12

Анализ капиллярно-электрофоретических систем секвенирования ДНК

© А. Степухович, А. Цуприк, О. Кособокова, Д. Гаврилов, Б. Горбовицкий, Г. Гудков, Г. Тышко, М. Черевишник, В. Горфинкель

Университет штата Нью-Йорк, Кафедра электроники и компьютерной техники, 11794 Стони Брук, США
e-mail: stepukhovich@mail.ru

(Поступило в Редакцию 5 июня 2007 г.)

Предложен метод характеризации и оптимизации систем секвенирования ДНК, основанных на капиллярном электрофорезе. Разработана модель, учитывающая взаимосвязь между длиной прочтения нуклеотидной последовательности и такими параметрами секвенирующих систем, как свойства разделяющей среды, длина капилляра, время инжекции, концентрация инжектированного образца ДНК, чувствительность детектирования, системный шум и мощность лазерного излучения. Оптимизация ДНК секвенирующих систем позволит определять наилучшие режимы пробоподготовки, инжекции и разделения, позволяющие прочитывать последовтельности требуемой длины. Предложенный метод применен для характеристики капиллярно-электрофоретических секвенирующих систем, использующих метод детектирования одиночных фотонов.

PACS: 82.45.Wx

Введение

Все используемые в настоящее время методы секвенирования ДНК можно подразделить на следующие основные группы [1,2]:

Электрофоретическое секвенирование является по сути классическим методом Сэнгера, основанным на реакции циклического секвенирования [3] (Sanger et al., 1977), с последующим капиллярным электрофорезом (КЭ). Опираясь на достижения в области полупроводниковых микротехнологий, исследовательские группы Матиса (Mathies) и BioMEMS разработали чип-устройство, интегрирующее амплификацию, очистку и секвенирование ДНК [4]. Использование подхода, основанного на успешных высокопроизводительных методах КЭ, позволило им добиться прочтения порядка 800 нуклеидов при сравнительно высокой степени точности [5].

Заметим, что основным преимуществом секвенирования ДНК при помощи КЭ является большая длина прочтения, достигающая 1000 и более нуклеотидных пар (base pairs, pb). К основным недостаткам данного метода можно отнести высокую стоимость, связанную с относительно невысоким числом одновременно обрабатываемых образцов (обычно такие секвенаторы ограничены несколькими сотнями каналов).

Гибридизационное секвенирование основано на гибридизации различных олигонуклеотидных зондов с целью восстановления последовательности ДНК-мишени. Существуют два подхода: первый заключается в иммобилизации ДНК-мишени; при втором — иммобилизуются олигонуклеотидные зонды (Affimetrix и Perlegen). Секвенирование путем гибридизации применяется как для повторного, так и для *de novo*-секвенирования [6–9]. К основным недостаткам данной технологии относятся довольно короткая длина прочтения (~ 25 bp), что связано с ограниченной длиной зонда [7], трудоемкая подготовка образцов, возможность кросс-гибридизации зондов из-за повторяющихся элементов или их случайного сходства. Последний фактор может явиться причиной значительной потери генетической информации, например, свыше 50% хромосомы 21 [9].

Параллельное циклическое секвенирование (cyclie array sequencing) амплифицированных и одиночных молекул ДНК основано на многократном повторении циклов ферментативных операций с пространственно разделенными фрагментами ДНК. В настоящее время ученые, использующие данные методы, идут по пути отказа от электрофореза [10–21]. В существующих версиях в каждом цикле ферментативных операций детектируется лишь один или несколько нуклеотидов; при этом одновременно обрабатывается от нескольких тысяч до нескольких миллионов фрагментов.

Очевидно, методы параллельного циклического секвенирования являются перспективными благодаря их умеренной стоимости. Однако, чтобы достить производительности, необходимой для *de novo*-секвенирования, необходимо увеличить длину прочтения последовательности, которая на данный момент не превышает 8 bp для метода FISSEQ [15] и 50 bp в случае пиросеквенирования [10]. Другая проблема, возникающая при использовании этих методов, — невысокая степень точности вследствие возрастающей потери синхронизации между матрицами (особенно в случае гомополимерных последовательностей).

За исключением КЭ, все вышеперечисленные технологии находятся на начальной стадии развития (стоит упомянуть также нециклические методы [22–24]).

Разработка этих технологий сопряжена с решением многочисленных научных и технологических проблем, на что может уйти 5–10 лет, прежде чем они будут запущены в производство. Следовательно, по крайней мере в ближайшем десятилетии, технологии, основывающиеся на капиллярном электрофорезе, останутся наиболее распространенными и востребованными.

Характеристика систем для секвенирования ДНК, основанных на КЭ

Для характеристики ДНК секвенирующих систем обычно применяются следующие основные параметры: длина прочтения, время прогона и чувствительность. Будучи удобными для пользователей, эти параметры, однако, непригодны для сравнительного анализа различных секвенирующих систем, поскольку ни один из них не имеет четкого определения. Вышеупомянутые параметры характеризуют систему секвенирования в целом, являясь зависящими как друг от друга, так и от многих других параметров системы.

В основанных на КЭ системах длина прочтения зависит от двух основных параметров: разрешения двух соседних групп фрагментов ДНК и качества распознавания последовательности. Если длина прочтения будет определена как максимальный размер фрагментов, которые можно разделить с разрешающей способностью R, большей, чем некое определенное значение [25,26], то эта длина будет зависеть от свойств среды разделения, режима секвенирования и дизайна детектирующей системы. Если длина прочтения определяется как максимальный размер фрагмента, который можно „прочитать“ с достаточным индексом качества $Q \geq 20$ [27,28], то она зависит от методов и алгоритмов обработки данных в дополнение ко всем вышеперечисленным параметрам системы. Кроме того, очевидно, что длина прочтения зависит от чувствительности метода обнаружения пика, так же как и от качества и эффективности электрокинетической инжекции образцов ДНК.

Не лучше дело обстоит и с определением чувствительности секвенирующих систем. Фактически общепринятая единица измерения чувствительности отсутствует как таковая. Поэтому, несмотря на очевидность зависимости длины прочтения от чувствительности, количественное соотношение между этими двумя параметрами в литературе отсутствует.

В настоящей статье предлагается модель, описывающая взаимосвязь длины прочтения, разрешения, чувствительности детектирования флуоресценции и параметров электрокинетической инжекции в системах секвенирования ДНК. Модель применена к основанной на КЭ системе, использующей метод счета фотонов [29], для оценки максимально возможной длины прочтения нуклеотидной последовательности и чувствительности системы.

1. Материалы и методы

1.1. Чувствительность и длина прочтения в системах секвенирования ДНК, основанных на КЭ

Чувствительность секвенатора определяется как минимальная (пороговая) концентрация маркированных фрагментов ДНК $n_{\mathrm{det\,th}}$, необходимых для обнаружения разделенных полос фрагментов, при соотношении сигнал/шум SNR выше $\mathrm{SNR_{th}}$. Это предполагает, что сигнал, обнаруженный в максимуме пика, должен удовлетворять

$$A_{\mathrm{det}} > \mathrm{SNR_{th}}\mathrm{Noise}. \qquad (1)$$

(Экспериментально было продемонстрировано, что достижение индекса качества $Q = 20$, с учетом разработанного нами программного обеспечения для распознавания последовательности, возможно при $\mathrm{SNR_{th}} \geq 4$).

Для характеристики чувствительности секвенатора необходимо измерить шум секвенирующей системы и определить сигнал ϕ от единичного фрагмента ДНК, регистрируемый системой на 1 mW лазерной мощности (P_{laser}). Используя полученное значение $n_{\mathrm{det\,th}}$, можно вычислить пороговую концентрацию маркированных фрагментов ДНК $n_{\mathrm{det\,th}}$ на максимальном участке выделенной полосы, необходимую для успешного обнаружения этой полосы:

$$n_{\mathrm{det\,th}} = \frac{A_{\mathrm{det\,th}}}{\phi P_{\mathrm{laser}} V_{\mathrm{det}}} \geq \frac{\mathrm{SNR_{th}\,Noise}}{\phi P_{\mathrm{laser}} V_{\mathrm{det}}}, \qquad (2)$$

где V_{det} — объем освещенного участка капилляра, в котором производится детектирование.

