ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2013, том 23, № 3, с. 20–25

ИСПОЛЬЗОВАНИЕ И МЕТОДИКИ ВЫСОКОТЕХНОЛОГИЧНЫХ ИЗМЕРЕНИЙ

УДК 577.2, 519.226, 57.087.1

© А. Л. Буляница, Д. Г. Сочивко, А. А. Федоров, В. Е. Курочкин

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ КОЛИЧЕСТВЕННОГО АНАЛИЗА С ИСПОЛЬЗОВАНИЕМ ПОЛИМЕРАЗНОЙ ЦЕПНОЙ РЕАКЦИИ В РЕАЛЬНОМ ВРЕМЕНИ (ПЦР-РВ) НА ОСНОВЕ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ

Представленная работа является продолжением цикла работ [1–3], в которых методом стохастического имитационного моделирования исследовались кинетика ПЦР и количественный анализ ДНК с ее помощью. В частности, были даны оценки соответствующего закона распределения в зависимости от начальной концентрации цепей ДНК, проанализирована кинетика реакции, исследованы погрешности количественного анализа. В данной работе рассматривается круг вопросов, связанных с интерпретацией результатов, а именно оценивание числа частиц (n) в исходной пробе на основе серии результатов (k_i, i =1, 2,…,m) количественного анализа для нескольких лунок, получение самих оценок k_i и интерпретация ложно- или истинно-отрицательного результата анализа, связанного с обнаружением нуля частиц в m повторных анализах.

Кл. сл.: полимеразная цепная реакция в реальном времени, вероятностная схема, формула Байеса, количественный анализ, ложно-отрицательный результат

ВВЕДЕНИЕ

Полимеразная цепная реакция (ПЦР) лежит в основе множества молекулярных методов, применяемых для качественного и количественного анализа специфических последовательностей нуклеиновых кислот. Процесс ПЦР представляет собой цепную реакцию размножения анализируемого фрагмента ДНК, который можно представить в виде дискретного ветвящегося случайного процесса с участием нескольких типов частиц [1]. Если в процессе реакции производить измерение концентрации ее продуктов на каждом цикле, то мы получим кинетическую кривую ПЦР, имеющую дискретный по времени характер. Анализ динамики накопления продуктов реакции позволяет определить исходное количество копий ДНК в пробе, что является основой наиболее распространенного метода количественного анализа нуклеиновых кислот — ПЦР с регистрацией продуктов реакции в режиме реального времени (ПЦР-РВ).

Стохастическая природа явлений, связанных с проведением количественного анализа методом ПЦР-РВ, определяется рядом факторов:

а) из исходного анализируемого объема, содержащего n копий, случайным образом отбирается m проб меньшего объема,

б) увеличение числа частиц (амплификация) моделируется цепочкой случайных событий (создание новой копии на цикле амплификации с вероятностью p, и отсутствие создания копии с вероятностью $q = 1 – p$, т. е. на основе биномиального распределения) и

в) результаты количественного анализа (k_i выявленных частиц) отличаются для различных проб (т. е. сформирована случайная выборка m элементов, требующая оценивания интегральных характеристик — математического ожидания и дисперсии или / и доверительного интервала).

Имитационная модель ПЦР-РВ на основе стохастического алгоритма детально описана в работе [1]. Базовая идея — вероятностный характер синтеза цепей ДНК. При этом задается вероятность P, зависящая от количества копий N в реакционном объеме. Зависимость P (N) также была предложена и обоснована ранее [2]. Ее вид:

$$P(N) = p_{\max} - \frac{N / N_{\max}}{1 / p_{\max} - r(1 + N / N_{\max})},$$

где $p_{\max}$ — начальная эффективность (вероятность удвоения); $N/N_{\max}$ — относительное число копий, лежащее в интервале [0;1]; $N_{\max}$ — максимально достижимое количество копий в реакционном объеме; r = 0.5–0.9 — параметр, определяющий форму кривой.

