвержен средовым модификациям. Однако, описанная модель не позволяет разделить эффекты среды и эффекты остаточных наследуемых факторов.

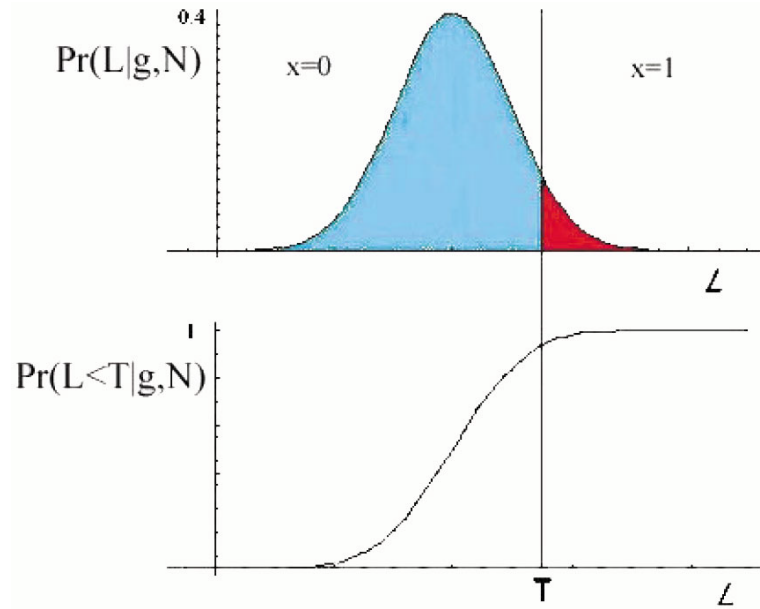


Рис. 1.1: Пороговая модификация: функция пенетрантности количественного признака и кумулятивная функция пенетрантности

Пороговая модификация модели наследования количественного признака

Пороговая модификация основывается на предположении, что в основе наблюдаемого бинарного признака $x \in \{0, 1\}$ лежит некоторая скрытая количественная характеристика $L \in (-\infty, +\infty)$. Считается, что если L не превышает некоторого порогового значения T , признак x принимает значение 0; если же $L > T$, бинарный признак принимает значение 1 (см. рис. 1.1).

Для скрытой характеристики L предполагается, что она является функцией различных факторов, как генетической, так и не генетической природы. Примером реализации подобной модели в природе может служить такой признак, как предрасположенность к инфаркту миокарда. В данном случае x принимает значения {здоров, болен}. В основе же признака лежат множественные количественные факторы, как-то: уровень холестерина в крови, возраст, пол, диета, курение, генетическая предрасположенность и т.д.

Пороговая модель требует видоизменить функцию пенетрантности: если в случае количественной характеристики вероятность реализации фенотипа L определяется функцией пенетрантности $Pr(L|g, \mathcal{N})$, выраженной через плотность распределения вероятности, то вероятность того, что особь будет иметь фенотип 0, равна *кумулятивной* функции пенетрантности $Pr(L < T|g, \mathcal{N})$, выражаемой через функцию

распределения вероятности (см. рис. 1.1).

Для описания наследования скрытой характеристики в родословной может быть применена любая из перечисленных выше моделей детерминации количественного признака.

Например, если применяется *смешанная модель наследования*, кумулятивная функция пенетрантности имеет вид

$$Pr(L < T | g, G) = \int_{-\infty}^T \varphi_x(m(g) + G, \sigma_e^2) dL = \Phi_T(m(g) + G, \sigma_e^2).$$

Многократное вычисление подобного интеграла занимает много времени, поэтому необходимо либо пользоваться таблицами функции нормального распределения, либо, что делается чаще, использовать логистическое приближение этой функции (Феллер, 1984):

$$Pr(L < T | g, G) = \frac{1}{1 + \exp \left\{ \frac{\pi(T - m(g) - G)}{\sqrt{3}\sigma_e^2} \right\}}.$$

Пороговая модификация смешанной модели наследования количественного признака позволяет определить не только эффекты майоргена, но и относительный вклад среды и полигенотипа в детерминацию признака.

Пороговая модификация майоргенной модели наследования количественного признака предполагает, что вклад полигенотипа в детерминацию популяционного полиморфизма признака мал по сравнению с вкладом майоргенотипа. Кумулятивная функция пенетрантности равна

$$Pr(L < T | g) = \frac{1}{1 + \exp \left\{ \frac{\pi(T - m(g))}{\sqrt{3}\sigma^2} \right\}}.$$

Хотя эта функция описывается пятью параметрами — $m(A_1A_1)$, $m(A_1A_2)$, $m(A_2A_2)$, σ и T , однако только три зависящих от майоргенотипа параметра $m(g)$ являются независимыми и могут быть оценены. Как дисперсию признака, так и пороговое значение T можно произвольно зафиксировать. В этом случае существует взаимоднозначное преобразование, позволяющее перейти от параметров $m(g)$ к пенетрантностям w_g и наоборот:

$$w_g = 1 - \frac{1}{1 + \exp \left\{ \frac{\pi(T - m(g))}{\sqrt{3}\sigma^2} \right\}}, \text{ а}$$

$$m(g) = \frac{\sqrt{3}\sigma^2}{\pi} \cdot \log_e \left\{ \frac{1 - w_g}{w_g} \right\} + T.$$

Следовательно, пороговая модификация майоргенной модели наследования количественного признака и майоргенная модель наследования бинарного признака являются практически эквивалентными⁹.

Пороговая модификация позволяет перенести богатый арсенал сложных моделей наследования, разработанных для анализа количественных признаков, на случай бинарных признаков. Эта модификация часто применяется при сегрегационном анализе бинарных признаков со сложным механизмом контроля, она позволила получить ряд важных результатов (Hecht et al., 1991; Nemana et al., 1992; Ray et al., 1993; Jarvik et al., 1994; Hanson, Knowler, 1998).

1.2 Тестирование гипотез

Основной вопрос сегрегационного анализа это вопрос о том, принимает ли некоторый менделевский ген участие в контроле признака. Для решения этого вопроса разработан ряд критериев, основанных на статистическом сравнении специальных генетических гипотез. Наиболее общим из них является критерий Эльстона-Стюарта-Мортон (Elston, Stewart, 1971; Morton, MacLean, 1974; Lalouel et al., 1983; Demenais et al., 1986).

Если участие майоргена в контроле признака подтверждается, решается ряд дополнительных вопросов: например, можно ли описать наследование признака в рамках доминантной модели, взаимодействует ли найденный ген с другими известными генами и т.д.

Критерий отношения правдоподобия и информационный критерий Акаике

Для тестирования генетических гипотез, как правило, используется критерий отношения правдоподобия, КОП (Neuman, Pearson, 1928). Пусть $L_1 = \log_e Pr(X|\hat{\theta} = \{\hat{\vartheta}_1, \hat{\vartheta}_2, \dots, \hat{\vartheta}_n\}, \mathcal{S})$ есть максимальное значение функции правдоподобия выборки родословных при справедливости гипотезы, определяемой n генетическими параметрами ϑ , для которых методом максимума правдоподобия найдены оценки; $L_2 =$

⁹Единственным исключением является то, при майоргенной модели наследования бинарного признака пенетрантности определены на отрезке $[0, 1]$, а при пороговой модификации — на отрезке $(0, 1)$. Однако для практических нужд сегрегационного анализа эта разница никакой роли не играет.

$\log_e Pr(X | \hat{\theta} = \{\hat{\vartheta}_1, \hat{\vartheta}_2, \dots, \hat{\vartheta}_m, \vartheta_{m+1}, \dots, \vartheta_n\}, \mathcal{S})$ — то же в случае, когда первые m параметров оцениваются, а значения остальных $(n - m)$ параметров заранее заданы. Тогда распределение $2 \cdot (L_1 - L_2)$ имеет в асимптотике распределение χ^2 с числом степеней свободы $df = n - m$.

Видно, что с помощью КОП можно тестировать только иерархические пары гипотез, когда одна из них является частным случаем другой. Иногда необходимо сравнивать гипотезы, не являющиеся иерархическими. В этом случае на помощь приходит информационный критерий Акаике, определенный как $AIC = -2 \cdot (\log_e L(X | \hat{\theta}, \mathcal{S}) - n)$, где n — число независимо оцениваемых параметров. Наилучшей считается гипотеза, которой соответствует минимум AIC (Akaike, 1974).

Информационный критерий Акаике не является статистическим критерием в традиционном смысле этого слова — из набора предложенных гипотез он позволяет выбрать лучшую, но не позволяет ответить на вопрос, насколько достоверно она отличается от остальных.

Критерий Эльстона-Стюарта-Мортонна

Этот критерий является вариантом КОП. Он предполагает сравнение четырех генетических гипотез. Одна из них ("неограниченная") предполагает оценку всех параметров модели, остальные являются частными случаями первой:

- **Менделевская гипотеза** накладывает только одно ограничение на модель: правило, определяющее наследование компоненты g , является менделевским, т.е. при формировании гаметы у особи любого генотипа сегрегация двух составляющих его аллелей равновероятна. Для диаллельного майоргена, гипотеза формализуется в виде ограничения

$$\tau_{A_1 A_1} = 1, \tau_{A_1 A_2} = 0.5 \text{ и } \tau_{A_2 A_2} = 0.$$

- **Внемайоргенная гипотеза**¹⁰ предполагает отсутствие эффекта майоргена, т.е. изменение майоргенотипа не меняет значение признака. Вариация признака

¹⁰Иногда (особенно в случае, когда остаточная наследуемая компонента не выражена в явном виде) эта гипотеза из рассмотрения исключается. Критерий, основанный на сравнении оставшихся трех гипотез, часто называют критерием Эльстона-Стюарта.

в родословной объясняются только немайоргенными компонентами (средовой и остаточной наследуемой). Гипотеза формализуется с помощью ограничения

$$m(g) = \hat{\mu} \text{ для любого } g.$$

- **Гипотеза сложного средового контроля** предполагает, что вариация признака в родословной определяется остаточной наследуемой компонентой и сложными средовыми эффектами. Гипотеза формализуется с помощью ограничения

$$\tau_{A_1 A_1} = \tau_{A_1 A_2} = \tau_{A_2 A_2} = \hat{q}.$$

Гипотеза о майоргенном контроле признака принимается, если менделевская гипотеза не отличается значимо от неограниченной, а гипотезы, предполагающие отсутствие майоргенной компоненты, достоверно хуже неограниченной гипотезы.

1.3 Вычислительные алгоритмы

Для тестирования генетических гипотез необходимо максимизировать функцию правдоподобия, что требует многократного ее вычисления. Время счета функции правдоподобия пропорционально числу необходимых математических операций. Предположим, что модель включает в себя только дискретную генетическую компоненту \mathcal{G} ¹¹. Если число особей в родословной равно N , то однократный подсчет функции правдоподобия по общей формуле (1.2) требует $\mathcal{L} = |\mathcal{G}|^N \cdot (2 \cdot N - 1)$ операций. Таким образом, при увеличении родословной трудоемкость возрастает экспоненциально¹².

Для родословной малого размера ($N=20$) и простейшей майоргенной гипотезы ($|\mathcal{G}| = 3$) время прямого подсчета функции правдоподобия равно 5 минутам (Pentium 166MMX, MS Fortran 32bit compiler). Уже для родословной среднего размера ($N \sim 50$) время подсчета абсолютно неприемлемо (более миллиарда лет).

Подобная ситуация породила многочисленные вычислительные алгоритмы, которые позволяют осуществить подсчет функции правдоподобия с разумными затрата-

¹¹Такой дискретной компонентой может быть майорген g , или набор из нескольких генов.

¹²Например, при введении одной дополнительной особи трудоемкость возрастает в $\frac{\mathcal{L}_{N+1}}{\mathcal{L}_N} = \frac{|\mathcal{G}|^{(N+1)} \cdot (2 \cdot (N+1) - 1)}{|\mathcal{G}|^N \cdot (2 \cdot N - 1)} \approx |\mathcal{G}|$ раз.

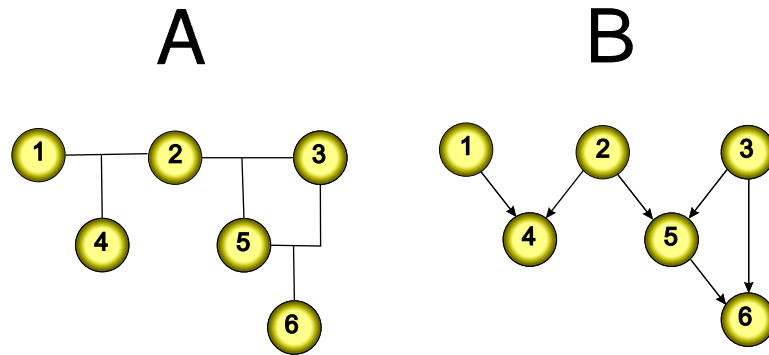


Рис. 1.2: Родословная, изображенная в стандартном виде (А) и в виде ориентированного графа (В)

ми компьютерных ресурсов и времени. Для описания этих алгоритмов необходимо ввести определенную терминологию.

Терминология

При работе с родословными часто используются термины теории графов. Действительно, родословную можно представить в виде ориентированного графа, в котором вершины обозначают особей, а дуги — родственные связи между ними, причем ориентация дуг соответствует направлению передачи генетического материала (рис. 1.2).

Говорят, что родословная *не содержит петель*, если соответствующий граф не содержит циклов, т.е. между любыми двумя вершинами можно проложить только один путь. Особь называется *родоначальником*, если в соответствующую ей вершину не входит ни одна дуга, т.е. родители этой особи не представлены в родословной (особи 1, 2 и 3 рис. 1.2). Потомок называется *свободным*, если из соответствующей ему вершины не выходит ни одна дуга (особи 4 и 6). *Шарнирным набором* для некоторой части родословной называется набор особей, при удалении которых эта часть отделяется от остальной родословной.

Родословная, состоящая только из пары родителей и их потомков, называется *ядерной родословной*. В родословной, изображенной на рисунке, можно выделить три ядерных родословных: (1, 2, 4), (2, 3, 5) и (3, 5, 6). *Терминальной* называется ядерная родословная, для которой шарнирный набор состоит из одной шарнирной особи (особь 2 для ядерной родословной (1, 2, 4)).

Свертка родословной

В основе алгоритма подсчета функции правдоподобия больших родословных лежит предположение о том, что при фиксированных генотипах шарнирных особей вероятности двух частей родословной, разделенных ими, независимы. Это предположение справедливо для широкого класса моделей, не рассматривающих негенетические фенотипические корреляции между дальними родственниками. Оно позволяет производить подсчет вероятности родословной по частям, т.е., фактически, проводить суммирование (интегрирование) по генотипам некоторых особей независимо от генотипов остальных особей в родословной. Вычислительная процедура, с помощью которой производится подсчет вероятности части родословной при фиксированных генотипах шарнирных особей, называется *сверткой* части родословной *на* шарнирных особей (Cannings et al., 1978; Thomas, 1986; Sham, 1998). Эти вероятности запоминаются, а свернутая часть родословной исключается из дальнейшего рассмотрения.