Чувствительность секвенатора ДНК, использующего метод счета фотонов

Основным шумовым компонентом в системах, основанных на детектировании одиночных фотонов, является, как правило, стохастический шум детектируемого потока фотонов [30]. Другой шумовой компонент ассоциируется с темновым фотоном детектора такого типа. Во всех детекторах, используемых в нашем секвенаторе (см. разд. 1.2), темновой фон был менее нескольких сотен импульсов в секунду, что на 1–2 порядка ниже типичного фонового сигнала, связанного с флуоресценцией разделяющей среды. Следовательно, этим шумом можно пренебречь. Для стохастического шума, ассоциируемого с потоком фотонов [30], соотношение сигнал–шум вычисляется следующим образом:

$$\mathrm{SNR} = \frac{\mathrm{Signal}}{\mathrm{Noise}} = \frac{ST_{\mathrm{int}}}{\sqrt{[(S+B)T_{\mathrm{int}}]}}, \qquad (3)$$

где S — количество импульсов в секунду, вызванное „полезной“ флуоресценцией фрагментов ДНК, B — количество импульсов в секунду, вызываемое естественной фоновой флуоресценцией, и T_{int} — время интеграции.

Естественный фон в секвенаторе возникает, главным образом, из-за флуоресценции полимера. К сожалению, выпускаемые промышленностью полиметры для КЭ (POP-5, POP-6, POP-7, Applied Biosystems, USA) флуоресцируют в том же спектральном диапазоне, что и большинство стандартных флуоресцентных маркеров (например, наборы BigDye, Applied Biosystems, USA), что делает этот фон неизбежным.

Величина детектируемого флуоресцентного сигнала зависит от нескольких факторов, а именно: длины волны и мощности источника света $P_{\text{laser}}(\lambda)$, эффективности сбора и эффективности фотодетектора $\varepsilon_{\text{det}}(\Delta\lambda)$ в спектральном диапазоне $\Delta\lambda$. Так как высота пика „полезной" флуоресценции уменьшается к концу типичного секвенсового прогона, фоновая флуоресценция становится основным источником шума в системе. Таким образом, с целью оценки минимальной величины детектируемых пиков, можно предположить, что $S \ll B$, и из соотношений (1) и (3) получим

$$a_{\text{det th}} = \text{SNR}_{\text{th}} \sqrt{b(\Delta\lambda)/(T_{\text{int}} P_{\text{laser}})}, \tag{4}$$

где a и b являются соответственно полезным и фоновым сигналами на 1 mW мощности лазера. Используя (1), (2) и (4), можно вычислить пороговую концентрацию фрагментов ДНК в максимуме детектируемой полосы

$$n_{\text{det th}} = \text{SNR}_{\text{th}} \frac{1}{\varphi V_{\text{det}}} \sqrt{\frac{\beta(\Delta\lambda)}{\varepsilon_{\text{det}} P_{\text{laser}} T_{\text{int}}}}, \tag{5}$$

где ϕ — сигнал, испускаемый единичным фрагментом ДНК на 1 mW мощности возбуждения P_{laser}, $\phi = \varphi(\Delta\lambda)\varepsilon_{\text{det}}(\Delta\lambda)$, а β — фоновый сигнал на 1 mW мощности возбуждения, $b = \beta(\Delta\lambda)\varepsilon_{\text{det}}(\Delta\lambda)$. Из соотношения (5) следует, что в секвенирующих системах, использующих метод счета фотонов, пороговая концентрация фрагментов ДНК, необходимая для обнаружения флуоресцентных полос с SNR_{th}, обратно пропорциональна квадратному корню из мощности лазера, времени интеграции и эффективности детектора.

Разрешающая способность секвенатора ДНК, основанного на КЭ

В качестве единицы измерения разрешения (R) нами используется отношение интервала между пиками ΔX и ширины пика W [25,26]:

$$R = \frac{2\Delta X}{W_1 + W_2} = \frac{t\Delta\nu}{W}, \tag{6}$$

где ΔX — расстояние между двумя соседними полосами, а W_1 и W_2 — ширина пиков на уровне полувысоты; $\Delta\nu$ — различие в скорости фрагментов в двух последующих полосах, t — время наблюдения, а W — ширина пика. Время наблюдения определяется следующим выражением: $t = L_{\text{det}}/\nu$, где L_{det} — длина капилляра до точки наблюдения. Ширина пика W вводится исходя из предположения, что для двух соседних полос $W_1 \approx W_2$. Селективность системы $\Delta\nu/\nu$ определяется миграционными механизмами фрагмента ДНК (см. [31]). Экспериментальным путем было показано, что для достижения индекса качества 20 нашему программному обеспечению требуется разрешение $R_{\text{th}} \geq 0.75$.

Длина прочтения в ДНК секвенирующей системе

Длина прочтения определяется как молекулярный размер K фрагментов ДНК, который может быть:

— разделен секвенирующей системой с разрешением R выше определенного порогового значения R_{th};

— зарегистрирован системой детектирования при соотношении сигнал–шум выше SNR_{th}

$$\begin{cases} R_K \geq R_{\text{th}} \\ A_{\text{det } K\,\text{th}} \geq \text{SNR}_{\text{th}}\,\text{Noise}. \end{cases} \tag{7}$$

Для определения максимально достижимой длины прочтения в определенной системе секвенирования ДНК при помощи (7) нам требуется определить SNR_{th} и R_{th}, необходимые для успешного распознавания пиков, оценить или измерить шум в системе, ввести модель эволюции ширины пика, интервала между пиками и амплитудами пика за период времени от электрокинетической инжекции до детектирования этого пика, а также связать разрешение R_K и амплитуду $A_{\text{det } K}$ пиков с характеристиками разделяющей среды и параметрами электрокинетической инжекции (см. (4) и (5)).

Электрокинетическая инжекция

В общем можно определить ширину K-й полосы фрагментов ДНК как

$$W_K = \int_0^{T_{\text{inj}}} v_{\text{inj}\,K}(t)dt, \tag{8}$$

где $v_{\text{inj}\,K}$ есть скорость растяжения полосы. По завершении электрокинетической инжекции фрагменты ДНК различного молекулярного размера K будут распределены вдоль капилляра с концентрацией $n_{\text{inj}\,K}(x)$. Для упрощения дальнейших вычислений предположим [25], что в момент инжекции группы фрагментов входят в капилляр с постоянной скоростью $\nu_{\text{inj}\,K}$ и образуют прямоугольные зоны с концентрацией $n_{\text{inj}\,K}$ и размерами W_K

$$W_K = v_{\text{inj}\,K} T_{\text{inj}}, \tag{9}$$

где E_{inj} — электрическое поле инжекции, а T_{inj} — длительность инжекции.

Электрофоретическое разделение

В разделяющей среде с момента инжекции до момента детектирования зоны фрагментов ДНК подвергаются

продольному уширению в силу ряда факторов. Предположим, что условия секвенирования выбраны таким образом, что уширение полос вызывается лишь диффузией; остальными механизмами, вызывающими уширение, можно пренебречь, поскольку они могут быть минимизированы выбором соответствующего режима разделения, геометрии капиллярных каналов и конфигурации системы детектирования [26]. В таком случае ширина отдельных полос в момент детектирования будет определяться первоначальной шириной инжекционных зон W_K и диффузионным уширением за период разделения t. Распространение полосы, содержащей фрагменты ДНК длиной в K bp, можно описать при помощи следующего уравнения:

$$\frac{\partial n_K}{\partial t} = \mu_K(E)\, E\, \frac{\partial n_K}{\partial x} + D_K \frac{\partial^2 n_K}{\partial x^2}, \qquad (10)$$

где n_K — концентрация фрагментов в полосе, μ_K — подвижность фрагментов, E — напряженность электрического поля, а D_K — коэффициент диффузии. Решить данное уравнение для инжекционной зоны, расположенной между $-W_K/2$ и $W_K/2$, можно следующим образом:

$$n_K(\chi, t) = \left\{ \int_{-W_K/2}^{W_K/2} \eta_{\mathrm{inj}\,K}(x-\chi)\, \frac{\exp[-(x-\chi)^2/4D_K t]}{\sqrt{4\pi D_K t}}\, d\chi \right\}, \qquad (11)$$

где $x = \chi - \mu_K \times E \times t$, а $\eta_{\mathrm{inj}\,K}(\chi)$ — количество фрагментов, инжектированных на единицу площади. Для прямоугольной инжекционной зоны при $\eta_{\mathrm{inj}\,K}(\chi) = \mathrm{const}$:

$$n_K(x, t) = \frac{n_{\mathrm{inj}\,K}}{2} \left[\mathrm{Erf}\left(\frac{W_K - 2x}{4\sqrt{D_K t}} \right) + \mathrm{Erf}\left(\frac{W_K + 2x}{4\sqrt{D_K t}} \right) \right]. \qquad (12)$$

