С. В. Малов
И. Ю. Малова

ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ
1886
СПбГЭТУ «ЛЭТИ»
У П

БАЗОВЫЕ МОДЕЛИ БИОСТАТИСТИКИ

АНАЛИЗ РЕЗУЛЬТАТОВ ГЕНЕТИЧЕСКИХ ИССЛЕДОВАНИЙ

САНКТ-ПЕТЕРБУРГ
2025

МИНОБРНАУКИ РОССИИ

Санкт-Петербургский государственный электротехнический
университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

С. В. МАЛОВ И. Ю. МАЛОВА

БАЗОВЫЕ МОДЕЛИ БИОСТАТИСТИКИ

АНАЛИЗ РЕЗУЛЬТАТОВ ГЕНЕТИЧЕСКИХ ИССЛЕДОВАНИЙ

Учебное пособие

Санкт-Петербург
Издательство СПбГЭТУ «ЛЭТИ»
2025

УДК 519.2(07)
ББК В 172я7
М19

Малов С. В., Малова И. Ю.

М19 Базовые модели биостатистики. Анализ результатов генетических исследований: учеб. пособие. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2025. 80 с.

ISBN 978-5-7629-3488-6

Рассмотрено использование методов математической статистики в анализе результатов генетических исследований.

Предназначено для поддержки дисциплины «Статистический анализ и основы биостатистики» на ФКТИ СПбГЭТУ «ЛЭТИ», будет интересно студентам и аспирантам, обучающимся по программам подготовки специалистов в области информационных технологий.

УДК 519.2(07)
ББК В 172я7

Рецензенты: кафедра высшей математики ВШТЭ СПбГУПТД; д-р техн. наук, проф. Л. В. Уткин (СПбПУ им. Петра Великого).

Утверждено
редакционно-издательским советом университета
в качестве учебного пособия

ISBN 978-5-7629-3488-6

Оглавление

Введение

Различные свойства живого организма в значительной мере определяются его генетическими особенностями, или генетическим кодом. Генетический код содержит в себе важную информацию для производства белков, из которых строится сам организм, а также все системы, обеспечивающие его функционирование. Наличие генетической информации дает возможность предсказывать многие жизненно важные свойства организма.

Хранение и передачу генетической информации обеспечивает дезоксирибонуклеиновая кислота (ДНК). Молекула ДНК состоит из двух комплиментарных цепочек, химически связанных между собой и закрученных сложным образом. Каждая цепочка ДНК составлена их четырех азотистых оснований (нуклеотидов) – аденина (A), гуанина (G), тимина (T) и цитозина (C). Молекула ДНК хранится в хромосомах. У диплоидных организмов (например, у человека) существуют парные хромосомы, ДНК в которых различаются незначительно. В результате секвенирования и последующей расшифровки исследователь получает генетический код, представляющий собой набор букв «A»,«G»,«T»,«C», разбитый на хромосомы. В процессе размножения происходят минимальные изменения генетического кода, называемые мутациями, которые в процессе эволюции приводят к значительному разнообразию видов. Среди мутаций наиболее распространены единичные замены, представляющие собой замену одного нуклеотида на другой в определенной позиции (локусе), — транзиции (A↔G; T↔C) и трансверсии (A↔T; A↔C; G↔C; G↔T). Встречаются и более сложные мутациии – инсерции/делеции, инверсии, дупликации, транслокации. Мутации происходят постоянно, однако в случае многоклеточных организмов большинство мутаций не влияют на эволюцию, так как касаются только одного организма. Лишь мутации, происходящие в стволовых клетках, могут закрепиться в потомстве и влиять на ход эволюции.

Важнейшую роль в генетических исследованиях играет статистика, позволяющая получить рабочие гипотезы для установления связи между генотипом и интересующим свойством организма (фенотипом). Следует отметить, что статистика не гарантирует абсолютной достоверности полученных выводов, поэтому для надежного подтверждения связи генотипа и фенотипа, найденной статистическим путем, желательно получить иные аргументы, основанные на понимании причин этой связи. К сожалению, абсолютно достоверные аргументы подтверждения установленных связей генотипа с фенотипом удается найти нечасто, поэтому статистические вы-

воды чаще всего остаются единственным средством установления и подтверждения установленных связей.

Все растущее число научных исследований порождает множество невоспроизводимых (ложноположительных) выводов, в связи с чем следует подходить с особой осторожностью к интерпретации результатов статистического анализа.