Функция правдоподобия родословной без петель

Эффективный алгоритм подсчета функции правдоподобия родословной без петель основывается на последовательной свертке терминальных ядерных родословных. При каждом шаге свертки информация о членах ядерной родословной переносится на шарнирную особь, вследствие чего родословная постепенно редуцируется. При этом шарнирная особь меняет свой статус: если она была потомком в свернутой ядерной родословной, то в редуцированной родословной она становится родоначальником; если она была родителем в свернутой ядерной родословной, то в редуцированной родословной она может стать свободным потомком.

При применении алгоритма последовательной свертки требования к машинной памяти минимальны, а время подсчета функции правдоподобия родословной растет примерно линейно с увеличением размера родословной. Таким образом, вычисление функции правдоподобия родословной без петель не представляется проблематичным.

Функция правдоподобия родословной с множественными петлями

Подсчет функции правдоподобия родословной с малым количеством петель также не представляет проблем.

Один из способов подсчета функции правдоподобия родословных с петлями был

предложен Оттом (Ott, 1976; Lange, Elston, 1975). Этот способ заключается в выделении небольшого числа особей, фиксация генотипов которых позволяет представить родословную в виде графа-дерева, не содержащего петель. Последовательный подсчет вероятностей родословной при всех возможных генотипах выделенных членов родословной с последующим их суммированием позволяет избежать потери информации. Ясно, что число особей K , у которых фиксируется генотип, не может быть большим, т.к. вероятность родословной вычисляется при всех возможных наборах генотипов фиксированных особей, а число таких наборов равно $|\mathcal{G}|$. Следовательно, время подсчета функции правдоподобия увеличивается в $|\mathcal{G}|^K$ раз. Применение этого алгоритма в родословных с множественными петлями невозможно.

Второе решение заключается в использовании обобщенных сверток (Cannings et al., 1978). В этом случае время счета функции правдоподобия сравнимо со временем счета функции правдоподобия родословной без петель, т.е. растет приблизительно линейно с увеличением размера родословной. Применение этого алгоритма в родословной с множественными петлями часто невозможно, так как обобщенные свертки могут требовать большой объем памяти. В общем случае свертка на шарнирный набор, состоящий из K особей, потребует $|\mathcal{G}|^K$ единиц памяти¹³. Подобная ситуация порождает проблему поиска оптимальной последовательности сверток (т.е. последовательности, которая минимизирует используемые ресурсы). Эта проблема относится к классу NP-полных задач теории графов: если родословная состоит из N особей, для нахождения оптимальной последовательности требуется перебор $N!$ возможных вариантов (Garvil, 1977; Thomas, 1986).

Таким образом, если число петель в родословной велико, описанные алгоритмы вычисления функции правдоподобия не применимы, поскольку требования к компьютерным ресурсам и времени счета стремительно возрастают с увеличением числа петель.

При анализе больших родословных сложной структуры многообещающими кажутся аппроксимационные, т.е. приближительные, методы подсчета ФП. Один из таких методов основан на Монте-Карловском подходе (Lange, Matthise, 1989; Lange, Sobel, 1991; Sheehan, 1992; Thompson, Guo, 1991; Thompson et al., 1993; Guo, Thompson, 1994). Этот метод имеет хорошие асимптотические свойства, однако он

¹³В том случае, если массивы сверток хранятся в формате с двойной точностью (double precision), потребуется $8 \cdot |\mathcal{G}|^K$ байт памяти.

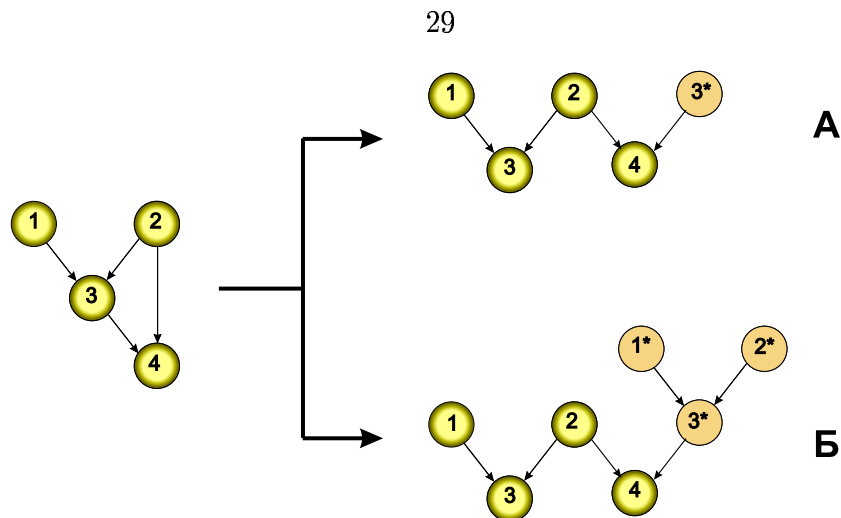


Рис. 1.3: Различные способы разрыва петель. А — дублирование особи, Б — дублирование с дополнительным расширением. i^* — фенотипическая копия особи i .

может потребовать огромных вычислительных ресурсов. Кроме того, его использование требует дополнительных теоретических разработок, которые проделаны только для некоторых классов моделей наследования. Все это ограничивает использование указанного метода (Lange, Matthysse, 1989; Sheehan, Thomas, 1993; Jensen et al., 1995).

Другой метод, применяемый для приблизительного подсчета функции правдоподобия родословных с множественными петлями, основывается на упрощении структуры родословной (Stricker et al., 1995; Wang et al., 1996). Этот метод предполагает разрыв всех петель путем дублирования некоторых особей (рис. 1.3), в результате чего преобразуется структура родословной. Получившаяся в результате родословная называется *расширенной*. Функция правдоподобия расширенной родословной может быть подсчитана с помощью существующего эффективного алгоритма, она является аппроксимацией реальной функции правдоподобия. Достоинством метода аппроксимации с помощью разрыва петель является то, что он позволяет анализировать наследование признаков в очень больших родословных сложной структуры (Stricker et al., 1995).

Было показано, что точность аппроксимации может быть улучшена с помощью ряда стратегий (Wang et al., 1996):

- условного подсчета — функция правдоподобия считается как логарифм условной вероятности распределения признака в расширенной родословной при известных фенотипах дублированных особей;

- дополнительного расширения — петля разрывается дублированием не только некоторой особи, но и ее родственников (рис. 1.3Б).

Почему эти приемы позволяют улучшить аппроксимацию, и какие способы могут быть предложены для еще большего увеличения точности аппроксимации?

Совместное распределение генотипов в родословной определяется двумя характеристиками: априорным совместным распределением генотипов кровных родственников, которое определяется только структурой родословной, и апостериорным распределением вероятностей индивидуальных генотипов, которое при заданном априорном совместном распределении определяется фенотипами особей в родословной.

Разрыв петли с помощью дублирования некоторой особи сопровождается двумя типами негативных эффектов: 1) искажением индивидуального апостериорного распределения вероятностей генотипов дублированной особи (и, как следствие, особей, связанных с нею близким родством) и 2) изменением структуры родословной, что искажает априорное совместное распределение генотипов.

Дополнительное расширение и условный подсчет функции правдоподобия позволяет минимизировать негативные эффекты первого типа. При достаточно больших дополнительных расширениях можно полностью восстановить индивидуальное апостериорное распределение генотипов (Wang et al., 1996), т.е. свести эффекты 1-го типа к нулю.

Негативные эффекты второго типа свести к нулю можно только в специальном случае¹⁴, во всех остальных их можно только минимизировать с помощью применения специальных стратегий (Wang et al., 1986).

При разрыве петли степень искажения априорного распределения пропорциональна числу и степени родства особей, родственные связи между которыми утрачиваются. Максимального искажения априорного распределения следует ожидать при разрыве коротких инбредных петель: именно в таких петлях априорное совместное распределение максимально отличается от независимого. Кроме того, при разрыве малых петель следует ожидать максимального искажения индивидуального апостериорного распределения, так как материала для его восстановления мало.

¹⁴Если дополнительное расширение позволяет выделить только *один* возможный вариант генотипа для каждой дублируемой особи, то функция правдоподобия, полученная с помощью алгоритма дополнительного расширения-обуславливания аналитически равна точной функции правдоподобия. Подобное возможно только для определенного и довольно узкого класса функций пенетрантности.

Таким образом, для улучшения точности аппроксимации ФП в родословных с множественными петлями необходимо разработать алгоритм, который позволит избежать разрыва коротких петель.

1.4 Заключение

При анализе данных гибридологического эксперимента ответ на вопрос, можно ли описать наследование изучаемого признака в рамках некоторой гипотезы, дается в явном виде.

При комплексном сегрегационном анализе тестирование проводится при помощи критерия отношения правдоподобия, который позволяет сравнивать гипотезы, *сформулированные в рамках определенной модели*.

Таким образом, формулировка корректных и как можно более полных моделей наследования составляет одну из ключевых задач комплексного сегрегационного анализа.

В разделе 1.1 мы описали общую генетико-математическую модель и ряд стандартных моделей, используемых в настоящее время при комплексном сегрегационном анализе. Эти модели разработаны в рамках набора явных и неявных предположений, которые ограничивают их применимость.

Так, стандартный вид функции пенетрантности предполагает, что исследуемый признак является характеристикой одной особи. Между тем известны признаки ("совместные"), которые правильнее рассматривать как результат взаимодействия двух генотипов: классическим примером такого признака является гемолитическая желтуха новорожденных, которая вызвана взаимодействием Rh-генотипов матери и плода (Zimmerman, 1973). Плодовитость некоторых млекопитающих может зависеть от генотипов пары родителей (Penrose, 1949; Bodmer, 1965; Feldman, Liberman, 1985; Kasser et al., 1986; Diltz, 1991).

При формализации популяционного распределения генотипов стандартные модели используют предположение о том, что все родоначальники являются представителями одной популяции, характеризуемой определенными частотами аллелей. Это предположение может не выполняться: родословные сельскохозяйственных и лабораторных животных часто получены при гибридизации двух родительских форм (Тихонов, 1990; Бекенев, 1997), длительная генетическая изоляция которых могла

привести к дивергенции частот аллелей.

Кроме того, стандартные модели в явном виде описывают только сегрегацию ядерных генов. Известны признаки, которые находятся под контролем не только ядерных, но и цитоплазматических генов (Хагеман, 1962; Сэджер, 1975).

Правило, формализующее процесс наследования (пункт 2 стр. 15), предполагает, что каждая особь является потомком двух родителей. Это не всегда так: например, у растений встречается самоопыление, апомиксис и т.д.

Существующие эффективные алгоритмы подсчета функции правдоподобия позволяют анализировать родословные без петель и родословные с малым количеством петель. Для многих родословных животных это требование не выполняется: в них часто встречаются петли, в том числе и короткие инбредные. Сегрегационный анализ признаков по таким родословным требует разработки новых алгоритмов.

Таким образом, расширение возможностей комплексного сегрегационного анализа и повышение его мощности требует пересмотра некоторых стандартных предположений и развития в свете переформулированных предположений новых моделей наследования признаков и алгоритмов анализа.

В данной работе были поставлены следующие конкретные задачи:

1. Разработать ряд новых методов, учитывающих специфику наследования признаков и особенности структуры родословных, полученных при разведении сельскохозяйственных и лабораторных животных. В рамках этой задачи выделяются три подзадачи:
 - Разработка методов анализа наследования признаков в родословных, полученных при гибридизации генетически изолированных форм.
 - Разработка методов анализа совместных признаков, характеризующих пару особей.
 - Разработка алгоритмов анализа родословных с множественными петлями.
2. На основе разработанных методов провести анализ наследования некоторых признаков животных.
3. Показать принципиальную возможность использования метода комплексного сегрегационного анализа признаков растений.

Глава 2

Разработка моделей наследования

Создание новых моделей наследования является насущной задачей современного комплексного сегрегационного анализа: оно позволит повысить корректность и мощность анализа, а также расширить спектр анализируемых признаков и объектов.

В данной главе представлены два аспекта комплексного сегрегационного анализа, нуждающихся в дальнейшей разработке:

1. Сегрегационный анализ признаков по родословным, в которых родоначальники являются представителями различных популяций (гибридных родословных). Для этих родословных были разработаны некоторые классы моделей (Аксенович, 1999), однако остались не формализованными многие важные частные случаи.
2. Анализ наследования совместных признаков. Практически все существующие модели наследования направлены на анализ *индивидуальных* признаков. В то же время существует ряд признаков, характеризующих группу особей. Простейшими групповыми признаками являются *совместные* признаки, характеризующие пару особей.

2.1 Модели наследования признаков в гибридных родословных

Целью исследования, представленного в данном разделе, является дальнейшая разработка моделей наследования признаков в родословных, полученных при скрещивании представителей двух популяций. Ранее (Аксенович, 1999), была сформулирована

смешанная модель наследования количественных признаков в гибридных родословных. Вычисление функции правдоподобия при этой модели чрезвычайно трудоемко, поэтому она применима только для анализа малых родословных. Анализ средних и больших родословных требует сформулировать варианты этой модели, которые позволяют быстро вычислять функцию правдоподобия. Анализ наследования бинарных признаков в гибридных родословных также требует разработки новых моделей наследования. В этом разделе описаны такие модели: майоргенная модель наследования признаков в гибридных родословных и пороговая модификация смешанной и майоргенной модели.

Известно, что в потомстве, полученном при гибридизации представителей разных популяций, могут наблюдаться генетические эффекты, не встречающиеся в родословных, сформированных представителями одной популяции, например, эффект гетерозиса. В этом разделе описано, как этот эффект может быть формализован и введен в смешанные и майоргенные модели наследования признаков.

Прежде чем перейти к описанию полученных результатов, кратко сформулируем основные положения смешанной модели наследования количественного признака в гибридных родословных, пользуясь терминологией, введенной в предыдущей главе.

Смешанная модель наследования признака в гибридной родословной

Рассмотрим родословную, полученную при скрещивании представителей двух популяций, P_1 и P_2 . Функция правдоподобия такой родословной (как и любой родословной) может быть подсчитана по формуле

$$L(\vec{x} | \mathcal{S}, \vec{\theta}) = \log_e \left(\sum_{g_1} \int_{\mathcal{N}_1} \dots \sum_{g_N} \int_{\mathcal{N}_N} \left[\prod_{i \in A} Pr_{\vec{\theta}}(x_i | g_i, \mathcal{N}_i) \cdot \prod_{j \in P} Pr_{\vec{\theta}}(g_j, \mathcal{N}_j) \cdot \prod_{k \in F} Pr_{\vec{\theta}}(g_k, \mathcal{N}_k | g_{k_m}, \mathcal{N}_{k_m}, g_{k_f}, \mathcal{N}_{k_f}) \right] d\mathcal{N}_1 \dots d\mathcal{N}_N \right)$$

Строго говоря, описание наследования количественного признака в двух популяциях и вычисление функции правдоподобия требует задания двух наборов генетических параметров. Однако, этого недостаточно, чтобы описать наследование признака у всех членов родословной — необходимо также ввести свой набор генетических параметров для каждого типа гибридов. Понятно, что этот способ описания наследования является тупиковым с точки зрения комплексного сегрегационного анализа:

число параметров стремительно возрастает и, следовательно, растет сложность их оценивания и интерпретации полученных результатов. Решить проблему можно, если ввести систему предположений, которая позволит описать наследование признака у всех членах родословной в рамках единой модели.