С учетом (11) и (12) дисперсия пика σ^2 может быть определена таким образом

$$\sigma_K^2 = \frac{1}{n_K W_K} \int_{-\infty}^{\infty} x^2 n_K(x, t_K) dx = \frac{W_K^2}{12} + 2D_K t_K. \qquad (13)$$

При небольшой ширине инжекции W, когда $W^2/12 \ll 2Dt$, пик, описанный соотношением (12), становится гауссовым

$$n_K(x, t) = \frac{\eta_K}{\sqrt{2\pi\sigma_K^2}} \exp[-x^2/2\sigma_K^2]. \qquad (14)$$

Используя выражения (12)–(14), можно вычислить ширину пика W_{HK} (полная ширина на уровне полувысоты) в момент детектирования. Если форма пика не может быть описана моделью Гаусса, W_{HK} можно определить численно. Для гауссовых пиков воспользуемся следующим выражением:

$$W_{HK} = 2\sigma_K \sqrt{2 \ln 2}. \qquad (15)$$

Объединив (6) с (12) либо с (15), можно вычислить разрешение R как для точной модели диффузионного распространения (diffusion propagation model, DPM), так и для модели Гаусса (GM).

Согласно (7), для того чтобы вычислить длину прочтения, необходимо выявить зависимость высоты пика $A_{\mathrm{det}\,K}$ в момент детектирования t_K от ширины инжекции W_K и концентрации $n_{\mathrm{inj}\,K}$. Величина $A_{\mathrm{det}\,K}$ может быть вычислена путем „выравнивания“ общего числа фрагментов ДНК $N_{\mathrm{total}\,K}$ в полосе K в моменты инжекции и детектирования $t_K = L_{\mathrm{det}}/v_K$. Для инжектированной прямоугольной полосы

$$N_{\mathrm{total}\,K}(W_K, t_K) = n_{\mathrm{inj}\,K} W_K\, \mathrm{Area}_{\mathrm{cap}}\,. \qquad (16)$$

Прибегнув к соотношению (16) и принимая во внимание (12) или (15), вычислим высоту регистрируемых флуоресцентных пиков $A_{\mathrm{det}\,K}$ как для DPM, так и для GM. Для модели диффузионного распространения

$$N_{\mathrm{total}\,K}(W_K, t_K) = \int_{-\infty}^{\infty} n_K(x, t_K) dx\, \mathrm{Area}_{\mathrm{cap}}; \qquad (17)$$

для гауссова пика:

$$N_{\mathrm{total}\,K}(W_K, t_K) = n_{\mathrm{det}\,K} \times \sqrt{2\pi\sigma_K^2}\, \mathrm{Area}_{\mathrm{cap}}, \qquad (18)$$

где $n_{\mathrm{det}\,K}$ — концентрация ДНК в максимуме пика в момент детектирования t_K.

Обратившись к выражениям (18) и (5), вычислим общее пороговое число фрагментов ДНК в полосе, необходимое для детектирования полосы при $\mathrm{SNR}_{\mathrm{th}}$ для секвенирующих систем, использующих метод счета фотонов

$$N_{\mathrm{total}\,K\,\mathrm{th}} = \mathrm{SNR}_{\mathrm{th}} \frac{1}{\phi V_{\mathrm{det}}} \sqrt{\frac{2\pi\sigma_K^2 \times b(\Delta\lambda)}{P_{\mathrm{laser}} T_{\mathrm{int}}}}\, \mathrm{Area}_{\mathrm{cap}}\,. \qquad (19)$$

Высота пика $A_{\mathrm{det}\,K}$ для DPM и GM может быть найдена при помощи (4), (5) и (18).

Для GM высота пика $A_{\mathrm{det}\,K}$ может быть выражена через высоту исходного инжектированного прямоугольного пика $A_{\mathrm{inj}\,K}$ путем преобразования соотношений (16), (17) и (1):

$$A_{\mathrm{det}\,K} = \frac{n_{\mathrm{inj}\,K} P_{\mathrm{laser}} \phi V_{\mathrm{det}} W_K}{\sqrt{2\pi\sigma_K^2}} = \frac{A_{\mathrm{inj}\,K} W_K}{\sqrt{2\pi\sigma_K^2}} \geq \mathrm{SNR}_{\mathrm{th}}\, \mathrm{Noise}. \qquad (20)$$

Последнее неравенство в (20) указывает на то, что для любого данного размера фрагмента K и любой данной ширины инжекционной зоны W_K существует минимальная концентрация $n_{\mathrm{inj}\,K\,\mathrm{th}}$, которую необходимо инжектировать в капилляр для обнаружения пика

$$n_{\mathrm{inj}\,K\,\mathrm{th}} \geq \frac{\mathrm{SNR}_{\mathrm{th}} \mathrm{Noise}\, \sqrt{2\pi\sigma_K^2}}{P_{\mathrm{laser}} \phi V_{\mathrm{det}} W_K}. \qquad (21)$$

Соотношение (20) указывает на то, что при уменьшении ширины инжекционной зоны W_K необходимо обратно пропорционально увеличить концентрацию $n_{\mathrm{inj}\,K}$ инжектируемого образца. Если условие (21) выполнено, то

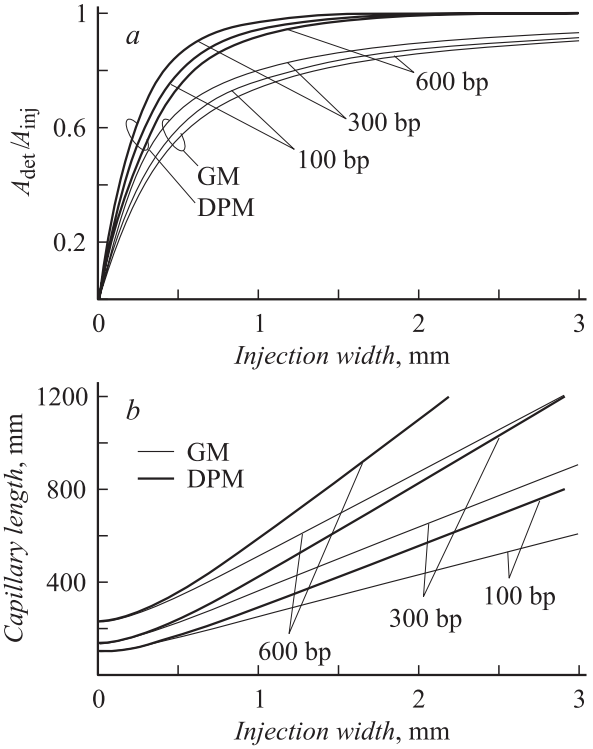

Рис. 1. A_{det}/A_{inj} (a) минимальная длина капилляра, обеспечивающая $R = 0.75$ (b) в зависимости от ширины инжекционной зоны, вычисленные для GM (тонкие линии) и DPM (жирные линии). POP-7, 50°C, $E = 120$ V/cm (остальные параметры вычисления приведены в табл. 1.)

для GM минимальная длина капилляра, необходимая для прочтения K bp, будет составлять

$$L_{det} = 2R_{th}\frac{\nu_K}{\Delta\nu_K}\sqrt{2\ln 2\left(\frac{W_K^2}{12} + 2D_K t_K\right)}. \qquad (22)$$

При $W^2/12 \gg 2Dt$ гауссова модель неприменима, что заставляет нас прибегнуть к DPM для вычисления длины капилляра.

С учетом (9)–(12) для данной разделяющей среды и мощности возбуждающего источника система (7) может быть представлена в следующем виде:

$$\begin{cases} R_K(D_K, W_K, L_{det}) \geq R_{th} \\ A_{det\,K}(n_{inj\,K}, D_K, W_K, L_{det}) \geq \mathrm{SNR_{th}} \times \mathrm{Noise}(A_{det\,K}, B) \end{cases} \qquad (23)$$

Система (23) позволяет оптимизировать условия инжекции и длину капилляра, необходимые для распознавания K-й полосы. Важно отметить, что для обеих моделей концентрация фрагментов в максимуме пика в момент детектирования n_{det} меньше инжектируемой концентрации n_{inj}. Поэтому высота детектируемого типа A_{det} всегда меньше высоты инжектируемого пика A_{inj}.