В работе [3] была промоделирована типичная ситуация анализа образца, содержащего некоторую концентрацию целевых последовательностей ДНК. Анализ состоит из отбора пробы, постанов-

ки ПЦР-РВ с получением кинетической кривой, определения порогового цикла реакции и расчета количества копий целевой ДНК в отобранной пробе. Было показано, что при анализе небольших концентраций ДНК на погрешность результатов влияет стохастическая природа ПЦР-РВ. В то же время даже при анализе единичных копий это влияние несущественно по сравнению с вариацией, порождаемой случайным процессом при отборе пробы. Более того, при концентрациях, выше 100 копий на пробу, вклад стохастического процесса реакции в общую погрешность анализа практически не выявляется.

Таким образом, можно принять, что количество копий ДНК в каждой анализируемой пробе определяется с помощью ПЦР-РВ достаточно точно. Тогда основной задачей интерпретации результатов анализа является оценка концентрации ДНК в исследуемом образце по результатам определения количества копий ДНК в одной либо нескольких пробах, отбираемых из объема образца. Статистический анализ различных аспектов этой задачи является предметом данной работы.

ЗАДАЧА 1. СЛУЧАЙНЫЙ ОТБОР *A*-Й ЧАСТИ АНАЛИЗИРУЕМОГО ОБРАЗЦА В ПРОБУ

При анализе образца проводится отбор части его объема — пробы для проведения ПЦР-РВ и определения количества k копий ДНК (далее называемых частицами) в данной пробе. Детерминистический подход предполагал бы следующую процедуру: результат k количественного анализа отдельной пробы, объем которой равен $1/A$ части объема исходного образца, просто пропорционально увеличивался бы в A раз. Вместе с тем собственно процедура отбора объектов анализа (частиц) в пробу является случайным процессом, адекватно описываемым в рамках схемы последовательных простейших независимых испытаний (схема Бернулли), поскольку:

а) отбор каждой из частиц не зависит от результатов отбора других частиц, т. к. их взаимное влияние друг на друга пренебрежимо мало;

б) имеется лишь два результата (исхода): положительный (успех) — отбор в пробу с вероятностью $p = 1/A$, отрицательный (неудача) — непопадание в пробу (вероятность $q = 1 - p$);

в) число испытаний — число частиц n в исходном образце.

В этом случае вероятность того, что в пробу попали k частиц есть $C_n^k p^k q^{n-k}$, $k = 0,1,...,n$. Однако в нашем случае приходится решать обратную задачу, а именно при известном результате анализа отдельной пробы (k частиц) определить число порождающих частиц n, т. к. именно число частиц в исходном объеме, а не в одной отдельной анализируемой пробе является требуемой аналитической информацией. При дальнейшем рассмотрении число частиц n в исходном анализируемом объеме будем называть числом порождающих частиц (порождается соответственно k частиц в пробе).

ЗАДАЧА 2. ВЫЯВЛЕНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ ЧИСЛА ПОРОЖДАЮЩИХ ЧАСТИЦ

Условная (апостериорная) вероятность того, что k частиц порождены n частицами, определяется с помощью соотношения

$$\pi[n / k] = C_n^k p^{k+1} q^{n-k};\ n = k, k+1, ..., +\infty . \qquad (1)$$

То есть, если в пробе обнаружены k частиц, то $\pi[n / k]$ есть вероятность того, что они были порождены именно n частицами.

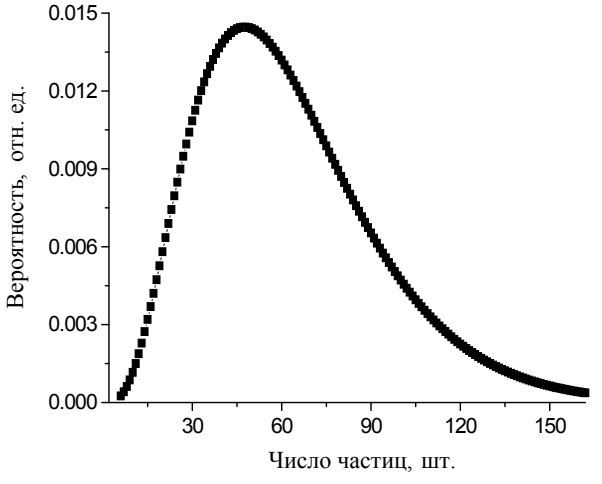
Зависимость строится на основе формулы апостериорной вероятности (формулы Байеса) в предположении о полной априорной неопределенности числа порождающих частиц (это может быть с равной вероятностью любое число частиц, не меньшее k). То есть принимается изначальная гипотеза: число частиц в исходном анализируемом объеме, породивших попадание k частиц в отбираемую пробу, может с равной вероятностью быть любым, не меньшим k. Данное утверждение носит название постулата Байеса или принципа равновероятности (закона недостаточного основания Лапласа): если о параметре ничего не известно, то в качестве априорного распределения принимают равномерное. Эта гипотеза обсуждалась, в частности, в [4].