Описание наследования признака в гибридной родословной с помощью конечного числа генетических параметров возможно лишь тогда, когда выполняются следующие допущения:

1. Вклад в признак отдельных наследственных факторов одинаков как для представителей обеих популяций, так и для их гибридов.
2. Правила передачи наследственных факторов из поколения в поколение одинаковы как для представителей обеих популяций, так и для их гибридов.
3. Средовые воздействия не направленно и одинаково модифицируют значение признака как у представителей обеих популяций, так и у гибридов.

Сам процесс гибридизации предполагает, что скрещиваемые особи находятся под воздействием одних и тех же условий среды. Поэтому третье допущение вполне реально. Первое и второе допущения выполняются в том случае, если правила наследования и реализации признака не меняются на разном генетическом фоне, не учитываемом моделью наследования. Заметим, что это предположение негласно принято и в классическом генетическом анализе: например, считается, что проявление майоргенотипа не зависит от того, принадлежит ли особь родительской породе или является гибридом, а ожидаемое распределение генотипов рассчитывается с помощью единого правила наследования (см., например, Elston, Stewart, 1973).

Из сформулированных предположений следует, что родительские популяции отличаются друг от друга только априорным распределением наследственных факторов.

Смешанная модель предполагает, что остаточная наследуемая компонента \mathcal{N} — это полиген аддитивного действия. Значение признака детерминировано суммой эффектов майоргенотипа, полигенотипа и среды: $x = m(g) + G + e$. Без потери общности можно считать, что математическое ожидание средового эффекта равно нулю для всех членов родословной; математическое ожидание эффекта полигенотипа в популяции P_1 равно нулю, а в P_2 — Δ . В рамках этих предположений $m(g)$ имеет смысл

среднего значения признака у особей с генотипом g в популяции P_1 . В популяции P_2 среднее значение признака у особей с генотипом g равно $m(g) + \Delta$.

Функция пенетрантности описывается плотностью нормального распределения в точке x :

$$Pr_{\vec{\theta}}(x|g, G) = \varphi_x(m(g) + G, \sigma_e^2),$$

где σ_e^2 — дисперсия, обусловленная действием среды.

Априорное популяционное распределение частот майоргенотипов, $Pr_{\vec{\theta}}(g)$, зависит от популяционной принадлежности родоначальника. Для описания распределения частот майоргенотипов в двух популяциях необходимо задать два набора частот аллелей — $p_1(A_i)$ и $p_2(A_i)$. Переходные вероятности $\tau(A_i|g)$, одинаковые для всех членов родословной согласно допущению (2), определяют распределение вероятностей майоргенотипов потомков.

Наследование полигенотипа в родительских популяциях описывается с помощью дисперсий полигенотипа в этих популяциях — σ_{G1}^2 и σ_{G2}^2 , а также математического ожидания полигенного эффекта в популяции P_2 — Δ . Как и в смешанной модели для представителей одной популяции, математическое ожидание полигенотипа потомка равно полусумме полигенотипов родителей. Описание наследования полигенной компоненты у гибридов требует введения дополнительного параметра γ^2 , который имеет смысл возрастания дисперсии полигенотипа в F_2 . Дополнительная дисперсия может возникать потому, что при гибридизации гены, находившиеся в альтернативных гомозиготных состояниях в исходных популяциях, начинают сегрегировать. Было показано, что математическое ожидание дисперсии полигенотипа у определенного типа гибридов описывается формулой (Аксенович, 1999)

$$\sigma_G^2 = (1 - \epsilon)\sigma_{G1}^2 + \epsilon\sigma_{G2}^2 + 2(\epsilon_m(1 - \epsilon_m) + \epsilon_f(1 - \epsilon_f))\gamma^2, \quad (2.1)$$

где ϵ — доля крови P_2 у рассматриваемого типа гибридов, а ϵ_m и ϵ_f — доля крови P_2 у отцов и матерей гибридов ($\epsilon = \frac{\epsilon_m + \epsilon_f}{2}$). Доля крови задается таким образом, что $\epsilon = 0$ у представителей популяции P_1 , $\epsilon = 1$ у представителей популяции P_2 , $\epsilon = 0.5$ для гибридов F_1 и F_2 , $\epsilon = 0.25$ или 0.75 для бэккроссов и т.д. При фиксированном полигенотипе родителей, дисперсия полигенотипа потомка равна $\sigma_G^2/2$.

Смешанная модель наследования признака в гибридной родословной описывается параметрами

$$\vec{\theta} = \{p_1(A_i), p_2(A_i), m(g), \tau(A_i|g), \Delta, \gamma^2, \sigma_{G1}^2, \sigma_{G2}^2, \sigma_e^2\}.$$

Майоргенная модель для гибридной родословной

По аналогии с логикой формализации майоргенной модели для одной популяции предположим, что разнообразие признака в гибридной родословной обеспечивается, в основном, майоргенной компонентой. Это предположение позволяет использовать математическое ожидание эффекта полигенотипа вместо его точного значения и исключить из формулы подсчета функции правдоподобия интегрирование по полигенотипу. Математическое ожидание эффекта полигенотипа равно нулю для представителей популяции P_1 и Δ — для представителей P_2 . Для гибридов математическое ожидание полигенной компоненты пропорционально доли крови P_2 :

$$E[G] = \epsilon \Delta. \quad (2.2)$$

Дисперсия признака у некоторого типа гибридов при фиксированном майоргенотипе равна сумме средовой и полигенной дисперсий. Средовая дисперсия, σ_e^2 , считается одинаковой для всех членов родословной. Математическое ожидание полигенной дисперсии в группе гибридов определенного типа рассчитывается по уравнению 2.1. Таким образом, функция пенетрантности

$$Pr(x|g) = \varphi_x(m(g) + E[G], \sigma^2), \text{ где}$$

$$\sigma^2 = (1 - \epsilon)(\sigma_{G1}^2 + \sigma_e^2) + \epsilon(\sigma_{G2}^2 + \sigma_e^2) + 2(\epsilon_m(1 - \epsilon_m) + \epsilon_f(1 - \epsilon_f))\gamma^2, \quad (2.3)$$

а $E[G]$ определено формулой 2.2.

Популяционное распределение и правило наследования майоргена задаются так же как и при смешанной модели наследования признака в гибридной родословной. Однако, может возникнуть ситуация, когда необходимо определить априорную (а не вычисленную при условии известных генотипов родителей) вероятность генотипа у гибридной особи. Такая ситуация возникает, например, при работе с родословными, содержащими множественные петли. Один из алгоритмов приближенного подсчета функции правдоподобия таких родословных предполагает разрыв петель (Stricker et al., 1995; Wang et al, 1996; см. также раздел 1.3), следствием которого может явиться потеря связи между гибридной особью и ее родителями. В этом случае гибридная особь рассматривается как один из основателей родословной и для нее необходим подсчет априорной вероятности генотипа.

В общем случае, частоты генотипов у определенного класса гибридов не подчиняются закону Харди-Вайнберга: например, в F_1 пары гамет объединяются не случайно — каждая из зигот сформирована гаметой из первой и второй популяций.

Математическое ожидание вероятностей майоргенотипов $Pr(g)$ для гибридов выражаются следующим образом:

$$Pr(g = A_i A_j) = \begin{cases} r_m(A_i) \cdot r_f(A_j) + r_m(A_j) \cdot r_f(A_i) & , \text{ если } i \neq j \\ r_m(A_i) \cdot r_f(A_j) & , \text{ если } i = j \end{cases} \quad (2.4)$$

где $r_k(A_i) = (1 - \epsilon_k) \cdot p_1(A_i) + \epsilon_k \cdot p_2(A_i)$, k принимает значения m или f .

В том случае, если рассматривается диаллельный майорген

$$Pr(A_1 A_1) = (1 - q_1)^2 + 2\epsilon(1 - q_1)(q_1 - q_2) + \epsilon_m \epsilon_f (q_1 - q_2)^2$$

$$Pr(A_1 A_2) = 2q_1(1 - q_1) + 2\epsilon(2q_1 - 1)(q_1 - q_2) - 2\epsilon_m \epsilon_f (q_1 - q_2)^2$$

$$Pr(A_2 A_2) = q_1^2 - 2\epsilon q_1(q_1 - q_2) + \epsilon_m \epsilon_f (q_1 - q_2)^2,$$

где q_1 — частота аллеля A_2 в популяции P_1 , а q_2 — в популяции P_2 .

Легко показать, что формулы 2.4 пригодны не только для гибридных потомков, но и для исходных основателей родословной.

Функция правдоподобия родословной равна

$$L(\vec{x}|\vec{\theta}, \mathcal{S}) = \sum_{g_1} \dots \sum_{g_N} \prod_{i \in A} \varphi_{x_i}(m(g_i) + E[G_i], \sigma_i^2) \cdot \prod_{j \in P} Pr(g_j) \cdot \prod_{k \in F} Pr(g_k | g_{m_k}, g_{f_k}). \quad (2.5)$$

Майоргенная модель для гибридной родословной описывается параметрами

$$\vec{\theta} = \{p_1(A_i), p_2(A_i), m(g), \tau(A_i|g), \Delta, \gamma^2, (\sigma_{G1}^2 + \sigma_e^2), (\sigma_{G2}^2 + \sigma_e^2)\}.$$

Заметим, что майоргенная модель для родословной, полученной из одной популяции, не позволяет разделить эффекты среды и полигенотипа (см. раздел 1.1). Эта же модель, примененная к гибридной родословной, дает возможность определить важные характеристики полигенной компоненты: разницу вкладов полигенотипа в детерминацию признака в двух рассматриваемых популяциях, а также добавочную полигенную дисперсию γ^2 .

Пороговая модификация смешанной и майоргенной моделей

Пороговая модификация смешанной модели, используемая при анализе бинарных признаков, может быть получена аналогично пороговым модификациям, описанным в разделе 1.1.

Предполагается, что в основе полиморфизма бинарного признака, $x \in \{0, 1\}$, лежит полиморфизм ненаблюдаемой количественной характеристики $L \in (-\infty, +\infty)$. Если значение L менее некоторого порога T , то бинарный признак принимает значение 0, иначе — 1.

В том случае, когда для описания наследования ненаблюдаемой характеристики используется смешанная модель наследования признака в гибридной родословной, тогда пенетрантность некоторого генотипа (g, G) , определенная как вероятность реализации фенотипа 1, равна

$$\begin{aligned} w_{g,G} &= 1 - Pr(L < T | g, G) = 1 - \int_{-\infty}^T \varphi_L(m(g) + G, \sigma_e^2) dL \approx \\ &\approx \frac{1}{1 + \exp\left\{\frac{\pi(T - (m(g) + G))}{\sqrt{3}\sigma_e^2}\right\}}. \end{aligned}$$

Как и в случае пороговой модификации смешанной модели для одной популяции (см. раздел 1.1), при оценивании параметров пороговое значение T и средовая дисперсия могут быть произвольно зафиксированы.

Пороговая модификация майоргенной модели наследования предполагает использование математического ожидания эффекта полигенотипа, $E[G]$, вместо точного значения G и позволяет исключить интегрирование по полигенотипу при подсчете функции правдоподобия. Пенетрантности майоргенотипов определяются формулой

$$\begin{aligned} w_g &= 1 - Pr(L < T | g) = 1 - \int_{-\infty}^T \varphi_L(m(g) + E[G], \sigma^2) dL \approx \\ &\approx \frac{1}{1 + \exp\left\{\frac{\pi(T - (m(g) + E[G]))}{\sqrt{3}\sigma^2}\right\}}, \end{aligned} \quad (2.6)$$

где σ^2 определена формулой 2.3.

При оценивании параметров могут быть зафиксированы пороговое значение T и сумма средовой и полигенной дисперсий в одной из популяций.

Ранее, для комплексного сегрегационного анализа бинарных признаков в гибридных родословных можно было применять только моногенную модель, которая предполагает отсутствие остаточной наследуемой компоненты. Следствием этого предположения является то, что пенетрантности майоргенотипов одинаковы для всех

членов родословной независимо от их происхождения и гибридного статуса. Такая модель реализована, например, в пакете программ SALP.

Описанные нами пороговые модификации смешанной и майоргенной моделей позволяют увеличить мощность сегрегационного анализа бинарных признаков в гибридных родословных и повысить шансы на нахождение майоргена.

Эффект гетерозиса

Хорошо известно, что у гибридных особей может наблюдаться эффект гетерозиса (Mather, Jinks, 1982; Lynch, Walsh, 1998). В рамках смешанной модели эффект гетерозиса h можно формализовать, предположив, что значение признака равно сумме эффектов майоргена, полигена, среды и гетерозиса: $x = m(g) + G + h + e$. Эффект гетерозиса предполагается пропорциональным гетерозиготности, возникающей при гибридизации (Mather, Jinks, 1982). Поэтому математическое ожидание эффекта гетерозиса

$$h = \eta \cdot (\epsilon_f(1 - \epsilon_m) + \epsilon_m(1 - \epsilon_f)) , \quad (2.7)$$

где η — максимальное изменение признака за счет гетерозиса, возникающее у гибридов F_1 . Из формулы 2.7 видно, что $h = 0$ для P_1 и P_2 , $h = 1/2\eta$ для F_2 и т.д. Этот вид формализации эффекта гетерозиса сходен с тем, что был введен Крессом (Cress, 1966).

В рамках смешанной модели, предполагающей эффект гетерозиса, функция пенетрантности описывается плотностью нормального распределения $\varphi_x(m(g) + G + h, \sigma_e^2)$. В случае майоргенной модели точное значение эффекта полигенотипа в функции пенетрантности заменяется на его математическое ожидание $E[G]$.

Введение эффекта гетерозиса в пороговую модификацию обеих моделей также не представляет труда.

2.2 Модели наследования совместных признаков

Стандартные модели комплексного сегрегационного анализа предполагают, что признак является характеристикой отдельной особи и определяется только ее генотипом.

В то же время существуют признаки, которые могут зависеть от генотипов пары особей: плодовитость некоторых млекопитающих (Penrose, 1949; Kasser et al., 1986; Dिल्ts et al., 1991), гемолитическая желтуха новорожденных (Zimmerman, 1973) и т.д. Назовем такие признаки совместными.

При формировании совместных признаков генотипы пары особей взаимодействуют опосредованно, через индивидуальные фенотипы — например, гемолитическая желтуха новорожденных возникает в следствии реакции иммунной системы матери на *Rh*-антиген плода, когда *Rh*-отрицательная мать вынашивает *Rh*-положительного ребенка (Фогель, Мотульски, 1990).

В идеале, основы генетической детерминации совместных признаков следует изучать в два этапа: во-первых, выделять и изучать взаимодействие индивидуальных признаков, которые детерминируют совместный признак и, во-вторых, изучать генетику этих индивидуальных признаков. Однако, выделение и изучение таких индивидуальных признаков может быть очень сложным, в то время как совместный признак легко наблюдаем.