1. ГЕНЕТИЧЕСКИЕ ИССЛЕДОВАНИЯ

Особенность постановки статистического эксперимента для генетических исследований состоит в том, что для проведения статистического анализа требуются наличие одной или нескольких наблюдаемых характеристик (фенотипов) и наборы генетических маркеров (генотипов) каждого объекта исследования.

Фенотипы могут представлять собой результаты проведения клинических исследований, значение определенного или нескольких внешних признаков объекта исследования, а также их этногеографические характеристики. В биомедицинских исследованиях для получения фенотипов реализуют стандартные планы биостатистики.

1.1. Генетические данные

Генотипы обычно представляют собой наборы генетических маркеров. Вариации в одной позиции исследуемого участка ДНК (локуса), закрепившиеся в популяции, получили название однонуклеотидных полиморфизмов (ОНП)[1]. Ввиду наличия парных хромосом у диплоидных организмов ОНП представляет собой неупорядоченный набор аллелей, где каждую аллель определяет вариант нуклеотида фиксированного индивида в данной позиции на соответствующей хромосоме. Вероятность закрепления в популяции более двух вариантов ОНП в одной позиции крайне мала. Хотя существует широкий класс мультивариантных ОНП, в пределах одной популяции мультивариантные ОНП обычно имеют два варианта. Наличие идентичных нуклеотидов в данной позиции на обеих парных хромосомах называется гомозиготным вариантом (например, AA, TT), тогда как при наличии различных вариантов (например, AT, AG) – гетерозиготным. При наличии двух вариантов аллелей ОНП представляет собой признак с тремя уровнями, включающими 2 гомозиготных варианта и один – гетерозиготный. Варианты более сложных мутаций тоже довольно часто сводят к биаллельным, группируя все варианты, отличные от одного выбранного.

[1]Single Nucleotide Polymorphism (SNP), *англ.*

Для определения местоположения генетического маркера используются его координаты, состоящие из номера хромосомы (парные хромосомы имеют один и тот же номер) и его позиции на хромосоме. Генетические варианты обычно определяются сравнением генетических данных изучаемой популяции с референсным геномом, представляющим собой характерную последовательность букв A, G, C, T. Большая часть последовательности всех представителей популяции совпадает с референсным геномом, в остальных же позициях определяются генетические варианты. Длина генома человека приближенно равна $3.2 \cdot 10^9$ пар нуклеотидов. На текущий момент описано более 500 млн ОНП в организме человека. Разумеется, число ОНП в ограниченной популяции случайно и зависит от ее размера, поэтому в рамках одного исследования находится гораздо меньше ОНП. В частности, число ОНП в популяции 1000 человек все еще достаточно велико и может достигать 10 млн. Более сложных мутаций, разумеется, гораздо меньше. В исследуемой популяции чаще встречающийся вариант аллели обычно называют общим («common»,«wild type»), а альтернативный – редким («minor»).

Существует 2 основных подхода для получения генетических данных – полногеномное секвенирование и ДНК-микрочипы. Полный набор вариантов получают в результате полногеномного секвенирования. В основе метода лежит репликация коротких фрагментов ДНК и «выравнивание» их на референсный геном. Общее число генетических вариантов зависит от размера популяции. К сожалению, в результате полногеномного секвенирования далеко не всегда удается точно восстановить каждый генетический вариант. Обычно ненадежные варианты отфильтровываются на базе специальных показателей, вычисляемых при выравнивании, но и это не гарантирует, что все остальные варианты генотипированы корректно. Кроме того, полногеномное секвенирование – очень дорогой способ получения генетических данных. Более дешевый способ – использование ДНК-микрочипов, позволяющих определить значения только фиксированного набора генетических маркеров. Современные стандартные ДНК-микрочипы позволяют определять до $5 \cdot 10^6$ генетических вариантов, тогда как более старые ДНК-микрочипы содержали менее $5 \cdot 10^5$ генетических вариантов. Возможен заказ специализированных ДНК-микрочипов, ориентированных на определенные исследования. Основной недостаток секвенирования на базе ДНК-микрочипов – невозможность определения новых вариантов. Как и в случае полногеномного секвенирования, при использовании ДНК-микрочипов возможны ошибки. Ввиду относительной дешевизны технологии ДНК-микрочипы позволяют секвенировать большие популяции.