Пример. Пусть, для определенности, анализируемый объем равномерно распределяется по $A = 16$ пробам (то есть $p = 1/16$). В одной из проб зарегистрировано 3 частицы. Требуется построить закон распределения числа частиц в исходном объеме, согласно (1). График апостериорной плотности вероятности числа порождающих частиц представлен на рисунке. Распределение имеет несимметричный характер, мода (вероятнейшее значение) не совпадает с математическим ожиданием. Интегральные характеристики распределения равны соответственно

$$\mathrm{M}\{n\} = (k+q)/p,\ \sigma^2\{n\} = q(k+1)/p^2, \qquad (2)$$

где $\mathrm{M}\{n\}$ — математическое ожидание, $\sigma^2\{n\}$ — дисперсия.

В нашем случае математическое ожидание равно 63, стандартное отклонение равно 30.98.



Плотность распределения вероятностей числа порождающих частиц n при k =3 порожденных частицах с вероятностью отбора p = 1/16

Тогда 90 %-му доверительному интервалу, вычисленному на основе вероятностей (1), соответствует диапазон числа порождающих частиц от 17 до 109, 50 %-му интервалу — от 43 до 83.

Определение наивероятнейшего числа порождающих частиц n проводится на основе системы двух неравенств аналогично нахождению вероятнейшего числа успехов по схеме Бернулли:

$$\frac{\pi[n/k]}{\pi[n-1/k]} \geq 1 \Rightarrow \frac{n}{n-k} q \geq 1 \Rightarrow n \leq k/p,$$

$$\frac{\pi[n/k]}{\pi[n+1/k]} \geq 1 \Rightarrow \frac{n-k+1}{(n+1)q} \geq 1 \Rightarrow n+1 \geq k/p.$$

Таким образом, $\frac{k}{p} - 1 \leq n \leq \frac{k}{p}$. Эта точечная оценка основывается на однократном измерении числа частиц k, обнаруженных в пробе. В нашем случае, модами (равновероятными порождающими условиями) будут 47 и 48 частиц.

Важнейшим частным случаем является порождение нуля частиц (истинно- или ложно-отрицательный результат анализа). При этом, согласно (2), математическое ожидание и дисперсия составят $\mathrm{M}\{0\} = 1/p - 1$, $\sigma^2\{0\} = q/p^2$.

Последние оценки непосредственно связаны с классической схемой испытаний до "первого успеха". Сдвиг математического ожидания на (–1) объясняется тем, что рассматривается не общее число проведенных испытаний, а число испытаний, предшествующих заключительному (успешному).

ЗАДАЧА 3. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ КОЛИЧЕСТВЕННОГО АНАЛИЗА НА ОСНОВЕ СЕРИИ НЕЗАВИСИМЫХ ИЗМЕРЕНИЙ ЧИСЛА ПОРОЖДЕННЫХ ЧАСТИЦ

В практическом применении регулярно проводится повторный анализ исследуемого образца путем отбора и количественного анализа нескольких проб. В этом случае на основе выборки измерений (результатов $k_1, k_2, \ldots, k_m$ анализа проб) строится соответствующая выборка оценок математических ожиданий числа порождающих частиц n (согласно (1)) и впоследствии доверительный интервал для математического ожидания. Тем самым получаем интервальную оценку числа порождающих частиц.

Определение наивероятнейшего числа частиц при неоднократном проведении анализа с результатами $k_1, k_2, \ldots, k_m$ может быть основано на методе максимального правдоподобия. Искомое число порождающих частиц n_0 — целая часть от решения алгебраического уравнения

$$\left(1 - \frac{k_1}{x}\right)\left(1 - \frac{k_2}{x}\right)\ldots\left(1 - \frac{k_m}{x}\right) = q^m,$$

$$x \in [n_0; n_0 + 1].$$

Приближенный поиск корня этого уравнения можно основывать на разложении логарифма левой и правой частей равенства в многочлен Маклорена второго порядка

$$\ln\left(1 - \frac{k_1}{x}\right) + \ln\left(1 - \frac{k_2}{x}\right) + \ldots + \ln\left(1 - \frac{k_m}{x}\right) \approx \approx -\frac{\sum_i k_i}{x} - \frac{\sum_i k_i^2}{2x^2}, \quad (3)$$

поскольку $\ln(1-p) \approx -p - p^2/2$.

