

ГРИШЕЧКИН С. А.

**О СВЯЗИ МЕТОДА ДНК-ДАКТИЛОСКОПИИ
С ЗАДАЧЕЙ О РАЗМЕЩЕНИИ ЧАСТИЦ ПО ЯЧЕЙКАМ**

1. Введение

Математические вопросы, которые будут изучаться в настоящей статье, имеют чисто практическое происхождение. Они возникли в результате попытки построить строгое теоретическое обоснование для ДНК-дактилоскопии. ДНК-дактилоскопия — это новый научный метод в криминалистике, который произвел революцию в расследовании наиболее тяжких преступлений против личности — убийств, изнасилований и т. п. Типичная ситуация, при которой применяется этот метод: совершено убийство, и на одежде подозреваемого найдены следы крови. Если это кровь жертвы, то это тяжчайшая, решающая улика, и подозреваемый наверняка будет признан виновным в убийстве.

Все, что могли сделать эксперты-криминалисты раньше, это сравнить, например, группы крови жертвы и крови на одежде подозреваемого. Однако групп крови всего 4, и совпадение не говорит еще о том, что кровь на одежде — это именно кровь жертвы. Кровь на одежде могла появиться совершенно в иной ситуации, и совпадение с группой крови жертвы имеет чисто случайный характер. В распоряжении криминалистов были и другие тесты, но даже все вместе они не были достаточно информативными в том смысле, что совпадение все равно можно было бы рассматривать как чисто случайное. Вероятность этого (т. е. процент населения с теми же самыми результатами тестов) был, скажем, более 2%.

Метод ДНК-дактилоскопии позволил снизить эту цифру во много раз — скажем, до 10^{-6} – 10^{-7} . Это количественное изменение привело к качественно другой роли экспертизы. Теперь стало возможным практически с полной уверенностью утверждать, что кровь на одежде подозреваемого — это именно кровь жертвы. Вероятность ошибиться, по существу, и совпадает с указанными выше цифрами 10^{-6} – 10^{-7} . При этом для проведения экспертизы достаточно микроскопических количеств образца — крови, спермы, волос и т. п. Описан случай, когда идентификация была проведена по мельчайшей капельке слюны, которую жертва оставила на очках преступника. Теоретически достаточно

лишь одной молекулы, так как используемая в настоящее время техника (так называемая полимеразная цепная реакция) позволяет осуществлять своеобразное клонирование — генерировать генетические копии образца в количестве, необходимом для проведения анализа.

Метод ДНК-дактилоскопии был впервые применен в Великобритании в 1986 г. и затем в 1988 г. в США. После этого он стал стремительно распространяться по всему миру. Однако в начале 90-х годов, после непродолжительного периода триумфального шествия, возникла бурная дискуссия в широких научных и общественных кругах о возможности применения этого метода в судопроизводстве [1], [6]–[13]. В центре критики оказались именно те цифры, о которых шла речь выше. Утверждалось, что процедура их получения опирается на многочисленные допущения, которые нельзя доказать теоретически и которые невозможно в настоящее время проверить на практике. Возможно, что вероятность ошибки во много раз выше, чем 10^{-6} – 10^{-7} . Поэтому, как утверждают противники ДНК-дактилоскопии, ее нельзя применять в судебной практике до тех пор, пока не будет создан надлежащий теоретический базис.

Как следствие этой бурной дискуссии стали возникать прецеденты, когда суды в США и Великобритании перестали принимать к рассмотрению данные ДНК-анализа, и возникла реальная опасность дискредитации этого исключительно важного метода. Главной причиной такой ситуации является некорректный, неквалифицированный анализ вероятностных аспектов ДНК-дактилоскопии, что и сделало ее уязвимой для критики.

Обратимся теперь к фактической стороне дела и опишем процедуру ДНК-дактилоскопии (разумеется, в упрощенной форме), а также методы подсчета вероятности ошибки.