На рис. 1, a проиллюстрирована зависимость величины A_{det}/A_{inj}, вычисленной при помощи (20) и (22), от ширины зоны инжекции для минимальной длины капилляра, обеспечивающей разрешение $R = 0.75$, а также длина капилляра, необходимая для достижения такого разрешения (рис. 1, b).

Как видно из рис. 1, для широкой инжекционной зоны при использовании GM возникает погрешность в оценке приблизительно 10–20% от высоты пика A_{det} и длины капилляра. Поэтому эта модель может быть использована для приблизительного расчета длины прочтения и производительности систем секвенирования ДНК. Отношение A_{det}/A_{inj} (рис. 1, a) возрастает по мере увеличения времени инжекции. Эта зависимость носит линейный характер для $W^2/12 \ll 2Dt$. Для $W_K \sim 2D_K t$ она становится сублинейной и достигает предела при $A_{det\,K} = A_{inj\,K}$. Результаты, полученные при помощи обеих моделей, указывают на то, что после того как инжекционная зона достигает определенной ширины $W_{K\,sat}$, любое дальнейшее увеличение W_K не приводит к сколь бы то ни было заметному увеличению высоты пика ($W_{K\,sat} \approx 1$ mm, рис. 1). Приведенный выше анализ свидетельствует о том, что в действительности существуют два ограничения на разделение в секвенирующих системах, основанных на КЭ: диффузионно ограниченное (ДО) разделение и инжекционно ограниченное (ИО) разделение.

Диффузионно ограниченное разделение возможно лишь при очень высокой концентрации образца ДНК в инжекционной зоне n_{inj} ($n_{inj} \gg n_{det\,th}$), такой что даже для очень коротких инжекционных зон n_{det} будет превышать $n_{det\,th}$. В этом случае длина разделения–прочтения будет определяться лишь диффузионным уширением полос. Режим ДО-разделения требует наименьших длины капилляра и времени разделения (см. рис. 2, штриховая линия) и поэтому обеспечивает наибольшую возможную производительность системы секвенирования.

Режим инжекционно ограниченного разделения может использоваться при секвенировании образцов с низким содержанием ДНК ($n_{inj} \to N_{det\,th}$). При работе с такими образцами с целью поддержания высоты пика и разрешения выше необходимого порога (система (23)) необходимо увеличивать как ширину инжекционной зоны W_K,

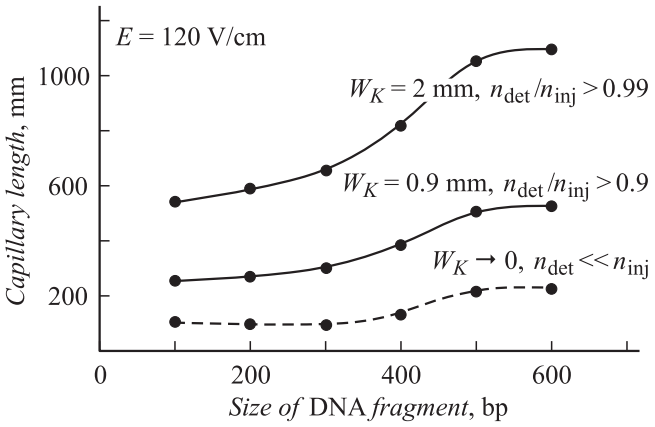

Рис. 2. Зависимость длины капилляра от размера фрагмента ДНК, вычисленная для трех значений ширины инжекции с использованием модели DPM (штриховые линии соответствуют ДО-разделению).

так и длину капилляра, по сравнению с режимом ДО-разделения. Для приблизительного расчета необходимой длины капилляра можно воспользоваться соотношением (16). Более точное вычисление длины капилляра можно произвести, используя (16) совместно с (2). В случае, когда W_K достигает $W_{K\,\mathrm{sat}}$, отношение $A_{\mathrm{det}}/A_{\mathrm{inj}}$ приблизительно равно единице, n_{det} приближается к n_{inj}, а высота пика A_{det} достигает максимального значения (см. рис. 1).

Как видно из рис. 2, минимально необходимая длина капилляра может весьма заметно варьироваться при переходе режима разделения от ДО к ИО. Таким образом, если концентрация инжектируемого материала ДНК $n_{\mathrm{inj}\,K}$ низкая, но все же выше $n_{\mathrm{det\ th}\,K}$, то можно вычислить ширину инжекционной зоны и длину капилляра, которые позволят осуществить как разделение, так и детектирование данного образца ДНК с необходимыми для данной секвенирующей системы разрешением и SNR. Очевидно, образцы с $n_{\mathrm{inj}\,K} < n_{\mathrm{det}\,K\,\mathrm{th}}$ невозможно секвенировать при помощи такой системы.

1.2. Описание секвенатора

В настоящей реботе использовался разработанный нашей группой [29] однокапиллярный секвенатор, основанный на детектировании флуоресценции методом счета фотонов. Схема и фотография секвенатора (модель SBS-2004) представлены на рис. 3.

Маркированные образцы ДНК подвергаются разделению в однокапиллярном разделительном модуле. Этот модуль содержит миниатюрный источник высокого напряжения (до 15 kV) со встроенным вольтметром и микроамперметром, систему замещения полимера, систему контроля температуры, позволяющую поддерживать температуру в капилляре в диапазоне 25–70°C с точностью до 0.01° (температурный контроллер 5C7-378, Mc.Shane Inc., USA), карусель для смены пробирок, содержащую образцы ДНК, воду, рабочий буфер и высокоточную оптоволоконную оптическую систему, разработанную авторами (см. *2* на рис. 3). В качестве источника возбуждающего света используется оптоволоконный аргон-ионный лазер (488 и 514 nm, EverGreen, USA), работающий в непрерывном волновом режиме. Флуоресценция собирается оптоволоконным микрообъективом (C 230260P-B, Thorlabs Inc., USA) и передается на оптоволоконный детектор, работающий в режиме счета фотонов.

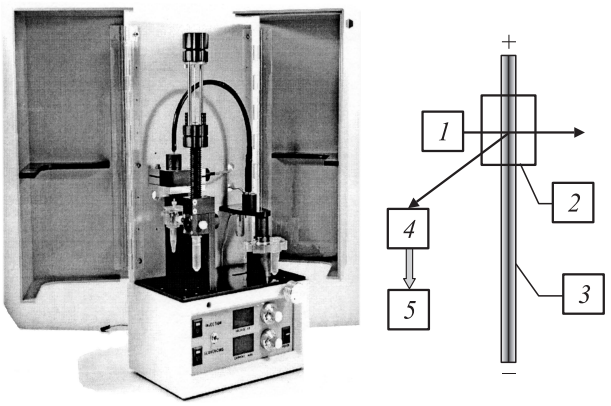

Рис. 3. Схема и фотография секвенатора ДНК (SBS-2004). Прибор состоит из лазерного источника (*1*), однокапиллярного разделительного модуля (фотография) со встроенной оптической системой (*2*), системой замещения полимера и системой контроля температуры (*3*), а также фотодетектора (*4*) и компьютера (*5*).

Флуоресцентный сигнал, отфильтрованный от лазерного света при помощи узкополосного режекторного либо ступеньчатого фильтра, подается на вращающееся колесо, содержащее четыре полосовых фильтра (10–20 nm, 3–4 OD OMEGA Optical, USA), соответствующих максимальной эмиссии четырех используемых красителей. После фильтрации флуоресцентный сигнал регистрируется оптоволоконным детектором, использующим PMT (H7464, H6240, Hamamatsu, Japan) или APD (SPCM-AQ4C, Perkin Elmer, USA). Накопленный за время прохождения света через каждый из четырех фильтров сигнал интегрируется и дает, таким образом, амплитуду в четырех каналах. Эти четыре значения передаются на компьютер во время каждого обращения колеса с фильтрами ($\sim$ 10–25 оборотов в секунду). По завершении этапа секвенирования обработка данных осуществляется либо на том же компьютере, либо на любом другом компьютере, подключенном к лабораторной сети.