Если A достаточно велико, то p и mp существенно меньше единицы и квадратичные слагаемые пренебрежимо малы. Величина x вычисляется только на основе первого слагаемого (3), т. е. как выборочное среднее k_i, а именно $x = \sum_i k_i / p$. Однако в общем случае оценка максимального правдоподобия несколько расходится с оценкой выборочного среднего. Эта оценка — точечная.

Например, пусть в трех лунках получены следующие результаты количественного ПЦР-анализа: k_1 = 3, k_2 =2, k_3 = 4. Отбор пробы проводился с A = 10. Оценка выборочного среднего трех измерений $\langle k \rangle$ = 3, из чего следует n = 30. Число частиц 30 — наиболее вероятное число порождающих частиц. Расчет по аппроксимирующей формуле (3) дает значение x, равное 30.101. В на-

шем случае из-за округления оценки максимального правдоподобия и выборочного среднего совпадут. В общем случае использованная оценка выборочного среднего будет иметь достаточно высокую эффективность, слабо расходясь с оценкой максимального правдоподобия.

Для получения интервальных оценок требуется учет случайных величин n_i — порождающих частиц, соответствующих k_i порожденным частицам. Каждая величина n_i имеет закон распределения (1) с различными в общем случае математическими ожиданиями и дисперсией, определяемыми (2). Величина доверительного интервала оценивается следующим образом: вычислением среднеквадратичного отклонения выборочного среднего с учетом независимости оценок числа порождающих частиц n_i, т. е. с суммированием соответствующих дисперсий, и квантиля распределения Стьюдента при известном числе степеней свободы m и заданной доверительной вероятности Q.

Математическое ожидание среднего арифметического оценок n_i есть $\frac{\langle k\rangle + q}{p}$, а ее среднеквадратичное отклонение составит $s\{n\} = \sqrt{\frac{q(\langle k\rangle + 1)}{mp^2}}$. Полуширина доверительного интервала оценки будет равна $s\{n\} * t(m,(1+Q)/2)$. Здесь $t(m,\alpha)$ — α-квантиль распределения Стьюдента с m степенями свободы.

В рассматриваемом случае используем 90 %-й доверительный интервал. Тогда α = 0.95 и при m = 3 квантиль равен 2.353 (соответствующий квантиль нормального распределения равен 1.645 [5]). Подстановка k_i дает: математическое ожидание равно 39, $s\{n\} = \sqrt{120} \approx 10.95$ и полуширина доверительного интервала 25.78. Таким образом, формальная оценка 90 %-го доверительного интервала есть [13.22; 64.78]. Очевидно, что как число порождающих частиц, так и границы являются целочисленными. Следовательно, границы интервала следует уточнить, а именно от 13 до 65 частиц.

Для сравнения, если те же результаты количественного анализа получены при двукратных повторах каждого измерения, т. е. значения 2, 3 и 4 получены при двух анализах каждый, то математическое ожидание и вероятнейшее число порождающих частиц, очевидно, не изменятся. Ширина доверительного интервала уменьшится не только благодаря удвоению числа степеней свободы (m = 6), но и в связи с уменьшением квантиля распределения Стьюдента до 1.943. В результате с учетом округления границ 90 %-й доверительный интервал составит [24; 54].

ЗАДАЧА 4. ИНТЕРПРЕТАЦИЯ ЛОЖНО-ОТРИЦАТЕЛЬНОГО РЕЗУЛЬТАТА НА ОСНОВЕ НЕСКОЛЬКИХ НЕЗАВИСИМЫХ ИЗМЕРЕНИЙ

Суть задачи состоит в следующем: одновременное измерение в m = 2, 3, …, 5 пробах дает отрицательный результат (ноль частиц). Как следует его интерпретировать, если принцип подбора пробы в лунку является случайным? Необходимо дать статистические оценки возможного исходного числа частиц в пробе, если осуществляется отбор $1/A$ ее части, и отрицательный результат независимо повторен в m пробах.