Двойная спираль ДНК представляет собой последовательность нуклеотидов, каждый из которых может быть одного из 4-х типов — аденин (А), цитозин (С), гуанин (G) или тимин (Т). Число возможных вариантов спиралей ДНК исключительно велико. Однако непосредственно установить, в какой последовательности располагаются нуклеотиды у данного человека, невозможно. Поступают иным, обходным путем. Существуют специальные реактивы (enzyme), которые разрывают спираль ДНК каждый раз, когда встречается определенная комбинация нуклеотидов. Например, enzyme HaeI разрывает спираль каждый раз, когда встречается комбинация AGGCCA. Образовавшиеся фрагменты ДНК могут быть упорядочены по длине при помощи процесса электрофореза. На фотографии геля, в котором происходит электрофорез, эксперт видит полосы, которые располагаются тем выше, чем меньше длина соответствующего участка ДНК.

Эта техника используется в криминалистике следующим образом [7]. Существуют участки ДНК (локусы в генетической терминологии) с так называемыми тандемными повторами. Одна и та же комбинация

нуклеотидов присутствует у всех людей, но у одного человека она встречается в единичном экземпляре, у другого она повторяется в этом локусе дважды подряд, у третьего — трижды и т. п. Максимальное количество K повторов этой комбинации нуклеотидов, которое может встретиться у людей, известно. Говорят, что в данном локусе может быть K возможных аллелей, подразумевая под аллелью определенное количество повторов данной комбинации нуклеотидов.

Специальные реактивы рассекают спираль ДНК в начале и в конце данного локуса. В результате образуется обрывок ДНК, который содержит $1, 2, \dots, K$ повторов рассматриваемой комбинации нуклеотидов. Эти обрывки занимают при электрофорезе строго определенное место, в соответствии со своей массой. Эксперт видит на фотографии черту в одном из K возможных мест.

Каждый человек получает одну из парных хромосом от отца, другую — от матери. Всего таких пар хромосом 23. Один и тот же locus будет встречаться дважды — в хромосоме отца и в хромосоме матери. Поэтому эксперт наблюдает на фотографии две аллели (возможно, они совпадут): одну, унаследованную от отца, а вторую — от матери. Общее число различных комбинаций двух аллелей равно $K(K + 1)/2$. Например, при типичном значении $K = 10$ получается 55 различных сочетаний (генотипов).

В реальности исследуются несколько локусов (обычно от 5 до 10). Это во много раз увеличивает число возможных генотипов (под генотипом условимся понимать определенное сочетание аллелей в исследуемых локусах). Обозначим это число N . Например, для 7 локусов и 55 возможных сочетаний по каждому из локусов мы получаем $N = 55^7 \simeq 1.52 \times 10^{12}$ возможных генотипов, что почти в 400 раз превышает население Земли. Обозначим все возможные генотипы цифрами $1, 2, \dots, N$.

Обратимся теперь к интерпретации совпадения генотипов. Пусть, например, на месте преступления обнаружен криминальный объект (кровь, волосы, семя и т. п.), который имеет генотип Γ_1 . Из обстоятельств дела ясно, что этот объект принадлежит преступнику. Пусть генотип подозреваемого совпал с Γ_1 . Как следует интерпретировать эту ситуацию?

Этот, казалось бы, простой вопрос таковым не является. Именно в этом месте начинается применение теории вероятностей к реальной жизни, т. е. построение соответствующей математической модели. Этих моделей (близких по духу) может быть несколько. Например, если желать, чтобы экспертное заключение имело вид «криминальный объект принадлежит подозреваемому с вероятностью 0.9999», то неминуемо привлечение байесовского подхода. В этом случае необходимо априори предположить, что криминальный объект принадлежит подозреваемому с некоторой вероятностью (т. е. фактически, что подозреваемый является преступником с некоторой априорной вероятностью). Однако в ре-

альной жизненной ситуации вряд ли можно найти разумные способы задания этой априорной вероятности.