Пакет программ для обработки данных

Нами разработан полный пакет программного обеспечения для обработки секвенсовых данных [29]. Пакет включает отдельные модули для записи данных (MONITOR), обработки и визуализации (BASE). MONITOR принимает данные, передаваемые посредством IEEE-1284 Parallel Interface, выполняет их визуализацию в реальном времени и запись в двоичный „сырой“ файл на жестком диске. Модуль обработки данных начинается по команде, принимая „сырой“ файл на входе, а на выход поступают файлы с обработанными данными, полностью совместимые со стандартной программой PHRED [27,28] (<www.phrap.org>). Процедура обработки включает в себя обнаружение и устранение ошибок, возникающих при передаче и записи данных, сглаживание, вычет фоновой флуоресценции, фильтрацию и выравнивание интервалов между пиками. Обработанные данные подвергаются повторной обработке с целью получения от 7 до 15 точек на пике и затем сохраняются в формате SCF. Для распознавания последовательности нуклеотидов и оценки качества секвенирования применяется коммерческий пакет PHRED. Модуль BASE позволяет осуществлять просмотр, ручную обработку, редактирование и распечатку как необработанных, так и готовых данных.

1.3. Реагенты и условия секвенирования

В качестве тестовых образцов ДНК использовались „линейки“: Fluorescein Ruler (FR, 100–1000 bp, Bio-Rad, USA) и Internal Lane Standard 600 (ILS-600, 60–600 bp, Promega, USA), а также стандартная последовательность, маркированная BigDye (BigDye Terminator v1.1 Sequencing Standard Kit, ~ 1500 bp, Applied Biosystems, USA). Перед инжекцией FR и $2\,\mu$l образца, денатурированного при 95°C в течение 3 min и затем остуженного на льду, добавлялось $18\,\mu$l формамида (Di-Hi Formamide, Applied Biosystems, USA). Образцы ILS-600 и Big Dye Sequencing Standard были проготовлены, согласно инструкции производителей. При проведении некоторых экспериментов ILS-600 и FR были скомбинированы для получения „смешанной линейки“. Все образцы инжектировались электрокинетически в капилляры (общая длина — 56 cm, длина детектирования L_{det} — 50 cm, $50/365\,\mu$m внутренний/внешний диаметр (PolyMicro Technologies, USA)), заполненные полимером POP-7. Секвенирование проводилось при рабочем напряжении 6.8–11.2 kV ($E = 120$–200 V/cm) и температуре 50°C.

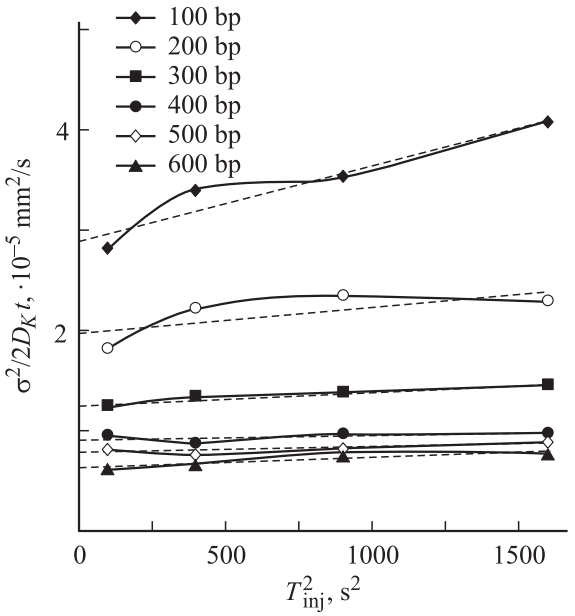

Рис. 4. Зависимость дисперсии пика для фрагментов 100–600 bp от квадрата времени инжекции. POP-7, 50°C, $E = 120$ V/cm.

2. Обсуждение результатов

2.1. Характеризация разделяющего полимера и применимость моделей

Характеризация разделяющего полимера

Чтобы применить модель, представленную выше, к нашему секвенатору ДНК, необходимо вычислить коэффициент диффузии D_K, скорость фрагментов ν_K при заданном рабочем напряжении, скорость в момент инжекции $\nu_{\text{inj}\,K}$ и селективность $\Delta\nu_K/\nu_K$, зависящие от молекулярного размера фрагментов ДНК.

Все необходимые величины можно получить, используя секвенсовые данные, зависимость дисперсии пика σ^2 от времени инжекции в квадрате T^2_{inj}. Из соотношений (9) и (13) получаем

$$\sigma_K^2 = \frac{(\nu_{\text{inj}\,K} T_{\text{inj}})^2}{12} + 2D_K t. \qquad (24)$$

Как следует из выражения (24), отрезок на оси Y, отсекаемый кривой, описывающей эту зависимость, соответствует коэффициенту диффузии, а наклон прямых линий представляет собой скорость K-го фрагмента ДНК в момент инжекции T_{inj}, возведенную в квадрат $\nu^2_{\text{in}\,K}$.

Чтобы определить эти величины экспериментально, мы провели секвенирование тестовых образцов ДНК при 50°C: для ограничения уширения пика, связанного с термальным градиентом [26], рабочее электрическое поле было в пределах 120–200 V/cm. Для определения коэффициентов диффузии для фрагментов 100–600 bp был использован образец ILS-600. Коэффициенты диффузии для фрагментов 700–1000 bp были получены с помощью образца FR. Так как ILS-600 содержит только отдельные фрагменты, отстоящие друг от друга на 20–25 bp, и позволяет получить на электроферрограмме четкие индивидуальные пики, нам удалось применить (24) для извлечения коэффициентов диффузии. К сожалению, электроферрограммы, полученные при использовании FR, содержали „двойные“ пики, а каждый пик дублета состоял из нескольких отчетливых небольших пиков. Поэтому образец FR был использован лишь для оценки скорости фрагментов.

Коэффициенты диффузии для фрагментов размером 700–1000 bp были определены с использованием диффузионной теории, описанной в [25,32]. Известно, что коэффициент диффузии прямо пропорционален электрофоретической подвижности

$$D_K = \frac{\mu_K k_{\text{boltz}} T}{Q_K X}, \qquad (25)$$

где Q_K — полный заряд рассматриваемой молекулы, а X — безразмерный фактор, корректирующий частичную нейтрализацию ионами противоположного знака. Взяв отношение D_K и D_{600}, нейденное экспериментальным путем, и предположив, что фактор X не зависит от размера фрагмента, получим

$$D_K = D_{600}\,\frac{\mu_K Q_{600}}{\mu_{600} Q_K}. \qquad (26)$$

На рис. 4 представлены зависимости дисперсии пика $\sigma^2/2D_K t$ от T^2_{inj} для фрагментов 100–600 bp, инжектированных при 1.5 kV и разделенных при 6.8 kV в капилляре длиной 56 cm. Параметры, полученные для

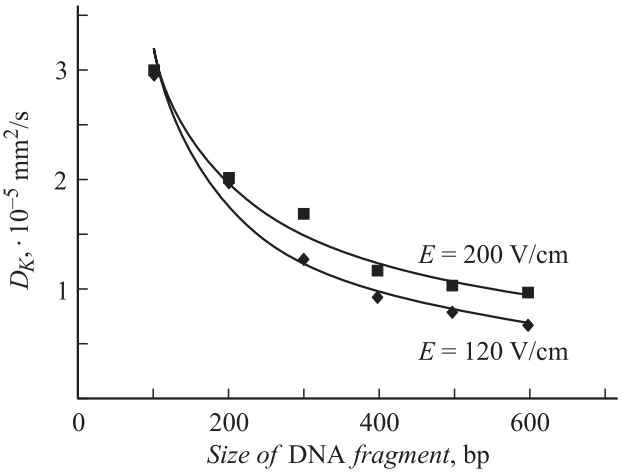

Рис. 5. Зависимость коэффициента диффузии от размера фрагмента ДНК для двух электрических полей. POP-7, 50°C.

Таблица 1. Распространение фрагментов ДНК в POP-7 ($E = 120$ V/cm, 50°C)

bp	t, s	v, mm/s	Δv, mm/s	D, mm^2/s
100	2000	0.25	0.000671	0.0000291
200	2631	0.19	0.000494	0.0000199
300	3333	0.15	0.000358	0.0000128
400	3909	0.1279	0.000228	0.0000094
500	4545	0.11	0.00016	0.0000081
600	5050	0.099	0.000129	0.0000066
700	5555	0.09	0.000108	0.0000051
800	5966	0.0838	0.0001	0.0000041
900	6288	0.079505	0.000094	0.0000033
1000	6666	0.075	0.000088	0.0000028

$E = 120$ V/cm, и зависимости коэффициентов диффузии от размера фрагментов приведены на рис. 5 и в табл. 1.