Ранее была выведена формула закона распределения числа порождающих частиц (1), которую следует применить для случая k = 0. При этом интерпретация понятий "успех" и "неудача", а также расчет соответствующих вероятностей требуют модификации.

Под "неудачей" будем понимать одновременное получение отрицательного результата во всех m пробах, "успехом" будет наличие хотя бы одного положительного результата. При доле объема пробы по отношению к исходному объему образца, равной $1/A$, вероятность неудачи $q = (1 - 1/A)^m$. Следовательно, $p = 1 - q$. Оценим предельное (максимальное) число частиц n^* в исходном анализируемом объеме, которое может порождать отрицательный результат одновременно в m пробах с заданной доверительной вероятностью $\eta = 1 - \alpha$. Традиционно выбирают α равной 5, 10, 25 и 50 %. Требуется вычислить сумму

$$\sum_{i=0}^{n^*} pq^i = p\frac{1-q^{n^*+1}}{1-q} = 1 - q^{n^*+1},$$

которая, в свою очередь, равна $\eta = 1 - \alpha$. Из данного равенства следует, что $q^{n^*+1} = \alpha$. Далее, применив логарифмирование и округление до целых значений, получим $n^* = \left[\ln(\alpha)/\ln(q)\right]$, где символ […] — целая часть числа.

Поскольку $q = (1 - 1/A)^m$, то при повторе измерений в m независимых пробах верхняя граница доверительного интервала $\eta = 1 - \alpha$ обратно пропорциональна m. С учетом возможных округлений и вычисления целой части эту закономерность следует считать приближенной, работающей в условиях $m << A$.

Для расчета математического ожидания числа порождающих частиц при m-кратных нулевых измерениях требуется использовать найденное выше значение q, рассчитать $p = 1 - q$ и воспользоваться формулой (2) при k = 0.

Максимальное число порождающих частиц n^* при m-кратном отрицательном результате анализа

m	Доверительная вероятность, %			
	50	75	90	95
2	17	34	56	74
3	11	22	37	49
4	8	17	28	37
5	6	13	22	29

Пример. Используем $A = 50$. Оценки среднего числа порождающих частиц n^* при 50, 75, 90 и 95 %-х доверительных вероятностях приведены в таблице. Число повторных независимых проб — от 2 до 5. Как видно из представленных данных, интерпретация повторных отрицательных анализов меняется.

Например, при четырехкратном достижении отрицательного результата нельзя достоверно утверждать, что исходный образец не содержит анализируемого объекта. Можно утверждать, что с вероятностью 90 % число частиц в исходном объеме не превосходит 28 (при отборе для каждого анализа 0.02 части). Математическое ожидание числа порождающих частиц будет 11.88. Следует отметить, что сформулированная выше закономерность $n^* \sim 1/m$ практически соблюдена.

В случае отбора в каждую пробу большей доли исходного объема (например, 0.03) все значения уменьшатся по сравнению с представленными в таблице величинами. Так, при $m = 4$ данные по доверительным верхним границам станут 5, 11, 18 и 24 частицы (вместо 8, 17, 28 и 38 соответственно).

ЗАКЛЮЧЕНИЕ

Предложенные в статье оценки, по нашему мнению, позволяют сделать более адекватной интерпретацию результатов количественного анализа ДНК методом ПЦР-РВ, учтя вероятностную (стохастическую) природу основных процессов, из которых определяющим является стадия отбора пробы.

На основании полученных оценок, предложен алгоритм интерпретации результатов ПЦР-РВ, описанный выше в задаче 3. Отметим, что он позволяет вычислить точечную оценку наиболее вероятного числа порождающих частиц на базе метода максимального правдоподобия. Однако вместо наиболее эффективной оценки (3) можно использовать более простую оценку, основанную на вычислении выборочного среднего от результатов анализа, имеющую малое отклонение от оценки максимального правдоподобия.