Модель наследования, которую мы описываем в этой главе, может помочь при изучении генетики совместных признаков. Она позволяет, по меньшей мере, показать, следует ли рассматривать изучаемый признак как действительно совместный, или этот признак определен, в основном, генотипом только одной особи.

Формализация совместной модели

Относительно контроля совместного признака мы делаем стандартное предположение о том, что значение признака модифицируется набором наследственных факторов (среди которых выделяется менделевский ген (гены) g и остаточная наследственная компонента \mathcal{N}) и средой. Для подсчета функции правдоподобия родословной требуется определить три распределения вероятностей — априорное распределение вероятностей генотипов в популяции, распределение генотипов потомков и вероятность набора фенотипов всех членов родословной при условии их генотипов.

Два первых распределения задаются стандартным образом, как и для всех опи-

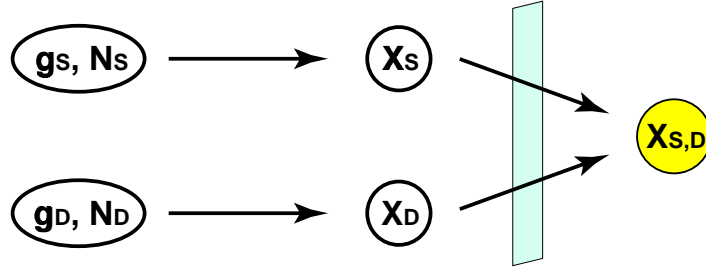


Рис. 2.1: Формирование совместного признака

санных выше моделей. Третье распределение не может быть стандартным, поскольку совместный признак определен для пары особей. Поэтому функция пенетрантности подвергается модификации — она должна формализовать правило построения совместного фенотипа пары особей на основе наследственных факторов этих двух особей.

Мы предполагаем, что взаимодействие генотипов пары особей происходит на уровне ненаблюдаемых признаков x_D и x_S , взаимодействие которых определяет наблюдаемое значение совместного признака $x_{S,D}$ (рис. 2.1). В общем случае, каждый из индивидуальных признаков x_D и x_S может иметь свою систему генетического контроля.

Пусть генотип, контролирующий ненаблюдаемый признак x_D , обозначается как (g_D, \mathcal{N}_D) , а x_S — (g_S, \mathcal{N}_S) . В общем случае, наборы генов в генотипах (g_D, \mathcal{N}_D) и (g_S, \mathcal{N}_S) могут быть разными. Тогда правило формирования наблюдаемого совместного признака на основе генотипов пары особей можно формализовать с помощью функции пенетрантности вида

$$Pr_{\vec{\theta}}(x_{S,D} | g_S, \mathcal{N}_S, g_D, \mathcal{N}_D)$$

— вероятность того, что значение признака равно $x_{S,D}$ при условии майоргенотипов и остаточных наследуемых факторов пары особей.

Функция правдоподобия может быть посчитана по формуле

$$\begin{aligned}
 L(\vec{x} | \mathcal{S}, \vec{\theta}) = & \log_e \left(\sum_{g_1} \int_{\mathcal{N}_1} \dots \sum_{g_N} \int_{\mathcal{N}_N} \left[\prod_{i \in C} Pr_{\vec{\theta}}(x_i | g_{S_i}, \mathcal{N}_{S_i}, g_{D_i}, \mathcal{N}_{D_i}) \cdot \right. \right. \\
 & \cdot \prod_{j \in P} Pr_{\vec{\theta}}(g_{S_j}, \mathcal{N}_{S_j}, g_{D_j}, \mathcal{N}_{D_j}) \cdot \\
 & \cdot \left. \prod_{k \in F} Pr_{\vec{\theta}}(g_{S_k}, \mathcal{N}_{S_k}, g_{D_k}, \mathcal{N}_{D_k} | g_{S_{m_k}}, \mathcal{N}_{S_{m_k}}, g_{D_{m_k}}, \mathcal{N}_{D_{m_k}}, g_{S_{f_k}}, \mathcal{N}_{S_{f_k}}, g_{D_{f_k}}, \mathcal{N}_{D_{f_k}}) \right] \\
 & \left. d\mathcal{N}_{S_1} d\mathcal{N}_{D_1} \dots d\mathcal{N}_{S_N} d\mathcal{N}_{D_N} \right),
 \end{aligned}
 \tag{2.8}$$

где C — множество пар особей, для которых определено значение совместного признака. Например, в случае гемолитической желтухи новорожденных это множество пар мать-потомок, а при анализе плодовитости — все пары родителей.

Выдвигая различные предположения о форме функции пенетрантности и конкретизируя вид остаточной компоненты \mathcal{N} , можно формализовать различные модели наследования совместных признаков.

Майоргенная модель наследования совместного признака

Предположим, что наследование как x_D , так и x_S можно описать в рамках майоргенной модели контроля индивидуального признака. Тогда значение совместного признака определяется как $x_{S,D} = m(g_S, g_D) + e$, где $m(g_S, g_D)$ — среднее значение признака при условии, что генотипы пары особей g_S и g_D , а e — случайная модификация, обусловленная влиянием среды (как внешней, так и внутренней). Для параметризации модели также требуется определить два набора частот аллелей и переходных вероятностей.

Модель формализуется с помощью параметров

$$\vec{\theta} = \{p_S(S_i), p_D(D_i), \tau(S_i | g_S), \tau(D_i | g_D), m(g_S, g_D), \sigma^2\}.$$

В том случае, если g_S и g_D являются не сцепленными менделевскими диаллельными генами, $m(g_S, g_D)$ описывается девятью параметрами, априорное распределение частот генотипов при справедливости распределения Харди-Вайнберга описывается двумя параметрами, а наследование майоргенов в родословной — двумя наборами переходных вероятностей.

Функция правдоподобия родословной записывается как

$$L(\vec{x} | \vec{\theta}, \mathcal{S}) = \sum_{g_1} \dots \sum_{g_N} \prod_{i \in C} \varphi_{x_i}(m(g_{S_i}, g_{D_i}), \sigma^2) \cdot \prod_{j \in P} Pr(g_{S_j}) \cdot Pr(g_{D_j}) \prod_{k \in F} Pr(g_{S_k} | g_{S_{m_k}}, g_{S_{f_k}}) \cdot Pr(g_{D_k} | g_{D_{m_k}}, g_{D_{f_k}}), \quad (2.9)$$

где C — множество пар особей, для которых определен совместный признак.

Иногда индивидуальные признаки x_D и x_S могут иметь общую систему контроля (как это происходит, например, в случае гемолитической желтухи новорожденных). В этом случае как параметрическое пространство, так и функция правдоподобия

упрощаются: модель описывается с помощью вектора параметров

$$\vec{\theta} = \{p(A_i), \tau(A_i|g), m(g_S, g_D), \sigma^2\},$$

а функция правдоподобия равна

$$\begin{aligned} L(\vec{x}|\vec{\theta}, \mathcal{S}) = & \sum_{g_1} \dots \sum_{g_N} \prod_{i \in C} \varphi_{x_i}(m(g_{S_i}, g_{D_i}), \sigma^2) \cdot \\ & \cdot \prod_{j \in P} Pr(g_j) \prod_{k \in F} Pr(g_k | g_{m_k}, g_{f_k}). \end{aligned} \quad (2.10)$$

2.3 Обсуждение

В этой главе мы описали модели наследования количественных и бинарных признаков в гибридных родословных, а также новый класс моделей комплексного сегрегационного анализа — совместные модели.

Модели наследования признаков в гибридных родословных

Родословные, полученные при гибридизации представителей различных пород или популяций, могут быть чрезвычайно информативными с точки зрения сегрегационного анализа. Высокая информативность таких родословных обусловлена возможным повышением генетического полиморфизма, происходящем при гибридизации.

Описание наследования и сегрегационный анализ признаков в гибридных родословных может быть проведен двумя способами. Первый способ заключается в следующем: части родословной, сформированные определенным типом родителей (P_1 и P_2) или полученные при скрещивании представителей одного и того же типа гибридов (например, $F_1 \times F_1$), формально можно рассматривать как отдельные родословные, каждая из которых получена из некоторой популяции. Для каждой из таких родословных может быть проведен свой сегрегационный анализ. Такой способ позволяет оценить модель наследования в каждой рассматриваемой группе, однако не позволяет рассматривать гибридную родословную как единое целое; статистически обоснованные выводы о модели контроля признака можно делать только в рамках каждой из рассмотренных групп, но не всей гибридной родословной.

Второй способ описания наследования признаков в гибридных родословных предполагает единую модель контроля признака у всех членов родословной. В рамках этой модели можно провести сегрегационный анализ, рассматривая родословную как единое целое и используя информацию о всех особях в родословной.

Первый способ не позволяет интегрировать результаты анализа и не использует информацию о некоторых членах родословной (F_1 , потомки бэккроссов). Поэтому второй способ является предпочтительным с точки зрения информативности сегрегационного анализа.

В рамках второго способа были разработаны новые модели комплексного сегрегационного анализа. Модели предполагали, что возможная разница в значении признака у представителей двух исходных популяций и разных типов гибридов объясняется только различным распределением наследственных факторов. Правила наследования генов, а также правило реализации признака, предполагались едиными для всех членов родословной. В рамках этих предположений была описана майоргенная модель наследования и пороговые модификации смешанной и майоргенной моделей; дана одна из возможных формализаций эффекта гетерозиса.

По сравнению с родословными, полученными из одной популяции, при описании наследования признака в гибридной родословной требуется ввести дополнительные параметры: набор частот майоргенотипов и дисперсию полигенотипа во второй популяции, а также параметры Δ и γ^2 . Известно, что чем больше число параметров модели, тем выше должна быть информативность данных, используемых для их оценивания. Ожидается, что гибридизация представителей различных пород или популяций животных повысит генетическую информативность родословных и обеспечит тем самым возможность оценки дополнительных параметров.

Важно отметить, что модификации, необходимые для описания модели наследования в гибридных родословных, не приводят к изменению формы представления функции правдоподобия. Это значит, что для анализа гибридных родословных можно использовать существующие алгоритмы точного или приближенного подсчета функции правдоподобия (Lang, Elston, 1975; Cannings et al., 1978; Lang, Boehnke, 1983). Реализация смешанной модели наследования признака в гибридной родословной в рамках какого-либо существующего пакета требует введения дополнительной информации о популяционном происхождении каждой особи, а также введения дополнительных параметров в модель, используемую при анализе. Все это возможно для пакетов, которые имеют открытый код и позволяют проводить анализ с дополнительными переменными (MENDEL, PAP, MAN).

Предложенный метод открывает новые возможности для тестирования значимости майоргенного эффекта в детерминации межпопуляционных или межпородных

различий количественного признака.

Модели наследования совместных признаков

Нами была описана формализация модели наследования совместного признака. Насколько нам известно, ранее проблема анализа совместных признаков систематически не рассматривалась.

В принципе, анализ совместных признаков может проводиться и с помощью стандартных моделей наследования. Например, влияние генотипа самки и самца на число новорожденных можно оценить, проведя два стандартных анализа: в первом анализе признак "число новорожденных" приписывается самке, а во втором — самцу. При определенных условиях¹ реальная модель контроля признака может быть установлена. Однако, если эффекты обоих генотипов и их взаимодействия велики, то скорее всего подобный метод анализа не позволит оценить реальную модель наследования. Следует отметить, что в рамках такого анализа сравнение гипотез об отцовском и материнском контроле невозможно: они не являются иерархическими, и построение критерия значимости невозможно.

Совместная модель наследования включает модели индивидуального контроля как частные случаи. Поэтому она позволяет статистически сравнивать гипотезу о совместном контроле признака генотипами пары особей с гипотезами о контроле признака генотипом только одной особи.

Если признак действительно контролируется генотипами пары особей, введение совместной модели наследования позволит повысить мощность генетического анализа, а в некоторых случаях получить принципиально новые результаты: оценить эффекты взаимодействия и относительный вклад генотипов пары при формировании совместного признака.

Реализация совместной модели требует принципиального изменения вида функции пенетрантности, и, как следствие, всей структуры представления функции правдоподобия и данных. Поэтому совместную модель нельзя реализовать в рамках уже существующих пакетов програм.

¹Например, если вклад одного из генов в детерминацию популяционного полиморфизма признака мал по сравнению с вкладом второго, или гены действуют аддитивно при контроле признака.

Разработанные модели реализованы в виде пакета компьютерных программ для комплексного сегрегационного анализа (МАІА), см. приложение В. Пакет позволяет анализировать наследование количественных и бинарных признаков в гибридных родословных. Кроме того, он позволяет проводить анализ наследования некоторых совместных признаков.

Глава 3

Разработка алгоритмов: анализ родословных с множественными петлями

В настоящее время не существует эффективных алгоритмов точного подсчета функции правдоподобия больших родословных с множественными петлями (раздел 1.3). Поэтому используются методы приближенного подсчета функции правдоподобия. Одним из них является метод разрыва петель. Метод обладает рядом привлекательных свойств: относительно проста его реализация, возможен быстрый подсчет функции правдоподобия для сколь угодно сложных родословных. Однако, свойства этого метода изучены слабо. Теоретический анализ предсказывает, что точность аппроксимации может быть увеличена, если разработать алгоритм, сохраняющий короткие инбредные петли.

Целью исследования, представленного в данной главе, является разработка такого алгоритма и сравнение его свойств со свойствами ранее предложенных алгоритмов.

3.1 Алгоритм подсчета функции правдоподобия родословной без петель

В основе алгоритма подсчета функции правдоподобия больших родословных лежит предположение о том, что при фиксированных генотипах шарнирных особей вероятности распределения фенотипов в двух частях родословной, разделенных ими, неза-

висимы. Это предположение позволяет производить подсчет вероятности родословной по частям.

Эффективный алгоритм подсчета функции правдоподобия родословной без петель основывается на последовательной свертке терминальных ядерных родословных. При каждом шаге свертки информация о членах ядерной родословной переносится на шарнирную особь, вследствие чего родословная постепенно редуцируется. При этом шарнирная особь меняет свой статус: если она была потомком в свернутой ядерной родословной, то в редуцированной родословной она становится родоначальником; если она была родителем в свернутой ядерной родословной, то в редуцированной родословной она может стать свободным потомком. Как видно, время подсчета функции правдоподобия родословной пропорционально числу ядерных родословных, т.е. растет примерно линейно с увеличением размера родословной.

Результатом свертки части родословной на шарнирную особь i является функция $C^{(S_i)}(\mathcal{G}_i)$, концентрирующая информацию о всех членах родословной, свернутых на i -ую особь за s шагов:

$$C^{(S_i)}(\mathcal{G}_i) = \begin{cases} Pr(\vec{x}_i^{(S_i)}, \mathcal{G}_i), & \text{если } i\text{-ая особь — родоначальник} \\ Pr(\vec{x}_i^{(S_i)} | \mathcal{G}_i), & \text{иначе.} \end{cases}$$

Здесь $\vec{x}_i^{(S_i)}$ — вектора фенотипов особи i и особей, свернутых на нее за s шагов.