Применимость модели пика

Чтобы проверить достоверность предложенной модели пика, нами были проведены измерения и теоретические расчеты зависимости амплитуды пика от времени инжекции для ILS-600 (см. рис. 6, *a*). Результаты, полученные экспериментальным путем (пунктир) и при помощи расчетов (сплошные линии), демонстрируют хорошее соответствие (см. также рис. 1). Для всех длин фрагментов обе кривые демонстрируют линейный рост амплитуды пика для короткого времени инжекции, за которым наступает насыщение при инжекции более $\sim 80-100$ s.

Используя полученные параметры разделяющего полимера, мы оценили применимость гауссовой модели для описания формы пиков. На рис. 6, *b* представлены зависимости диффузионной и инжекционной составляющих дисперсии пика от времени инжекции. Если предположить, что GM справедлива, если $W^2/12$ составляет менее 10% дисперсии пика, тогда, как видно из графиков, эта модель применима лишь при инжекционном времени меньше ~ 10 s (длина инжекционной зоны между 175 и 75 μm для фрагментов ДНК 100 и 600 pb соответственно). С другой стороны, GM довольно качественно описывает поведение пика для гораздо большей ширины зоны инжекции (см. рис. 6, *a*, а также рис. 1), и она может быть использована для приблизительного расчета длины прочтения и времени секвенирования с точностью до 10–20%.

2.2. Характеристика чувствительности секвенатора

Определение $a_{\text{det th}}$ и $n_{\text{det th}}$

Наши экспериментальные данные свидетельствуют о том, что индекс качества 20 может быть достигнут при $\text{SNR}_{\text{th}} \geq 4$ в максимуме пика, а также о том, что для сохранения исходной формы и высоты пика время интеграции T_{inj} должно быть меньше 1/10 ширины пика. В нашем секвенаторе T_{inj} может варьироваться в пределах $0.005-0.1$ s (обычно $T_{\text{inj}} = 0.025$ s). Как видно из уравнений (4) и (5), чтобы вычислить $\alpha_{\text{det th}}$ и $n_{\text{det th}}$, необходимо измерить параметры ϕ и b.

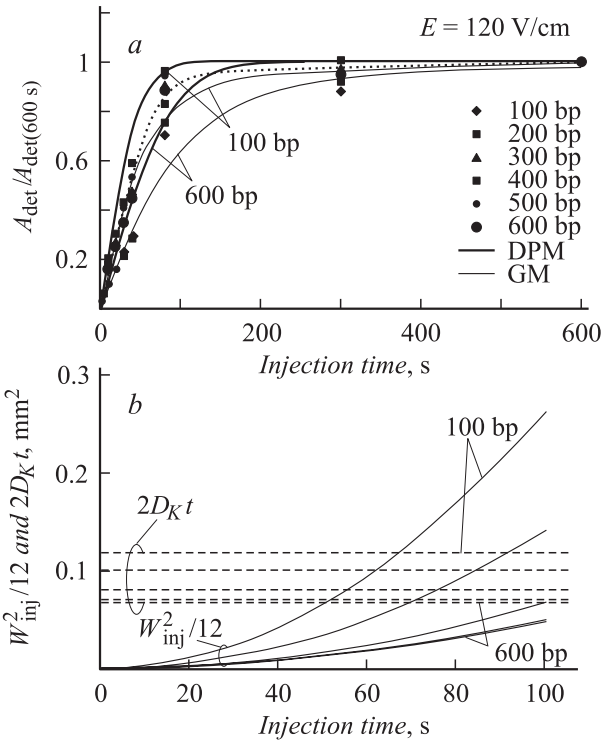

Рис. 6. Зависимость амплитуды пика (*a*): амплитуда пика нормирована на величину, полученную при инжекции длительностью 600 s и составляющих дисперсии пиков — диффузионной и инжекционной (*b*) — от времени инжекции, рассчитанного для $L_{\text{det}} = 50$ cm.

Таблица 2. Экспериментальный фоновый сигнал для PMT и APD детекторов

Тип детектора	b, $s^{-1}mW^{-1}$ (610 ± 5 nm, 'C')	b, $s^{-1}mW^{-1}$ (580 ± 5 nm, 'T')	b, $s^{-1}mW^{-1}$ (560 ± 5 nm, 'A')	b, $s^{-1}mW^{-1}$ (540 ± 5 nm, 'G')
Охлажденный модуль APD APSPCM-AQ4C	6.7	4	2.6	2
PMT H7467-01	737	548	420	360

Определение ϕ

Для оценки сигнала ϕ, испускаемого единичным фрагментом ДНК на 1 mW мощности возбуждения в нашем секвенаторе, мы использовали величину ϕ_{fl} — сигнал, регистрируемый от одной молекулы флюоресцеина на 1 mW мощности возбуждения в диапазоне 540 ± 5 nm. Чтобы измерить ϕ_{fl}, сериальные разведения красителя флуоресцеина в буфере TSR (Applied Biosystems, USA) были последовательно помещены в капилляр, закрепленный в считывающей головке секвенатора, после чего флуоресцентный сигнал регистрировался детектором SPCM-AQ4C (рис. 7).

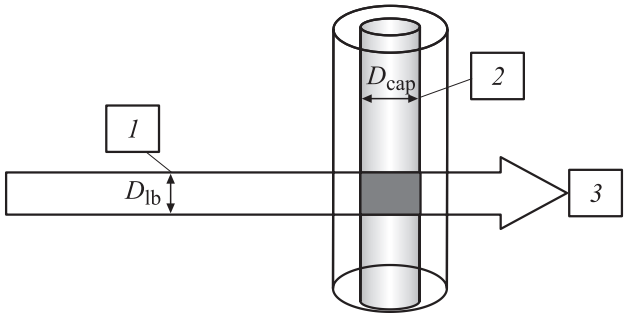

Рис. 7. Схема измерений сигнала φ_{fl}. *1* — луч лазера, D_{lb}, *2* — капилляр, D_{cap}, *3* — фотодетектор.

Величина ϕ_{fl}, соответствующая одной молекуле флюоресцеина, определяется следующим образом:

$$\phi_{\text{fl}} = \frac{r_{CM}\, B}{P_{\text{laser}} M}, \quad \text{где} \quad M = C_M 6 \cdot 10^{23}\, \frac{\pi D_{\text{lb}}^2}{4}\, D_{\text{cap}}. \quad (27)$$

Здесь r_{CM} — сигнал, соответствующий раствору известной молекулярной концентрации C_M, B — фоновая флуоресценция буфера TSR бездобавления флюоресцеина, P_{laser} — мощность возбуждения, M — количество молекул флюоресцеина, освещенных лазерным лучом диаметром D_{lb} в капилляре внутреннего диаметра D_{cap}. Измерения показывают, что для $D_{\text{cap}} = 50$ и $D_{\text{lb}} = 20\,\mu$m сигнал ϕ_{fl}, соответствующий одной молекуле флюоресцеина, составлял 20 импульсов в секунду на 1 kW мощности возбуждения.

Вычисление b

В табл. 2 представлены значения фоновой флуоресценции в нашем секвенаторе, измеренной двумя различными детекторами в четырех спектральных диапазонах $\Delta\lambda$, используемых для детектирования маркеров BigDye для четырех типов фрагментов, оканчивающихся нуклеотидами C, T, A и G (POP-7, при 50°C и возбуждении аргоновым лазером мощностью 1 mW).