Наиболее естественной нам представляется следующая трактовка. Речь идет, по существу, о том, что метод ДНК-дактилоскопии может привести к осуждению невиновного человека. Пусть, для определенности, генотип $\Gamma_1 = i$. Под вероятностью того, что невиновный человек будет осужден, станем понимать вероятность того, что случайно выбранный индивидум из данной популяции имеет генотип $\Gamma_2 = i$. Обозначим эту вероятность $MP(i) = \mathbf{P}\{\Gamma_2 = \Gamma_1 \mid \Gamma_1 = i\}$ (MP = match probability). Тогда надежность метода ДНК-дактилоскопии определим как $1 - MP(i)$.

Такая трактовка вполне согласуется со стандартами классической математической статистики. По существу, мы скопировали идеологию, используемую, например, при построении доверительных интервалов, когда выражение вида $\mathbf{P}\{\alpha \in (1, 2)\} = 0.95$ означает, что данный метод построения доверительного интервала приводит к правильным границам для неизвестного параметра α в 95% случаев. Иначе говоря, 0.95 — это надежность рассматриваемого метода построения доверительного интервала.

Введенная вероятность ошибки $MP(i)$ является общеупотребительным параметром в литературе по ДНК-дактилоскопии. Однако при попытке ее вычислить возникает несколько серьезных вопросов, по которым и развернулась бурная дискуссия. Упомянем два из них.

1. Генотип Γ_2 определен выше как генотип случайно выбранного человека из данной популяции. Но какая именно популяция должна быть рассмотрена? Если придерживаться приведенной выше трактовки $MP(i)$ как вероятности ошибки метода, т.е. вероятности того, что невиновный (=случайно взятый) человек будет иметь генотип i , то естественно в качестве этой популяции выбирать, например, все население России (коль скоро нас интересует надежность этого метода в России).

Представим, однако, гипотетическую ситуацию, когда вероятность $MP(i)$, подсчитанная для всего населения России, равна 10^{-6} , а для лиц некоторой национальности A величина $MP(i)$ равна 0.1. Какое из двух значений следует использовать? Конечно, более правильно использовать вторую цифру, которая есть, фактически, условная вероятность ошибки метода при условии, что (а) $\Gamma_1 = i$ и (б) подозреваемый имеет национальность A . К сожалению, такие рассуждения быстро приводят в тупик, так как аналогичным образом можно требовать подсчета, например, условной вероятности ошибки при условии, что (а) $\Gamma_1 = i$ и (б) подозреваемый есть уроженец, скажем, Пензы.

С математической точки зрения это означает, что имеющиеся в нашем распоряжении средства не дают возможности вычислить подобные условные вероятности. В этих случаях, по нашему мнению, будет правильно ограничиться безусловной вероятностью ошибки как

некой усредненной цифрой, отражающей надежность метода ДНК-дактилоскопии при использовании его в многочисленных случаях по всей территории России. Более того, как будет показано ниже, даже условную вероятность $MP(i) = \mathbf{P}\{\Gamma_2 = \Gamma_1 \mid \Gamma_1 = i\}$, по-видимому, хорошо оценить при современном уровне знаний невозможно. Это ведет к заключению, что наиболее правильным выходом будет оценка безусловной вероятности $MP = \mathbf{P}\{\Gamma_2 = \Gamma_1\}$, где Γ_1 — это случайная величина, совпадающая с генотипом криминального образца в случайно взятой экспертизе, в Γ_2 — генотип случайно взятого человека. Разумно предполагать, что Γ_1 и Γ_2 — независимые и одинаково распределенные случайные величины. Величина $1 - MP$ измеряет общую, интегрированную надежность метода ДНК-дактилоскопии по всем экспертизам. Действительно, MP , в рамках сделанных допущений, есть вероятность того, что невиновный человек будет осужден в некотором наудачу выбранном криминальном случае.