При $s = 0$,

$$C^{(0)}(\mathcal{G}_i) = \begin{cases} Pr(x_i, \mathcal{G}_i) = Pr(x_i | \mathcal{G}_i) \cdot Pr(\mathcal{G}_i), & \text{если } i\text{-ая особь — родоначальник} \\ Pr(x_i | \mathcal{G}_i), & \text{иначе.} \end{cases}$$

Для подсчета функции $C^{(S_i)}(\mathcal{G}_i)$ используются две процедуры свертки:

1. Свертка ядерной родословной на одного из родителей:

$$C^{(S_1+1)}(\mathcal{G}_1) = C^{(S_1)}(\mathcal{G}_1) \sum_{\mathcal{G}_2} C^{(S_2)}(\mathcal{G}_2) \prod_{j=3}^n \sum_{\mathcal{G}_j} C^{(S_j)}(\mathcal{G}_j) \cdot Pr(\mathcal{G}_j | \mathcal{G}_1, \mathcal{G}_2) \quad (3.1)$$

здесь индексами 1 и 2 обозначены родители ядерной родословной, причем шарнирной особью является родитель 1; n — число членов ядерной родословной (см. рис. 3.1).

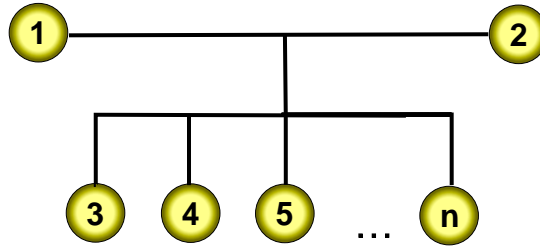


Рис. 3.1: Ядерная родословная

2. Свертка ядерной родословной на одного из потомков:

$$C^{(S_3+1)}(\mathcal{G}_3) = C^{(S_3)}(\mathcal{G}_3) \sum_{\mathcal{G}_1} C^{(S_1)}(\mathcal{G}_1) \sum_{\mathcal{G}_2} C^{(S_2)}(\mathcal{G}_2) \cdot Pr(\mathcal{G}_3|\mathcal{G}_1, \mathcal{G}_2) \prod_{j=4}^n \sum_{\mathcal{G}_j} C^{(S_j)}(\mathcal{G}_j) \cdot Pr(\mathcal{G}_j|\mathcal{G}_1, \mathcal{G}_2), \quad (3.2)$$

здесь шарнирной особью является потомок 3.

3.2 Свертка коротких инбредных петель

Для того, чтобы уменьшить потерю генетической информации при аппроксимации функции правдоподобия, мы ввели дополнительную процедуру свертки. Эта свертка может использоваться, если родословная, максимально редуцированная процедурами (3.1) и (3.2), содержит определенные структуры, схематически изображенные на рис. 3.2. Как видно, эта структура состоит из нескольких ядерных родословных, среди которых выделяется *первичная* (на рис. 3.2 — это ядерная родословная с родителями 1 и 2).

Основная характеристика рассматриваемой структуры заключается в том, что все потомки первичной ядерной родословной либо являются свободными (особь 4), либо скрещиваются только с членами этой ядерной родословной, вступая в брат-сестринские (особи 5 и 6) или в возвратные (особи 3 и 7) скрещивания. Кроме того, необходимо, чтобы только один из родителей первичной ядерной родословной был шарнирной особью для сворачиваемой части родословной (особь 1).

Общая формула свертки коротких инбредных петель имеет вид

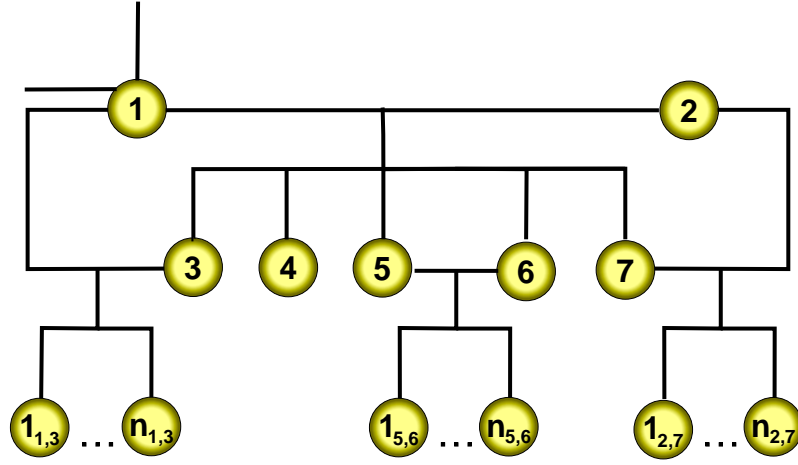


Рис. 3.2: Пример части родословной, которая может быть свернута процедурой (3.3).

$i_{j,k}$ обозначает i -го потомка от скрещивания особей j и k .

$$C^{(S_1+1)}(\mathcal{G}_1) = C^{(S_1)}(\mathcal{G}_1) \sum_{\mathcal{G}_2} C^{(S_2)}(\mathcal{G}_2) \prod_{i=2}^m \mathcal{A}_i(\mathcal{G}_1, \mathcal{G}_2), \quad (3.3)$$

здесь m — число ядерных родословных, входящих в состав сворачиваемой части ($m = 4$ для примера, приведенного на рис. 3.2); \mathcal{G}_1 и \mathcal{G}_2 — генотипы родителей первичной ядерной родословной. \mathcal{A} -функция является специальным видом \mathcal{R} -функции Каннингса (Cannings et al., 1978). Она описывает результат свертки i -ой ядерной родословной на родителей первичной ядерной родословной. Мы вводим три типа \mathcal{A} -функции:

1. \mathcal{A} -функция свободного потомка:

$$\mathcal{A}_j(\mathcal{G}_1, \mathcal{G}_2) = \sum_{\mathcal{G}_j} C^{(S_j)}(\mathcal{G}_j) Pr(\mathcal{G}_j | \mathcal{G}_1, \mathcal{G}_2),$$

здесь j — номер свободного потомка первичной ядерной родословной ($j = 4$ на рис. 3.2).

2. \mathcal{A} -функция бэкрасса:

$$\mathcal{A}_{p,f}(\mathcal{G}_1, \mathcal{G}_2) = \sum_{\mathcal{G}_f} C^{(S_f)}(\mathcal{G}_f) Pr(\mathcal{G}_f | \mathcal{G}_1, \mathcal{G}_2) \prod_{j=1_{p,f}}^{n_{p,f}} \sum_{\mathcal{G}_j} C^{(S_j)}(\mathcal{G}_j) Pr(\mathcal{G}_j | \mathcal{G}_1, \mathcal{G}_2),$$

здесь f — номер потомка первичной ядерной родословной, участвующего в возвратном скрещивании со своим родителем, имеющим номер p . Для одного из бэкроссов, представленных на рис. 3.2, $p = 1$ и $f = 3$, а для другого — $p = 2$ и $f = 7$.

3. \mathcal{A} -функция брат-сестринского скрещивания:

$$\mathcal{A}_{f_1, f_2}(\mathcal{G}_1, \mathcal{G}_2) = \sum_{\mathcal{G}_{f_1}} C^{(S_{f_1})}(\mathcal{G}_{f_1}) Pr(\mathcal{G}_{f_1} | \mathcal{G}_1, \mathcal{G}_2) \sum_{\mathcal{G}_{f_2}} C^{(S_{f_2})}(\mathcal{G}_{f_2}) Pr(\mathcal{G}_{f_2} | \mathcal{G}_1, \mathcal{G}_2) \prod_{j=1, f_1, f_2}^{n_{f_1, f_2}} \sum_{\mathcal{G}_j} C^{(S_j)}(\mathcal{G}_j) Pr(\mathcal{G}_j | \mathcal{G}_{f_1}, \mathcal{G}_{f_2}),$$

здесь f_1 и f_2 — номера потомков первичной ядерной родословной, вступающих в брат-сестринское скрещивание ($f_1 = 5$ и $f_2 = 6$ на рис. 2).

Легко убедиться, что в том случае, когда все потомки первичной ядерной родословной свободные, свертка (3.3) идентична свертке (3.1). Как видно, процедура (3.3) сворачивает часть родословной на единственную шарнирную особь, как и процедуры (3.1) и (3.2), применяемые для эффективного подсчета функции правдоподобия родословной без петель. Это позволяет легко ввести свертку (3.3) в существующие пакеты программ. Относительная трудоемкость свертки (3.3) немногим больше трудоемкости стандартных сверток (3.1) и (3.2).

3.3 Точность генетического анализа при различных аппроксимациях функции правдоподобия

Для сравнения точности генетического анализа при различных способах аппроксимации функции правдоподобия использовали генетико-стохастический эксперимент.

Структура родословной, в которой моделировалось распределение признака, была идентична структуре реальной родословной свиней, полученной в экспериментальном хозяйстве Института цитологии и генетики СО РАН (см. раздел 4, рис. 4.2 на стр. 63). Эта родословная включала 568 особей, образующих 60 ядерных родословных; она содержала много петель, в том числе и коротких инбредных. Некоторые особи участвовали во многих скрещиваниях — например, один из самцов скрещивался с двадцатью четырьмя самками. Только три ядерные родословные могли быть свернуты с помощью стандартных сверток (3.1) и (3.2).

При моделировании альтернативного признака предполагалось, что он контролируется аутосомным диаллельным геном, для которого в популяции выполняется закон Харди-Вайнберга. Параметры генетической модели были следующими: $p(a) = 0.5$, $w(AA) = 0.0$, $w(Aa) = 0.5$ и $w(aa) = 1.0$, где $p(a)$ — популяционная частота аллеля a , $w(g)$ — пенетрантность генотипа g , $g \in \{AA, Aa, aa\}$. Моделирование распределения признака в родословной и последующий сегрегационный анализ повторяли 500 раз.

Подсчет функции правдоподобия осуществлялся с помощью модифицированного пакета компьютерных программ МАН-А1 (Аксенович, Гинзбург, 1987), в который была введена процедура свертки (3.3).

Мы сравнивали следующие способы аппроксимации функции правдоподобия:

- (i) Аппроксимация с помощью разрыва всех петель: петлю разрывали путем дублирования некоторой особи, при этом копия имела тот же фенотип, что и оригинал (Stricker et al., 1995). В результате дублирования число особей в родословной возросло до 608.
- (ii) Условная аппроксимация (Wang et al., 1996): петли разрывали как в (i), но функцию правдоподобия считали при условии фенотипов дублированных особей, т.е. как разность логарифмов вероятностей распределения фенотипов в расширенной родословной и фенотипов дублированных особей.
- (iii) Условная аппроксимация с дополнительным расширением. Дополнительное расширение родословной осуществляли в соответствии со стратегией, предложенной Вангом и др. (1996). Мы использовали относительно простой вариант метода, заключающийся в следующем: если особь, по которой разрывалась петля, была родоначальником, она просто дублировалась, иначе дублировались также сибь и родители этой особи (см. рис. 1.3Б). Расширенная таким образом родословная состояла из 754 особей и 74 ядерных родословных. Четырнадцать ядерных родословных появилось в процессе дополнительного расширения. Функция правдоподобия считалась как и в (ii) при условии фенотипов дублированных особей.

Варианты (i)–(iii) предполагают разрыв всех петель родословной, поэтому для подсчета функции правдоподобия достаточно сверток (3.1) и (3.2). Следующие варианты

алгоритма использовали дополнительную свертку (3.3), которая позволяет обрабатывать короткие инбредные петли, не разрывая их:

- (iv) Аппроксимация функции правдоподобия с сохранением коротких петель. Разрывали только длинные петли, используя такой же способ как в (i). Распиренная таким образом родословная состояла из 587 особей. Функцию правдоподобия считали как в (i).
- (v) Условная аппроксимация с сохранением коротких петель. Родословную расширяли как в (iv), функцию правдоподобия считали при условии фенотипов дублированных особей как в (ii).

Результаты, полученные с помощью пяти аппроксимаций, сравнивали с результатами, которые давал точный подсчет. Точное значение функции правдоподобия вычисляли с помощью трудоемкого алгоритма, основанного на создании "генетических копий" (Ott, 1976, см. также обзор литературы). Время счета при использовании этого алгоритма растет экспоненциально с увеличением числа копируемых особей.

Свойства различных аппроксимаций характеризовали точностью оценки генетических параметров и правильностью тестирования генетических гипотез. Точность оценки каждого из генетических параметров измеряли двумя величинами: средним значением оценки параметра, полученным в 500 повторностях, и стандартным отклонением среднего. Точность тестирования гипотез характеризовали частотой ложного отвержения менделевского наследования модельного признака. Для проверки менделизма применяли критерий Эльстона-Стюарта, основанный на сравнении гипотез, предполагающих:

- (A) менделевскую сегрегацию аллелей предполагаемого гена,
- (B) произвольную сегрегацию аллелей,
- (C) отсутствие наследования.

Менделевская модель не отвергается, если гипотеза (B) статистически не отличается от (A) и значимо лучше гипотезы (C) (Elston, Stewart, 1971, см. также обзор литературы). Тестирование гипотез проводили при уровне значимости $\alpha = 0.05$. Число степеней свободы при сравнении гипотез (A) и (B) варьирует между $df_{max} = 3$ и

Таблица 3.1: Средние значения оценок параметров и стандартное отклонение среднего при различных способах подсчета функции правдоподобия

СПОСОБ ПОДСЧЕТА ФП	ПАРАМЕТР			
	$p(a)$	$w(AA)$	$w(Aa)$	$w(aa)$
Точный	0.507 ± 0.095	0.022 ± 0.055	0.496 ± 0.066	0.978 ± 0.055
(i) разрыв всех петель	0.506 ± 0.153	0.048 ± 0.089	0.493 ± 0.097	0.954 ± 0.093
(ii) разрыв всех петель и условный подсчет ФП	0.507 ± 0.137	0.041 ± 0.082	0.494 ± 0.096	0.959 ± 0.084
(iii) дополнительное расширение и условный подсчет ФП	0.509 ± 0.138	0.039 ± 0.080	0.493 ± 0.087	0.958 ± 0.084
(iv) сохранение коротких петель	0.509 ± 0.108	0.028 ± 0.059	0.497 ± 0.074	0.973 ± 0.064
(v) сохранение коротких петель, условный подсчет ФП	0.508 ± 0.104	0.028 ± 0.060	0.496 ± 0.075	0.972 ± 0.065

$df_{min} = [3 - \text{число переходных вероятностей, оценки которых равны граничному значению параметра}]$ (Self, Liang, 1987).

В таблице показаны средние значения оценок генетических параметров. Как видно, все они незначительно отклоняются от модельных значений. В то же время стандартное отклонение существенно зависит от способа аппроксимации: при аппроксимации (i) стандартное отклонение максимально; условный подсчет функции правдоподобия (ii) и дополнительное расширение (iii) несколько уменьшают разброс оценок параметров, а аппроксимации, сохраняющие короткие петли (iv и v), снижают стандартное отклонение почти до уровня, получаемого при точном подсчете функции правдоподобия (см. таблицу).