Чувствительность секвенатора

Чувствительность секвенатора может быть вычислена с помощью соотношений (4) и (5) с учетом $A_{\text{det th}} = P_{\text{laser}} a_{\text{det th}}$ (рис. 8). Как видно, пороговая высота пика $A_{\text{det th}}$ пропорциональна квадрату мощности возбуждения для заданного отношения сигнал–шум. Кроме того, пороговая концентрация в максимуме пика $n_{\text{det th}}$ уменьшается согласно той же зависимости (обратите

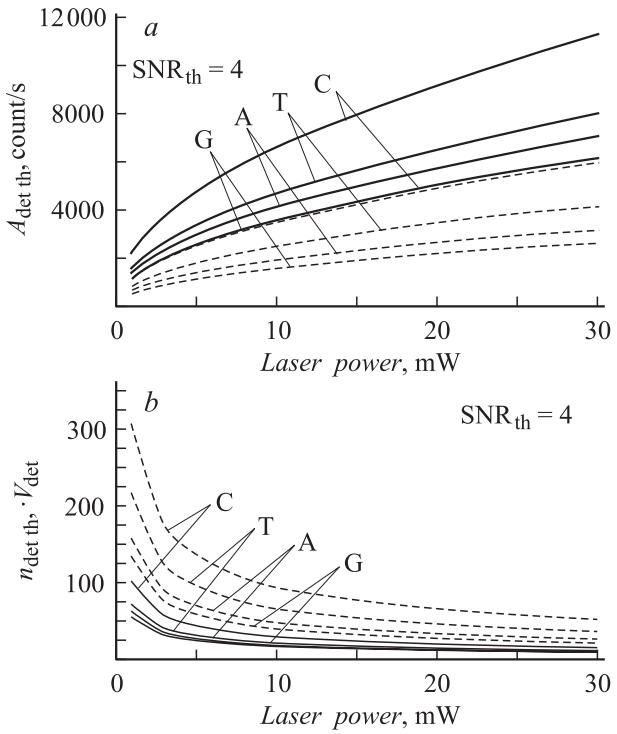

Рис. 8. Зависимость пороговой высоты пика (a) и пороговой концентрации четырех типов фрагментов ДНК (оканчивающихся соответственно нуклеотидами C, T, A, G) в максимуме пика (b) от мощности возбуждающего лазерного излучения: APD-детектор SPCM AQ4C — сплошные линии, PMT-детектор H7467-01 — штриховые линии; для вычислений использовались данные табл. 1 и 2, а также уравнения (4), (5); $T_{\text{int}} = 0.025$ s, $\text{SNR}_{\text{th}} = 4$.

Таблица 3. Чувствительность секвенатора ДНК (модель SBS-2004) при $P_{\text{laser}} = 30\,\text{mW}$, $E = 120\,\text{V/cm}$

$R = 0.75$	$N_{\text{det th}}/V_{\text{det}}$, APD-детектор	$N_{\text{det th}}/V_{\text{det}}$, PMT-детектор
C	19	56
T	13	39
A	12	30
G	10	24

внимание, что $N_{\text{det th}} = n_{\text{det th}} V_{\text{det}}$). Экспериментально было показано, что максимальная мощность возбуждения, не приводящая к перегреву капилляра в нашей системе, составляет 25–30 mW. Поэтому наивысшая чувствительность нашего секвенатора достигает при $P_{\text{laser}} = 30\,\text{mW}$. В табл. 3 представлено минимальное количество маркированных фрагментов ДНК $N_{\text{det th}}$ в области детектирования V_{det}, необходимое для детектирования пика при $\text{SNR} = 4$.

Как видно из рис. 8, *b* и табл. 3, существует зависимость чувствительности системы от типа фотодетектора. В зависимости от типа флуоресцентного маркера пороговая концентрация фрагментов ДНК $n_{\text{det th}}$ в 2–3 раза выше для PMT-детектора по сравнению с APD-детектором. Это различие вызвано более низкой эффективностью PMT-детектора по сравнению с APD (в диапазоне 540–610 nm эффективность наилучшего PMT H7467-01 в 5.5–9 раз ниже, чем эффективность наилучшего охлажденного APD SPCM-AQ4C). Значительное различие в $A_{\text{det th}}$ и $n_{\text{det th}}$, полученное для четырех маркеров BigDye, связано с различием как в уровне фона (см. табл. 2), так и в спектральной чувствительности названных детекторов.

Полученные пороговые концентрации $n_{\text{det th}}$ позволяют определить общее число фрагментов ДНК $N_{\text{total th}\,K}$, необходимое для детектирования K-й группы фрагментов ДНК в режимах ДО- и ИО-разделения. На рис. 9 представлены зависимости $N_{\text{total th}}$ от размера фрагментов, вычисленные при помощи (17) для $L_{\text{det}} = 50\,\text{cm}$. Для вычисления $N_{\text{total th}\,K}$ была определена максимальная ширина зоны инжекции $W_K(R_{\text{th}})$, которая позволяет достичь необходимой разрешающей способности R_{th} при использовании капилляра данной длины L_{det}. С учетом полученных значений W_K было вычислено $N_{\text{det th}\,K}$ для режима ИО-разделения. Значения $N_{\text{total th}\,K}$ для режима ДО-разделения были получены из соотношения (18), где положено $W_K = 0$. Как видно, $N_{\text{det th}\,K}$ в 1.5–2 раза больше для режима ДО-разделения. Однако, как следует из (14), в зависимости от соотношения ширины зоны инжекции в режимах ДО и ИО для ДО-разделения может потребоваться гораздо большее $n_{\text{inj}\,K}$, чем при ИО-разделении. Поэтому для образцов с низкой концентрацией ДНК режим ОИ-разделения может оказаться единственной возможностью для достижения успешного распознавания последовательности.

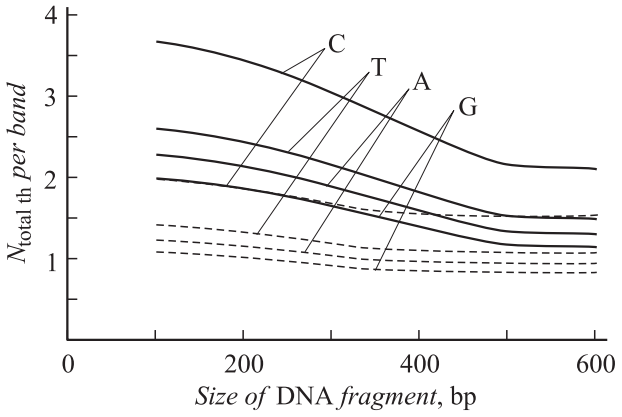

Рис. 9. Зависимости $N_{\text{total th}}$ от размера фрагмента ДНК для режимов разделения ДО (пунктирные) и ИО (сплошные линии), вычисленные для APD-детектора. Для вычислений использовались данные табл. 3; $\text{SNR}_{\text{th}} = 4$; $R_{\text{th}} = 0.75$.

2.3. Вычисление размера инжектированных зон по данным секвенирования

Разработанный нами подход позволяет при помощи полученных экспериментальным путем секвенсовых данных вычислить ширину зоны инжекции W_K и концентрацию ДНК $n_{\text{inj}\,K}$ в ней. Эта информация может оказаться весьма полезной для оптимизации процесса приготовления образца, а также для оценки ожидаемой длины прочтения для образцов ДНК, приготовленных стандартным способом, но содержащих малую концентрацию фрагментов ДНК (например, образцы ДНК с высоким молекулярным весом).

Рассмотрим один пик электроферограммы, соответствующий определенному размеру фрагмента K и имеющий амплитуду $A_{\text{det}\,K}$ и дисперсию σ_K^2. Ширина зоны инжекции как для GM, так и для DPM может быть вычислена следующим образом:

$$W_K = \sqrt{12(\sigma_K^2 - 2D_K t)}. \tag{28}$$

Очевидно, что точность определения W_K зависит от точности определения (экспериментально) коэффициента диффузии δD_K и дисперсии $\delta\sigma_K^2$. Согласно (28), относительная погрешность в определении ширины пика будет составлять

$$\frac{\delta W_K}{W_K} = \pm\frac{|\delta\sigma_K^2| + 2t|\delta D_K|}{2(\sigma_K^2 - 2D_K t)\sqrt{12(\sigma_K^2 - 2D_K t)}}. \tag{29}$$

Следует особо отметить, что неточность в определении ширины пика может быть особенно значительной в режиме ДО-разделения, в случае которого ширина пика значительно меньше, чем расплывание пика вследствие диффузии, а σ_K^2 примерно равно $2D_K t$. Для более точного определения ширины инжекционной зоны вместо соотношения (29) можно обратиться к (11) и попытаться