Также в задаче 3 описан и проиллюстрирован примером алгоритм нахождения доверительного интервала для числа порождающих частиц. Расчеты подтверждают ранее известные закономерности: а) для уменьшения ширины доверительного интервала требуется увеличить число независимых измерений (проб); б) при этом отношение объема исходного образца к объему пробы A должно быть уменьшено, насколько возможно, т. к. дисперсия оценки в первом приближении пропорциональна A^2, или, что то же самое, обратно пропорциональна p^2.

Особое значение имеет оценка отрицательных результатов анализа, полученных в одной либо нескольких пробах. Алгоритм оценивания математического ожидания числа порождающих частиц и верхней границы доверительного интервала описан выше в рамках задачи 4 и пояснения к данным таблицы. Эти оценки не зависят от способа обработки данных ПЦР-РВ и применимы к результатам работы с любыми тест-системами, необязательно количественными. Приведенный в задаче 4 пример соответствует типичной ситуации, когда из выделенного очищенного образца ДНК объемом 100 мкл берется проба объемом 2 мкл и проводится ПЦР. Допустим, образец ДНК получен из пробы воды объемом 1 мл. Если целью анализа является выявление возбудителя опасной инфекции с минимальной инфицирующей дозой 50 бактерий/мл, то отрицательный результат в трехкратной постановке ПЦР означает, что проба воды является опасной с вероятностью 5 %. Приведенная схема оценки доверительного интервала должна обязательно учитываться при разработке алгоритмов выявления инфекционных агентов в окружающей среде и клинических образцах.

Работа выполнена при финансовой поддержке Программ фундаментальных исследований Президиума РАН № 24 "Фундаментальные основы технологий наноструктур и наноматериалов" и № 9 "Создание и совершенствование методов химического анализа и исследования структуры веществ и материалов".

СПИСОК ЛИТЕРАТУРЫ

1. *Сочивко Д.Г., Федоров А.А., Курочкин В.Е., Петров Р.В.* Моделирование реакции амплификации ДНК в рамках теории ветвящихся процессов с двумя типами частиц // ДАН. 2010. Т. 434, № 2. С. 265–268.
2. *Сочивко Д.Г., Федоров А.А., Лавров В.В., Курочкин В.Е., Петров Р.В.* Стохастическое моделирование кинетических кривых полимеразной цепной реакции // ДАН. 2011. Т. 439, № 5. С. 696–699.
3. *Сочивко Д.Г., Федоров А.А., Варламов Д.А., Курочкин В.Е., Петров Р.В.* Точность количественного анализа ДНК с использованием полимеразной цепной реакции в реальном времени // ДАН. 2013. Т. 449, № 5. (В печати).
4. *Гуров С.И.* Интервальное оценивание на основе принципа согласованности // Вестник Тверского государственного университета. Серия "Прикладная математика". 2008. № 14, вып. 9. С. 77–93.
5. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. М.: Наука, 1983. 178 с.

Институт аналитического приборостроения РАН, г. Санкт-Петербург (*Буляница А.Л., Федоров А.А., Курочкин В.Е.*)

ЗАО "Синтол", г. Москва *(Сочивко Д.Г.)*

Контакты: *Сочивко Дмитрий Гарриевич*
sochivko@yahoo.com

Материал поступил в редакцию 14.03.2013

INTERPRETATION OF QUANTITATIVE REAL-TIME PCR ANALYSIS RESULTS BASED ON THE PROBABILISTIC MODELS

A. L. Bulyanitsa[1], D. G. Sochivko[2], A. A. Fedorov[1], V. E. Kurochkin[1]

[1]*Institute for Analytical Instrumentation of RAS, Saint-Petersburg*
[2]*Joint Stock Company "Syntol", Saint-Petersburg*

The present work develops our results obtained in recent studies [1–3] employing stochastic imitation modeling to analyze real-time PCR kinetics and DNA quantitative assay performance. In particular, PCR threshold cycle probability distribution depending on initial DNA copy number was estimated, reaction kinetics was studied, quantitative assay errors were analyzed. In the present paper, a number of topics related to the quantitative assay results interpretation is discussed: estimation of copy number (n) in the sample based on a series of quantitative assay results (k_i, $i = 1, 2,\ldots, m$) for a number of assay runs; obtaining estimates k_i; interpretation of negative or false-negative assay results in the case of multiple (m) runs with no copies found.

Keywords: real-time polymerase chain reaction, stochastic scheme, Bayes formula, quantitative analysis, false-negative result