На рис. 3.3 показаны ошибки I рода критерия Эльстона-Стюарта, определенные как частоты ложного отвержения менделевского наследования. При точном подсчете функции правдоподобия максимальная ошибка I рода несколько меньше α . При аппроксимации (i) она значительно превышает α и составляет 0.34. Условный подсчет функции правдоподобия (ii) и дополнительное расширение (iii) несколько исправляет ситуацию, но ошибки I рода по-прежнему в несколько раз больше α . Эти результаты показывают, что аппроксимация с помощью разрыва всех петель родословной может привести к ложному отвержению гипотезы о менделевском наследовании признака. В то же время, при аппроксимациях, сохраняющих короткие петли, ошибки I рода

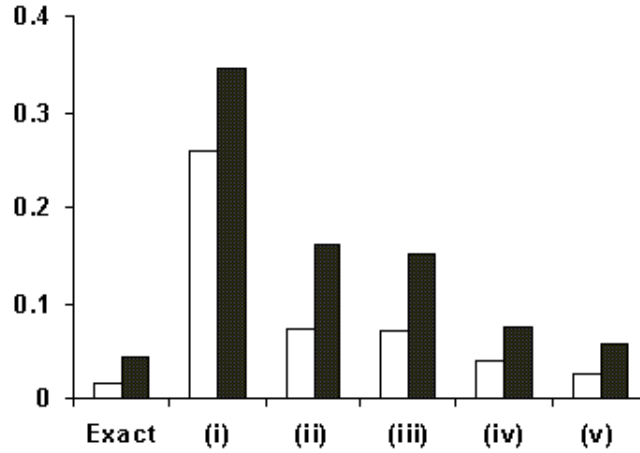


Рис. 3.3: Частота ложного отвержения менделевского наследования при различных способах подсчета функции правдоподобия. (i)–(v) обозначают то же самое, что и в таблице. Белые столбцы — тестирование при числе степеней свободы, равном df_{max} . Черные — при df_{min} .

близки к α и составляют 0.08 и 0.06 для аппроксимаций (iv) и (v), соответственно. Следовательно, ложное отвержение менделевского наследования возникает, в основном, за счет разрыва коротких инбредных петель, и использование дополнительной свертки (3.3) позволяет существенно повысить корректность генетического анализа.

3.4 Обсуждение

Короткие инбредные петли, возникающие в результате бэккроссов и брат-сестринских скрещиваний, часты в родословных сельскохозяйственных и лабораторных животных. Разрыв таких петель, как мы показали, может значительно повысить риск ложного отвержения гипотезы о менделевском наследовании признака. Причиной этого является изменение структуры анализируемой родословной, возникающее в результате разрыва петель. Это изменение сопровождается искажением распределения вероятностей генотипов дублированных особей и особей, связанных с ними близким родством. Степень искажения тем больше, чем меньше размер разрываемой петли и выше степень инбредности, т.е. именно при разрыве коротких инбредных петель происходит максимальное искажение совместного распределения генотипов в родословной. Следствием последнего может быть смещение оценок генетических па-

раметров.

Мы показали, что средние оценки параметров, полученные в 500 повторностях, хорошо совпадают с моделируемыми значениями при всех способах аппроксимации функции правдоподобия. Однако при разрыве всех петель повышается стандартное отклонение. Это значит, что в каждом конкретном случае, оценки параметров могут быть значительно смещены в ту или иную сторону. Как известно, смещение оценок генетических параметров может явиться причиной ложного отвержения менделизма. Результаты нашей работы подтверждают это: для шести рассмотренных способов подсчета функции правдоподобия наблюдается положительная корреляция между повышением стандартного отклонения и частотой ложного отвержения менделевской сегрегации генов.

Существуют стратегии, которые позволяют уменьшить искажение распределения генотипов, которое возникает в ходе разрыва петель. Наиболее эффективной из них является стратегия дополнительного расширения и условного подсчета функции правдоподобия (iii), поскольку она позволяет реконструировать родословную вблизи точки разрыва и, следовательно, частично восстановить индивидуальное апостериорное распределение генотипов. Однако результаты нашего анализа показывают, что если разрываются все петли родословной, применение этой стратегии не дает существенного улучшения точности оценок генетических параметров и уменьшения ошибок при проверке менделевской сегрегации. Это может быть обусловлено двумя причинами. Во-первых, исследуемая нами родословная была насыщена короткими инбредными петлями, для которых даже полное восстановление распределения генотипов разорванной особи не может компенсировать последствий нарушения априорного совместного распределения вероятностей генотипов. Во-вторых, мы использовали достаточно простой вариант дополнительного расширения, при котором разрыв инбредной петли по некоторой особи сопровождался дублированием только ее родителей и сибов. Стратегия дополнительного расширения, вероятнее всего, наиболее эффективна в том случае, когда разрываются не слишком короткие петли.

Использование алгоритмов, сохраняющих короткие петли, снижает частоту ложного отвержения менделизма и обеспечивает практически ту же ошибку I рода, что и точный метод подсчета функции правдоподобия. Поэтому интегрированный подход, предполагающий дополнительное расширение при разрыве длинных и средних петель и прямую свертку структур, содержащих короткие петли, может оказаться

ценным при анализе наследования признаков по родословным сельскохозяйственных и лабораторных животных. Мы реализовали один из вариантов этого подхода в пакете компьютерных программ МАІА (см. приложение В).

Список литературы

1. Аксенович Т.И., Гинзбург Э.Х. (1987) Система для менделевского анализа альтернативных признаков (МАН-A1) // Генетика. Т.23. С. 268-273.
2. Аксенович Т.И. (1999) Описание наследования количественных признаков в гибридных родословных. Смешанные модели // Генетика. Т. 35. С. 530-539.
3. Балков И.Я. (1990) ЦМС сахарной свеклы. М.: Агропромиздат.
4. Бекенев В.А. (1997) Селекция свиней. Новосибирск: РАСХН. Сиб. отделение.
5. Вепрев С.Г., Дикалова А.Э., Мглинец А.В. и др. (1997) Изменение типа утиоплазмы у сахарной свеклы (*Beta Vulgaris* L.) при инбридинге. Гибридологический анализ и тестирование типа митохондриальной ДНК // Генетика. Т. 33. С. 934-942.
6. Вепрев С.Г., Дударева Н.А., Малецкий С.И., Мглинец А.В. (1991) Женская двудомность: генетические системы контроля стерильности пыльцы // Генетический контроль размножения сахарной свеклы. Новосибирск: Наука. С. 114-166.
7. Гинзбург Э. Х. (1984) Описание наследования количественных признаков. Новосибирск: Наука.
8. Жимулев И.Ф. (1998) Общая и молекулярная генетика (курс лекций для студентов 3 курса, версия 1998-4). <http://www.nsu.ru/biology/courses/genetics/index.html>
9. Инге-Вечтомов С.Г. (1989) Генетика с основами селекции. Москва: Высшая школа.
10. Кендалл М.Д., Стьюарт А. (1973) Статистические выводы и связи. Москва: Наука.

11. Кольчик О.А., Душкин В.А., Горбунова Н.А. (1974) Миниатюрная свинья в медико-биологических исследованиях // Лабораторные животные в медицинских исследованиях. Москва: АМН СССР. С. 188-190.
12. Малецкий С.И., Коновалов А.А. (1991) Внутривидовая само- и перекрестная несовместимость // Генетический контроль размножения сахарной свеклы. Новосибирск: Наука. С. 15-49.
13. Познахирев П.Р. (1979) Мини-свиньи, как объект биомедицинских экспериментов на животных. Москва: АМН СССР. С. 32-40.
14. Серебровский А.С. (1970) Генетический анализ. Москва: Наука.
15. Сэджер Р. (1975) Цитоплазматические гены и органеллы. Москва: Мир.
16. Тихомирова М.М. (1990) Генетический анализ. Ленинград: ЛГУ.
17. Тихонов В.Н. (1990) Генетика мини-свиней. Новосибирск: ИЦиГ СО АН СССР.
18. Феллер В. (1984) Введение в теорию вероятностей и ее применение в 2-х т. Москва: Мир.
19. Фогель Ф., Мотульски А. (1990) Генетика человека. Проблемы и подходы в 3-х т. Москва: Мир.
20. Хагеман, Р. (1962) Плазматическая наследственность. Москва: Мир.
21. Akaike H. (1974) A new look at the statistical model identification // IEEE Trans. Autom. Control. V. AC-19. P. 716-723.
22. Amos C.I. (1994) Robust variance-components approach for assessing genetic linkage in pedigrees // Am. J. Hum. Genet. V. 54. P. 535-543.
23. Bergsma D.R., Brown K.S. (1971) White fur, blue eyes, and deafness in the domestic cat // J. Hered. V. 62. P. 171-185.
24. Bliss F., Gabelman W. (1965) Inheritance of male sterility in beets // Crop Sci. V. 5. P. 403-406.
25. Bodmer W.F. (1965) Differential fertility in population genetic models // Genetics. V. 51. P. 411-424.

26. Bonney G.E. (1984) On the statistical determination of major gene mechanisms in continuous human traits: regressive models // *Am. J. Med. Genet.* V. 18. P. 731-749.
27. Borecki I.B., Rao D.C., Third J.L., Laskarzewski P.M., Glueck, C.J. (1986) A major gene for primary hypoalphalipoproteinemia // *Am. J. Hum. Genet.* V. 38. P. 373-381.
28. Cannings C., Thompson E.A., Skolnik M.H. (1978) Probability function on complex pedigrees // *Adv. Appl. Probab.* V. 10. P.26-61.
29. Clark A.G., Feldman M.W. (1986) A numerical simulation of the one-locus multiple-allele fertility model // *Genetics.* V. 113. P. 229-253.
30. Clementi M., Tenconi R., Forabosco P., Calzolari E., Milan M. (1997) Inheritance of cleft palate in Italy. Evidence for a major autosomal recessive locus // *Hum. Genet.* V. 100. P. 204-209.
31. Collins F.S. (1995) Positional cloning moves from perditional to traditional // *Nat. Genet.* V. 9. P. 347-350.
32. Cress C.E. (1966) Heterosis of the hybrid related to gene frequency differences between two populations // *Genetics.* V. 113. P. 269-274.
33. Demenais F., Lathrop M., Lalouel J.-M. (1986) Robutness and power of the unified model in the analysis of quantitative measurements // *Am. J. Hum. Genet.* V. 38. P. 228-234.
34. Dilts R. B., Famula T. R., Bradford G. E. (1991) Ovulation rate and pre- and postimplantation survival in mice with a major gene for rapid postweaning gain // *J. Anim. Sci.* V. 69. P. 3590-3596.
35. Dürr M. (1937) Die vererbung der glasäugigkeit beim schwein // *Zeitschrift für Zuchtung.* V. 37. P. 129-158.
36. Eisen E.J., Jonson B.H. (1981) Correlated responses in male reproductive traits in mice selected for litter size and body weight // *Heredity.* V. 99. P. 513-524.

37. Elston R.C., Stewart J. (1971) A general model for the genetic analysis of pedigree data // Hum. Hered. V. 21. P. 523-542.
38. Elston R.C., Stewart J. (1973) The analysis of quantitative traits for simple genetic models from parental, F_1 and backcross data // Genetics. V. 73. P. 675-693.
39. Elston R.C., Namboodiri, K.K., Nino, H.V., et al. (1974) Studies on blood and urine glucose in Seminole Indians: indications for segregation of a major gene // Am. J. Hum. Genet. V. 26. P. 13-34.
40. Elston R.C., Yelverton K.C. (1975) General model for segregation analysis // Am. J. Hum. Hered. V. 27. P. 31-45.
41. Elston R.C., Namboodiri K.K., Glueck C.G., et al. (1975) Study of the genetic transmission of hypercholesterolemia and hypertriglyceridemia in a 195 member kindred // Ann. Hum. Genet. V. 39. P. 67-87.
42. Elston R.C., Namboodiri K.K., Kaplan E.B. (1978) Resolution of major loci for quantitative traits // In Genetics Epidemiology. Eds. Morton N.E., Chung C.S.. New York: Academic Press. P. 223-240.
43. Elston R.C. (1981) Segregation analysis // Adv. Hum. Genet. V. 11. P. 63-63.
44. Elston R.C., George V.T., Severtson F. (1992) The Elston-Stewart algorithm for continuous genotypes and environmental factors // Hum. Hered. V. 42. P. 16-27.
45. Fain P.R. (1978) Characteristics of simple sibship variance tests for the detection of major loci and application to height, weight and spatial performance // Ann. Hum. Genet. V. 42. P. 109-120.
46. Falconer D.S. (1963) Quantitatively different responses to selection in opposite directions // In Statistical Genetics and Plant Breeding. Eds. Hanson W.D., Robinson H.F.. Washington: National Academy of Sciences - National Research Council. Publication N. 982. P. 487-490.
47. Falconer D.S. (1989) Introduction to Quantitative Genetics. Harlow: Logman Scientific and Technical.

48. Feldman M.W., Liberman U. (1985) A symmetric two-locus fertility model // Genetics. V. 109. P. 229-253.
49. Fisher R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance // Trans. Royal Soc. Edinburgh. V. 52. P. 399-433.
50. Frankel R., Galun E. (1978) Pollination mechanisms, reproduction and plant breeding. Berlin et al.: Springer-Verlag.
51. Garvil F. (1977) Some NP-complete problems on graphs // In Proceedings from the Conference on Information Sciences and Systems. John Hopkins.
52. Gelatt K.N., Rempel W.E., Makamera T.P., Anderson J.F. (1973) Heterochromia irides in miniature swine // J. Hered. V. 64. P. 343-347.
53. Ginsburg E., Livshits G. (1999) Segregation analysis of quantitative traits // Ann. Hum. Biol. V. 26. P. 103-129.
54. Grune P. (1976) Cytoplasmic genetics and evolution. Columbia Univ. Press.
55. Guo S.W., Thompson E.A. (1994) Monte Carlo estimation of mixed models for large complex pedigrees // Biometrics. V. 50. P. 417-432.
56. Hadeler K.P., Liberman U. (1975) Selection model with fertility differences // J. Math. Biol. V. 2. P. 19-32.
57. Hall J.M., Lee M.K., Newman B., et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21 // Science. V. 250. P. 1684-1689.
58. Hanson R.L., Knowler W.C. (1998) Analytic strategies to detect linkage to a common disorder with genetically determined age of onset: diabetes mellitus in Pima Indians // Genet. Epidemiol. V. 15. P. 299-315.
59. Hasstedt S.J. (1982) A mixed-model likelihood approximation on large pedigrees // Comput. Biomed. Res. V. 15. P. 295-307.
60. Hasstedt S.J. (1991) A variance components/major locus likelihood approximation on quantitative data // Genet. Epidemiol. V. 8. P. 113-125.

61. Hecht J.T., Yang P., Michels V.V., Buetow K.H. (1991) Complex segregation analysis of nonsyndromic cleft lip and palate // *Am. J. Hum. Genet.* V. 49. P. 674-681.
62. Hoeschele I. (1988) Statistical techniques for detection of major genes in animal breeding data // *Theor. Appl. Genet.* V. 76. P. 311-319.
63. Hogaboam G.J. (1957) Factors influencing phenotypic expression of cytoplasmic male sterility in the sugar beet, *Beta vulgaris* L. // *J. Amer. Soc. Sugar Beet Technol.* V. 9. P. 457-465.
64. Hutterer R. (1993) *Suncus murinus* (Linnaeus, 1766) // In *Mammal Species of the World: a Taxonomic and Geographic Reference*. Eds. Wilson D. E., Reeder D. M.. Washington, D.C.: Smithsonian Institution.
65. Ishikawa A., Akadama I., Namikawa T, et al. (1989) Development of a laboratory line (SRI line) derived from the wild house musk shrew, *Suncus murinus*, indigenous to Sri Lanka // *Experim. Anim.* V. 38. P. 231-237.
66. Ishikawa A., Yamagata T., Namikawa T. (1995) Relationships between morphometric and mitochondrial DNA differentiation in laboratory strains of musk shrews (*Suncus murinus*) // *Japan J. Genet.* V. 70. P. 57-74.
67. Islam A.B.M.M., Hill W.G., Land R.B. (1976) Ovulation rate of lines of mice selected for testis weight // *Genet. Res.* V. 27. P. 23-32.
68. Jarvik G.P., Brunzell J.D., Austin M.A., et al. (1994) Genetic predictors of FCHL in four large pedigrees. Influence of ApoB level major locus predicted genotype and LDL subclass phenotype // *Arterioscler. Thromb.* V. 14. P. 1687-1694.
69. Jarvik G.P. (1998) Complex segregation analyses: uses and limitations // *Am. J. Hum. Genet.* V. 63. P. 942-946.
70. Jensen C.S., Kong A., Kjarulff U. (1995) Blocking Gibbs sampling in very large probabilistic expert systems // *Int. J. Hum. Comput. Stud.* V.42. P.647-666.
71. Karlin S., Carmelli D., Williams R. (1979) Index measures for assessing the mode of inheritance of continuously distributed traits: I. Theory and justification // *Theor. Pop. Biol.* V. 16. P. 106.

72. Kasser T.G., Mabry J.W., Benyshek L.L., Martin R.J. (1986) Heterotic and maternal effects on L and S growth strain rats. I. Reproductive performance // Growth. V. 50. P. 102-108.
73. Kaul M.L.H. (1988) Male sterility in higher plants: Monographs on Theoretical and Applied Genetics. Berlin et al.: Springer-Verlag.
74. Kinoshita T. (1971) Genetic studies on cytoplasmic male sterility of sugar beets (*Beta vulgaris* L.) and its related species // J. Fac. Agric. Hokkaido Univ. V. 56. P. 435-441.
75. Lalouel J.M., Rao D.C., Morton N.E., Elston R.C. (1983) A unified model for complex segregation analysis // Am. J. Hum. Genet. V. 35. P. 816-826.
76. Land R.B., Falconer D.S. (1969) Genetic studies of ovulation rate in mouse // Genet. Res. V. 13. P. 25-46.
77. Land, R. B. (1972) Is mammalian fertility sex limited? // J. Reprod. Fertil. V. 31. P. 512-513.
78. Land R.B. (1973) The expression of female sex limited characteristics in the male // Nature. V. 241. P. 208-209.
79. Land R. B., Carr W. R., Thompson R. (1979) Genetic and environmental variation in the LH response of ovariectomized sheep to LH-RH // J. Reprod. Fertil. V. 56. P. 243-248.
80. Land R.B., Carr W.R., Lee G.J. (1980) A consideration of physiological criteria of reproductive merit in sheep // In Selection Experiment in Laboratory and Domestic Animals. Ed. Robertson A.. Slough: Commonwealth Agriculture Bureau. P. 147-160.
81. Lander E., Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results // Nat. Genet. V. 11. P. 241-247.
82. Lange K., Elston R.C. (1975) Extensions to pedigree analysis I. Likelihood calculations for simple and complex pedigrees // Hum. Hered. V.25. P.95-105.
83. Lange K., Westlake J., Spence M.A. (1976) Extensions to pedigree analysis. III. Variance components by the scoring method // Ann. Hum. Genet. V. 39. P. 485-491.

84. Lange K., Boehnke M. (1983) Extensions to pedigree analysis. V. Optimal calculation of Mendelian likelihoods // *Hum. Hered.* V. 33. P. 291-301.
85. Lange K., Matthysse S. (1989) Simulation of pedigree genotypes by random walks // *Am. J. Hum. Genet.* V. 58. P. 1323-1337.
86. Lange K., Sobel E. (1991) A random walk method for computing genetic location scores // *Am. J. Hum. Genet.* V. 49. P. 1320-1334.
87. Lange K. (1997) An approximate model of polygenic inheritance // *Genetics.* V. 147. P. 1423-1430.
88. Lanneluc I., Drinkwater R. D., Elsen J. M., Hetzel D. J., Nguyen T. C., Piper L. R., Thimonier J., Harrison B., Gellin J. (1994) Genetic markers for the Booroola fecundity (Fec) gene in sheep // *Mamm. Genome.* V. 5. P. 2633.
89. Leipold H.W., Huston K. A herd of glass-eyed albino Hereford cattle // *J. Hered.* 1966. V. 57. P. 179-182.
90. Lynch, M. and Walsh B. (1998) *Genetics and analysis of quantitative traits.* Sinauer Associates.
91. Mather K, Jinks J.L. (1982) *Biometrical genetics.* 3rd Ed. NY: Chapman and Hall.
92. Michaelis P. (1967) The investigation of plasmon segregation by the pattern analysis // *Nucl. Calcutta.* V. 10. P. 1-14.
93. Mikami T., Kinoshita T., Takahashi M. (1974) The genetic relations between pollen restoration and the monogerm character in the crossing with male sterile mutants of sugar beet // *Proc. Amer. Soc. Sugar Beet Technol.* V. 16. P. 75-82.
94. Montgomery G. W., Crawford A. M., Penty J. M. et al. (1993) The ovine Booroola fecundity gene (FecB) is linked to markers from a region of human chromosome 4q // *Nat. Genet.* V. 4. P. 410414.
95. Montgomery G. W., Penty J. M., Lord E. A., Broom M. F. (1995) The search for the Booroola (FecB) mutation // *J. Reprod. Fertil.* V. 49. Supplement. P. 113-121.
96. Morton N.E., MacLean C.J. (1974) Analysis of family resemblance. III. Complex segregation of quantitative traits // *Am. J. Hum. Genet.* V. 26. P. 489-503.

97. Nemana L.J., Marazita M.L., Melnick M. (1992) Genetic analysis of cleft lip with or without cleft palate in Madras, India // *Am. J. Med. Genet.* V. 42. P. 5-9.
98. Newman B., Austin M.A., Lee M., King M.C. (1988) Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families // *Proc. Natl. Acad. Sci. U.S.A.* V. 85. P. 3044-3048.
99. Neyman J., Pearson E.S. (1928) On the use and interpretation of certain test criteria for the purposes of statistical inference // *Biometrika.* V. 20A. P. 175-263.
100. Ollivier L., Sellier P. (1982) Pig genetics: a review // *Ann. Génét. Sél. Anim.* V. 14. P. 481-544.
101. OMIA. Online Mendelian Inheritance in Animals.
<http://www.angis.su.oz.au/Databases/BIRX/omia/>.
102. OMIMTM. Online Mendelian Inheritance in Man.
<http://www.ncbi.nlm.nih.gov/Omim/>.
103. Ott J. (1976) A computer program for linkage analysis of general human pedigrees // *Am. J. Hum. Genet.* V. 28. P. 528-529.
104. Ott J. (1991) Analysis of human genetic linkage. Revised Ed. John Hopkins.
105. Owen F.V. (1945) Citoplasmically inherited male sterility in sugar beet // *Proc. Amer. Soc. Sugar Beet Technol.* V. 71. P. 423-440.
106. Penrose L.S. (1949) The meaning of fitness in human populations // *Ann. Eugen. (London).* V. 14. P. 301-304.
107. Proud C., Donovan D., Kinsey R., Cunningham P.J., Zimmerman D.R. (1976) Testicular growth in boars as influenced by selection for ovulation rate // *J. Anim. Sci.* V. 42. P. 1361-1362.
108. Ray A.K., Field L.L., Marazita M.L. (1993) Nonsyndromic cleft lip with or without cleft palate in west Bengal, India: evidence for an autosomal major locus // *Am. J. Hum. Genet.* 52:1006-1011.
109. Rhoades M.M. (1946) Plastid mutation // *Cold Spring Harbor Symp. Quant. Biol.* V. 11. P. 202-207.

110. Rothschild M.M., Ruvinsky A. (1998) The genetics of the pig. CAB International.
111. Ruedi M., Courvoisier C., Vogel P., Catzefflis F.M. (1996) Genetic differentiation and zoogeography of Asian *Suncus murinus* (Mammalia: Soricidae) // Biol. J. Linnean Society. V. 57. P. 307-316.
112. Ruiz A., Blanco R., Arcos M., Santander J., San M.A. (1997) Complex segregation analysis of schizophrenia in Santiago, Chile // Schizophr. Res. V. 26. P. 65-69.
113. Self S.G., Liang K-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions // J. Am. Stat. Assoc. V. 82. P. 605-610.
114. Sham P. (1998) Statistics in human genetics. John Wiley and Sons.
115. Sheehan N. (1992) Sampling genotypes on complex pedigrees with phenotypic constraints: the origin of the B-allele among the polar Eskimos // IMA J. Math. Appl. Med. Biol. V. 9. P. 1-18.
116. Sheehan N., Thomas A. (1993) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme // Biometrics. V.49. P.163-175.
117. Sokal R.R, Rohlf F.J. (1995) Biometry. NY: W. H. Freeman and Co. 2nd Ed.
118. Stricker C., Fernando R.L., Elston R.C. (1995) An algorithm to approximate the likelihood for pedigree data with loops by cutting // Theor. Appl. Genet. V. 91. P. 1054-1063.
119. Stroup D. (1970) Genic induction and maternal transmission of variegation in *Zea mays* // J. Hered. V. 61. P. 139-141.
120. *Suncus murinus*. Biology of laboratory shrew. Eds. Oda S.I., Kitoh J., Ohta K., Isomura G.. Tokyo: Japan Scientific Societies Press. 1985.
121. Tan W.Y., Chung, W.C. (1972) Convolution approach to genetic analysis of quantitative characters of self-fertilized populations // Biometrics. V. 28. P. 1073-1090.
122. Thoday J.M. (1961) Location of polygenes // Nature. V. 191. P. 368-370.

123. Thomas A. (1986) Optimal computation of probability functions for pedigree analysis // IMA J Math. Appl. Med. Biol. V. 3. P. 167-178.
124. Thompson E.A., Guo S.W. (1991) Evaluation of likelihood ratios for complex genetic models // IMA J. Math. Appl. Med. Biol. V. 8. P. 149-169.
125. Thompson E.A., Lin S., Olshen A.B., Wijsman E.M. (1993) Monte Carlo analysis on a large pedigree // Genet. Epidemiol. V. 10. P. 677-682.
126. Tsubota Y., Namikawa T. (1988) Alleles at plasma alpha-amylase locus (Amy-1) of the house musk shrew (*Suncus murinus*): frequency distribution and new variants in laboratory lines and local Asian populations // Experimental Anim. V. 37. P. 159-64.
127. Wang T., Fernando R.L., Stricker C., Elston R.C. (1996) An approximation to the likelihood for a pedigree with loops // Theor. Appl. Genet. V. 93. P. 1299-1309.
128. Williams W.R., Anderson D.E. (1984) Genetic epidemiology of breast cancer: segregation analysis of 200 Danish pedigrees // Genet. Epidemiol. V. 1. P. 7-20.
129. Wintz H. (1994) Analysis of heteroplasmy in a cytoplasmic mutant of maize // Plant Physiol. Biochem. V. 32. P. 649-653.
130. Wooster R., Neuhausen S.L., Mangion J., et al. (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13 // Science. V. 265. P. 2088-2090.
131. Yamagata T., Tanaka Y., Ishikawa A., Namikawa T., Tomita T. (1990) Genetic relationship among the musk shrews, *Suncus murinus* Insectivora, inhabiting islands and the continent based on mitochondrial DNA types // Biochem. Genet. V. 28. P. 185-195.
132. Yoshikawa T. (1935) On heterochromia irides in swine // Proc. Imper. Acad. Japan. V. 11. P. 125-128.
133. Zimmerman D. (1973) Rh. The intimate history of a disease and its conquest. NY: Macmillan.

Приложение А

Пример подсчета функции правдоподобия

В качестве примера рассмотрим родословную, изображенную на рис. А.1. Допустим, что анализируемый бинарный признак {здоровый/больной} контролируется ауто-сомным диаллельным локусом. Функция пенетрантности описывается тремя параметрами — вероятностями нормального фенотипа у особей с каждым из возможных генотипов:

$$Pr(x = \text{здоров} | g) = \begin{cases} w_{aa} & , \text{ если } g = aa \\ w_{Aa} & , \text{ если } g = Aa \\ w_{AA} & , \text{ если } g = AA . \end{cases}$$

Предположим, что в популяции выполняется распределение Харди-Вайнберга, а частота мутантного аллеля a равна q . Тогда априорное распределение генотипов в популяции $Pr(g)$ задается следующим образом:

$$Pr(g) = \begin{cases} q^2 & , \text{ если } g = aa \\ 2 \cdot q \cdot (1 - q) & , \text{ если } g = Aa \\ (1 - q)^2 & , \text{ если } g = AA . \end{cases}$$

Распределение вероятностей генотипов потомков определяется менделевскими правилами, формализованными в виде условного распределения вероятности генотипов потомка при известных генотипах родителей, $Pr(g_{\text{потомка}} | g_{\text{отца}}, g_{\text{матери}})$. Это распределение описывается тремя параметрами — переходными вероятностями τ_g , или вероятностями получения потомком аллеля A от родителя с генотипом g . При справедливости менделевского наследования $\tau_{AA} = 1$, $\tau_{Aa} = 0.5$, а $\tau_{aa} = 0$.

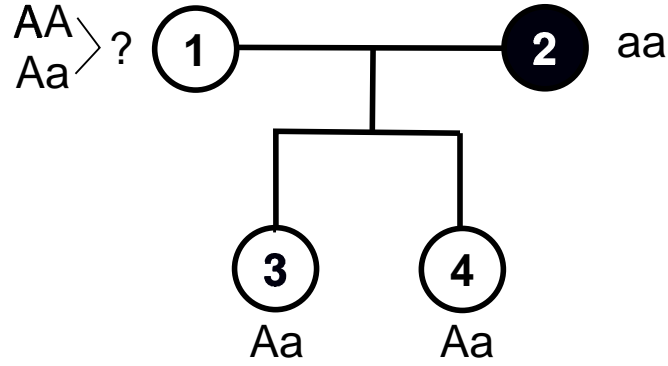


Рис. А.1: Родословная с моногенной болезнью. \circ — здоровые, \bullet — больной. Показаны генотипы, возможные при рецессивной модели контроля

В общем виде функция правдоподобия рассматриваемой родословной равна:

$$L(\vec{x} | \mathcal{S}, \vec{\theta}) = \log_e \sum_{g_1} \sum_{g_2} \sum_{g_3} \sum_{g_4} Pr(g_1) \cdot Pr(g_2) \cdot Pr(g_3 | g_1, g_2) \cdot Pr(g_4 | g_1, g_2) \cdot \\ \cdot Pr(x_1 = \text{здоров} | g_1) \cdot Pr(x_2 = \text{болен} | g_2) \cdot \\ \cdot Pr(x_3 = \text{здоров} | g_3) \cdot Pr(x_4 = \text{здоров} | g_4) .$$

Ее подсчет требует суммирования вероятностей родословной при всех возможных комбинациях генотипов членов родословной ($3^4 = 81$ суммирование).

При некоторых моделях наследования число суммирований может быть уменьшено. Например, при рецессивном контроле признака, когда $w_{aa} = 0$, а $w_{Aa} = w_{AA} = 1$, функция правдоподобия может быть представлена как сумма только двух слагаемых. Это обусловлено тем, что при рецессивной модели наследования у трех членов рассматриваемой родословной генотипы определены однозначно, а у четвертого (особь 1) — возможно 2 генотипа. Поэтому функция правдоподобия имеет вид:

$$L(\vec{x} | \mathcal{S}, \vec{\theta}) = \log_e \left(Pr(Aa) \cdot Pr(aa) \cdot Pr(Aa | Aa, aa) \cdot Pr(Aa | Aa, aa) \cdot \right. \\ \cdot Pr(x_1 = \text{здоров} | Aa) \cdot Pr(x_2 = \text{болен} | aa) \cdot \\ \cdot Pr(x_3 = \text{здоров} | Aa) \cdot Pr(x_4 = \text{здоров} | Aa) + \\ \left. + Pr(AA) \cdot Pr(aa) \cdot Pr(Aa | AA, aa) \cdot Pr(Aa | AA, aa) \cdot \right. \\ \cdot Pr(x_1 = \text{здоров} | AA) \cdot Pr(x_2 = \text{болен} | aa) \cdot \\ \left. \cdot Pr(x_3 = \text{здоров} | Aa) \cdot Pr(x_4 = \text{здоров} | Aa) \right) =$$

$$\begin{aligned}
&= \log_e \left(2 \cdot q \cdot (1 - q) \cdot q^2 \cdot \tau_{Aa} \cdot (1 - \tau_{aa}) \cdot \tau_{Aa} \cdot (1 - \tau_{aa}) \cdot \right. \\
&\quad \cdot w_{Aa} \cdot (1 - w_{aa}) \cdot w_{Aa} \cdot w_{Aa} + \\
&\quad + (1 - q)^2 \cdot q^2 \cdot \tau_{AA} \cdot (1 - \tau_{aa}) \cdot \tau_{AA} \cdot (1 - \tau_{aa}) \cdot \\
&\quad \cdot w_{AA} \cdot (1 - w_{aa}) \cdot w_{Aa} \cdot w_{Aa} \Big) = \\
&= \log_e \left(1/2 \cdot q \cdot (1 - q) \cdot q^2 + (1 - q)^2 \cdot q^2 \right).
\end{aligned}$$

Параметр q может быть оценен методом максимума правдоподобия, с помощью решения уравнения

$$\frac{\partial L(\vec{x} | \mathcal{S}, \vec{\theta})}{\partial q} = \frac{4 \cdot q^2 - 9 \cdot q + 4}{q \cdot (q^2 - 3 \cdot q + 2)} = 0.$$

Полученная таким образом оценка q равна 0.61.

Понятно, что как только родословные становятся больше или модели наследования усложняется, вычисление функции правдоподобия и нахождение ее максимума становится посильным только при привлечении вычислительной техники и использовании сложных вычислительных алгоритмов.

Приложение В

Описание пакета программ МАІА-1.0

Пакет компьютерных программ МАІА разработан для комплексного сегрегационного анализа признаков животных. Пакет позволяет анализировать наследование количественных и бинарных признаков по материалу, представленному родословными, полученными при скрещивании представителей одной или двух популяций. Анализируемые родословные могут содержать множественные петли. В настоящее время в пакете реализовано несколько моделей наследования признаков: майоргенная модификация смешанной модели (для количественных и бинарных признаков) и майоргенная модель наследования совместных признаков, характеризующих пару родителей.

Пакет программ написан на языке FORTRAN-77. Основой его послужил ранее разработанный в лаборатории пакет программ МАН-1 (Аксенович, Гинзбург, 1987). В настоящее время пакет реализован для платформы MS-DOS, планируется его перенос на операционную систему UNIX.

Пакет состоит из четырех программ (`recode`, `chstruct`, `peelseq` и `analysis`, см. рис. В.1). Три первых являются служебными программами. Первая из них (`recode`) находит в исходных данных отдельные родословные, делает перекодировку и сортировку, т.е. преобразует данные в формат, удобный для дальнейшего анализа. Вторая (`chstruct`) проверяет, присутствуют ли в родословной петли и, при необходимости, преобразует структуру родословной. Третья (`peelseq`) находит последовательность свертки родословной.

Последняя программа (`analysis`) позволяет проводить комплексный сегрегационный анализ: она производит подсчет и максимизацию функции правдоподобия.

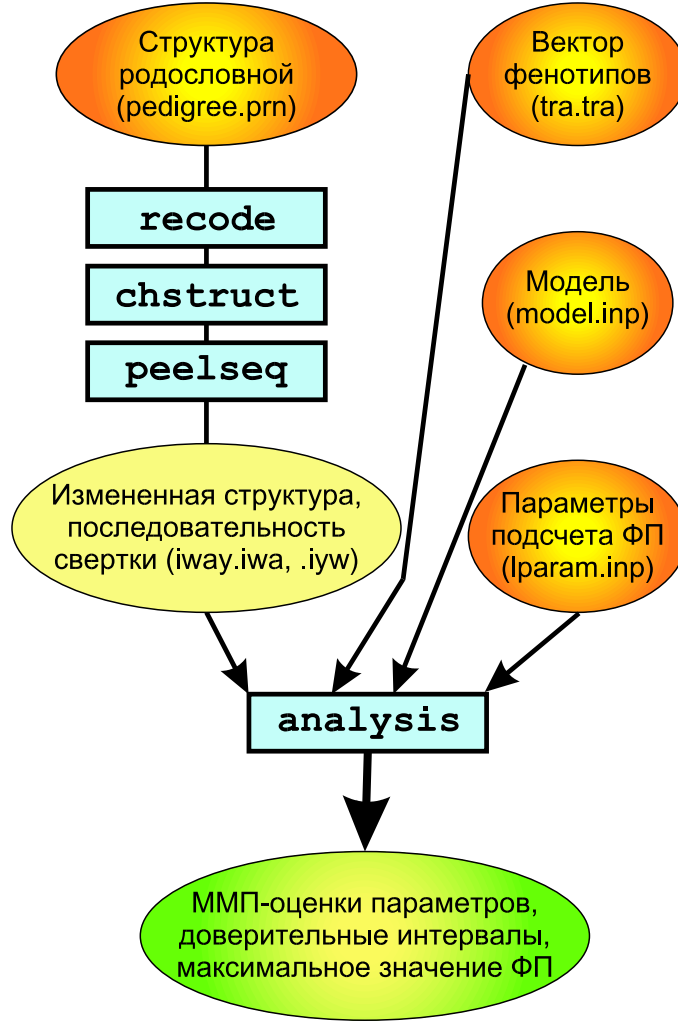


Рис. В.1: Структура пакета программ МАІА-1.0. Прямоугольники — программные модули, овалы — блоки данных и параметров.

Подсчет ФП, $L(\vec{x}|\mathcal{S}, \vec{\theta})$, требует знания вектора фенотипов членов родословной \vec{x} , структуры родословной \mathcal{S} и вектора параметров $\vec{\theta}$. Кроме того, для применения эффективного алгоритма вычисления ФП необходимо знать последовательность свертки \vec{P}_s . Эти векторы хранятся в различных файлах, используемых программой (см. рис. В.1). Вектор фенотипов членов родословной хранится в файле `tra.tra`, структура родословной и последовательность свертки хранятся в файлах `iway.iwa` и `iway.iyw`, а набор параметров, ограничения, наложенные на них и их начальные значения — в файле `model.inp`. Кроме того, отдельный файл `lparam.inp` содержит указания о способе подсчета ФП: применяется ли точный метод или приближительный, если приближительный, то какой именно.

Максимизация ФП проводится численно, с помощью подпрограммы METHGI, разработанной в нашей лаборатории совместно с ВЦ СО АН СССР.

Пакет программ (исполняемые модули и исходный код), а также его более подробное описание можно найти на Internet-сайте Сектора рекомбинационного и сегрегационного анализа <http://mga.bionet.nsc.ru>

Приложение С

Вычисление апостериорных вероятностей генотипов по гену-ингибитору масти

Для того, чтобы определить влияние гена-ингибитора масти на гетерохромию, в модель было введено математическое ожидание эффекта этого гена:

$$E[b(\mathcal{I})] = \sum_{\mathcal{I}=Ii,ii} b(\mathcal{I}) \cdot Pr_a(\mathcal{I}).$$

Для вычисления этого математического ожидания надо найти $Pr_a(\mathcal{I})$ — апостериорную вероятность того, что особь имеет генотип \mathcal{I} .

Модель наследования белой масти формализуется в виде вектора параметров $\vec{\theta}_{\mathcal{I}}$. Апостериорная вероятность генотипа $\mathcal{I} \in \{II, Ii, ii\}$ у j -й особи равна

$$Pr_a(\mathcal{I}_j) = \text{Exp}\{L(\vec{x}, \mathcal{I}_j | \vec{\theta}_{\mathcal{I}}, \mathcal{S}) - L(\vec{x} | \vec{\theta}_{\mathcal{I}}, \mathcal{S})\},$$

где $L(\vec{x}, \mathcal{I}_j | \vec{\theta}_{\mathcal{I}}, \mathcal{S})$ — функция правдоподобия родословной, вычисленная при генотипе особи j , фиксированном в значении \mathcal{I}_j .

Таким образом, для вычисления $Pr_a(\mathcal{I})$ необходимо знать $\vec{\theta}_{\mathcal{I}}$ — модель наследования.

Известно, что белая масть свиньи контролируется одним аутосомным геном с полным доминированием (Spillman, 1906; Wentworth, Lush, 1923; Hetzer, 1945; цит. по: Ollivier, Sellier, 1982).

Таким образом, пенетрантности генотипов, вообще говоря, могли быть априорно зафиксированы. Однако, частота аллеля i в популяции СМ не была известна. Для

их оценки использовали комплексный сегрегационный анализ.

Все 13 свиноматок КБ имели белую масть. Из 6-ти хряков МС один был окрашенным, масть остальных не была зафиксирована. В поколениях F_1 и в бэкроссе на КБ все потомство имело белую масть. В бэкроссе на МС происходило расщепление: 263 особи имело белую масть, а 60 были окрашенными.

Сегрегационный анализ подтвердил гипотезу о том, что белая масть контролируется одним аутосомным геном с полным доминированием. Оценка частоты аллеля i у СМ составила 0.4, КБ была мономорфна по аллелю I . Апостериорные вероятности генотипов по гену-ингибитору масти подсчитывались при условии этой модели.

Приложение D

Формализация

ген-цитоплазматического

наследования

Фенотипическая реализация генотипа формализуется функцией $Pr(x|g_{MUT}, g_{FR_1}, g_{FR_2}, C_m)$ — вероятностью формирования фенотипа x при условии генотипа g_{MUT} по гену-мутатору, g_{FR_1}, g_{FR_2} по генам FR и унаследованной материнской цитоплазмой типа C_m . Предполагаемый ген-мутатор может вызвать конверсию цитоплазмы, унаследованной от матери. Значение признака $x \in \{\text{мужскостерильное, гермафродитное}\}$ определяется типом цитоплазмы и генотипом по генам FR . (см. рис. 6.2).

При отсутствии сцепления между геном-мутатором и генами FR функция пенетрантности $Pr(x|g_{MUT}, g_{FR_1}, g_{FR_2}, C_m)$ может быть выражена следующим образом:

$$Pr(x|g_{MUT}, g_{FR_1}, g_{FR_2}, C_m) = \sum_{C_i \in \{\mathbf{S}, \mathbf{N}\}} Pr(x|g_{FR_1}, g_{FR_2}, C_i) \cdot Pr(C_i|g_{MUT}, C_m),$$

где C_i — возможный тип цитоплазмы у растения ($C_i = \mathbf{S}$ или \mathbf{N}), g_{FR_1} и g_{FR_2} — его генотип по генам FR_1 и FR_2 , g_{MUT} — генотип по гену-мутатору. $Pr(C_i|g_{MUT}, C_m)$ — вероятность C_i -го типа цитоплазмы при условии, что материнское растение имеет C_m тип цитоплазмы. При предположении об отсутствии реверсий цитоплазмы из \mathbf{S} -типа в \mathbf{N} -тип эта вероятность определяется пенетрантностями генотипов гена-мутатора w_{MutMut} , w_{Mutmut} и w_{mutmut} :

$$Pr(C_i = \mathbf{S}|g_{MUT}, C_m = \mathbf{S}) = 1,$$

$$Pr(C_i = \mathbf{N} | g_{MUT}, C_m = \mathbf{S}) = 0,$$

$$Pr(C_i = \mathbf{S} | g_{MUT}, C_m = \mathbf{N}) = w_{g_{MUT}},$$

$$Pr(C_i = \mathbf{N} | g_{MUT}, C_m = \mathbf{N}) = 1 - w_{g_{MUT}}.$$

$Pr(x | g_{FR_1}, g_{FR_2}, C_i)$ — вероятность фенотипа x у растения с генотипом g_{FR_1}, g_{FR_2} и C_i типом цитоплазмы. При аддитивной модели восстановления фертильности она определяется одним параметром ω_{FR} . Этот параметр соответствует вкладу, который делает аллель Fr_1 в восстановление фертильности. Учитывая, что наличие всех доминантных аллелей генов FR приводит к полному восстановлению фертильности, вклад аллеля Fr_2 может быть представлен как $\left(\frac{1}{2} - \omega_{FR}\right)$. Пусть $N(g_{FR_1})$ и $N(g_{FR_2})$ — число доминантных аллелей в каждом из локусов. Тогда пенетрантность некоторого генотипа g_{FR_1}, g_{FR_2} (вероятность восстановления фертильности на фоне \mathbf{S} -цитоплазмы) выразится следующим образом:

$$w_{g_{FR_1}, g_{FR_2}} = N(g_{FR_1}) \cdot \omega_{FR} + N(g_{FR_2}) \cdot \left(\frac{1}{2} - \omega_{FR}\right),$$

а вероятности $Pr(x | g_{FR_1}, g_{FR_2}, C_i)$ запишутся как

$$Pr(x = \text{гермафродитное} | g_{FR_1}, g_{FR_2}, C_i = \mathbf{N}) = 1,$$

$$Pr(x = \text{мужскостерильное} | g_{FR_1}, g_{FR_2}, C_i = \mathbf{N}) = 0,$$

$$Pr(x = \text{гермафродитное} | g_{FR_1}, g_{FR_2}, C_i = \mathbf{S}) = w_{g_{FR_1}, g_{FR_2}},$$

$$Pr(x = \text{мужскостерильное} | g_{FR_1}, g_{FR_2}, C_i = \mathbf{S}) = 1 - w_{g_{FR_1}, g_{FR_2}}.$$