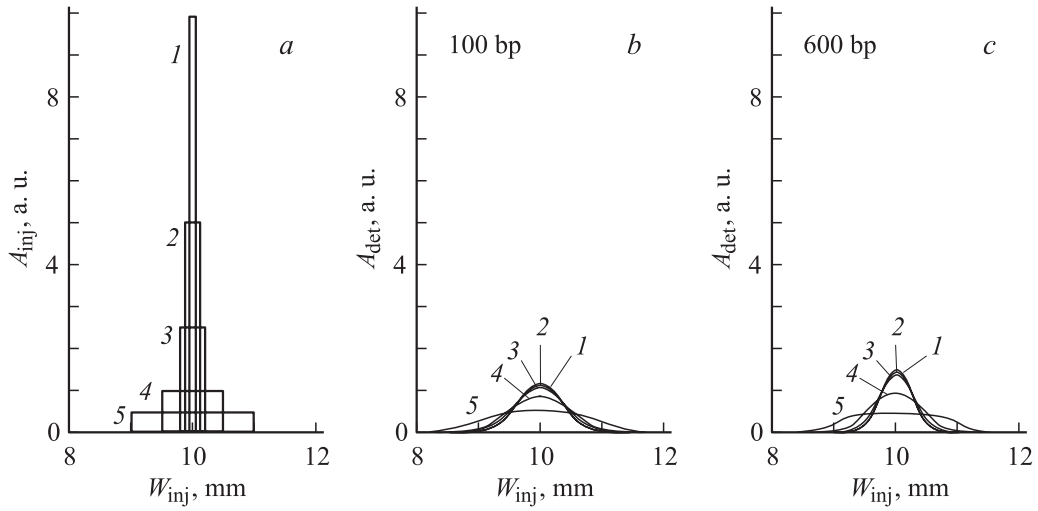

Рис. 10. Рассчитанные формы регистрируемых пиков A_{det} (b и c) для инжекционных зон $W_K = 0.1$ (*1*), 0.2 (*2*), 0.4 (*3*), 1 (*4*), 2 mm (*5*). Соответствующие инжекционные пики $A_{inj} = 1/W_K$ (a). $L_{det} = 50$ cm, POP-7, 50°C, $E = 120$ V/cm, $R_{th} = 0.75$, $SNR_{th} = 4$.

восстановить первоначальную форму зоны исходя из формы регистрируемого пика, воспользовавшись при этом наиболее подходящим методом. Рис. 10, *b, c* иллюстрирует формы регистрируемых пиков, рассчитанные при помощи (11) для инжекционных зон, имеющих одинаковую площадь $A_{inj} \times W_{inj}$ (рис. 10, *a*). Однако для узких инжекционных зон ($W_K < 0.4$ mm) регистрируемые пики имеют почти одинаковую форму (кривые *1*–*3*).

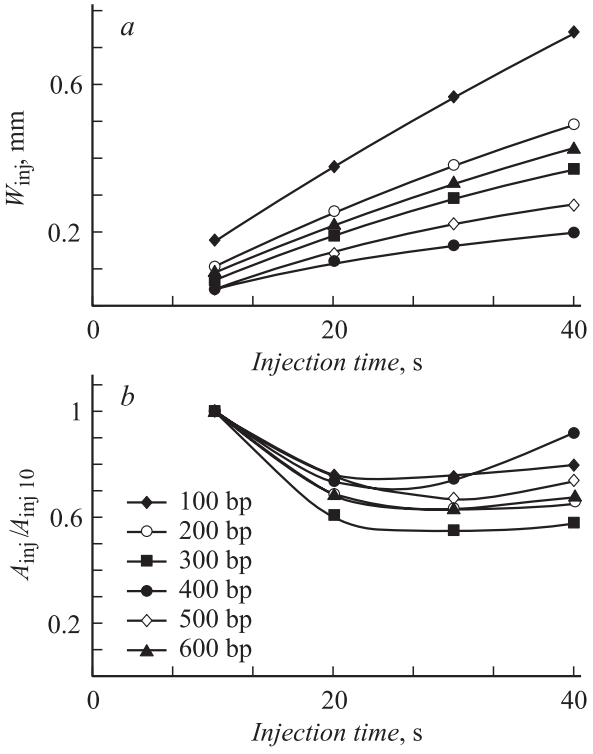

Рис. 11. Зависимость ширины (a) и относительной амплитуды инжекционной зоны от времени инжекции. ISL-600, $L_{det} = 50$ cm, POP-7, 50°C, $E = 120$ V/cm, $E_{inj} = 25$ V/cm.

Это означает, что для узких инжекционных зон форма регистрируемых пиков скорее зависит от площади зоны ($A_{inj} \times W_K$), чем от ее ширины.

Более точно W_K можно определить, проанализировав зависимость дисперсии пиков от времени инжекции (см. рис. 4). Погрешность оценки σ^2 можно снизить, вычислив линейную аппроксимацию $\sigma^2(T_{inj}^2)$ (штриховые линии на рис. 4) и принимая значения аппроксимирующей линии за „улучшенную оценку“ значения σ^2. Ширина инжекционной зоны может быть вычислена из уравнения (28), где σ^2 получена путем линейной аппроксимации. Высота $A_{inj\,K}$ инжекционной зоны, соответствующей детектируемой полосе высотой $A_{det\,K}$, может быть вычислена из (20). На рис. 11, *a, b* представлены зависимости ширины и амплитуды инжекционной зоны от времени инжекции, полученные для фрагментов 100–600 bp (ILS-600). Воспользовавшись этими зависимостями, можно весьма точно определить ширину и амплитуду инжекционных зон для небольшого времени инжекции.

Определив ширину инжекционной зоны W_K, можно вычислить отношение $n_{inj\,K}/n_{inj\,M}$ для режимов ДО (diffusion limited, DL) и ИО (injection limitec, IL) разделения

$$\left(\frac{n_{inj\,K}}{n_{inj\,M}}\right)_{DL} = \frac{A_{det\,K} W_M \sqrt{2\pi\sigma_K^2}}{A_{det\,M} W_K \sqrt{2\pi\sigma_M^2}} \quad (a),$$

$$\left(\frac{n_{inj\,K}}{n_{inj\,M}}\right)_{IL} = \frac{A_{det\,K}}{A_{det\,M}} \quad (b). \qquad (30)$$

Отношение $A_{det\,K}/A_{det\,100}$ ($100 \le K \le 600$), определенное экспериментально для типичных электроферограмм тестового образца ДНК, полученных на нашем секвенаторе, и отношение $n_{inj\,K}/n_{inj\,100}$, вычисленное из уравнений (28) и (29), представлены двумя нижними кривыми на рис. 12. Изображенное на графике отношение

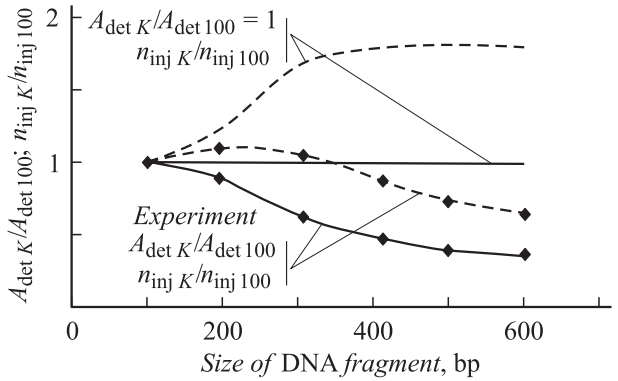

Рис. 12. Зависимость $A_{\text{det }K}/A_{\text{det }100}$ (сплошные линии) и $n_{\text{inj }K}/n_{\text{inj }100}$ (пунктир) от размера фрагмента ДНК (DNA); POP-7, 50°C, $L_{\text{det}} = 50$ cm, $T_{\text{inj}} = 20$ s, $E_{\text{inj}} = 25$ V/cm, $E = 120$ V/cm.

амплитуд пиков рассчитано путем усреднения амплитуд 8 пиков, соседних с K-ым пиком. Можно отметить, что инжектированная концентрация $n_{\text{inj }K}$ не является постоянной величиной и уменьшается по мере увеличения размера фрагментов ДНК. Это служит причиной примерно 50% общего уменьшения амплитуды пиков для больших фрагментов ДНК. Остальные 50% уменьшения амплитуды связаны с диффузионным расплыванием пика. Вышеприведенный анализ показывает, что уменьшение амплитуды регистрируемых пиков можно вообще избежать, повысив концентрацию инжектированных образцов ДНК. Прямая линия на графике соответствует идеальной электроферограмме с постоянной амплитудой пиков на протяжении всего секвенсового прогона. Чтобы получить такого рода данные, необходимо увеличить отношение $n_{\text{inj }K}/n_{\text{inj }100}$ примерно в 3 раза для больших фрагментов ДНК (ср. верхние и нижние пунктирные линии). Естественно, чтобы добиться такого увеличения, необходимо оптимизировать как процесс приготовления, так и процедуру инжекции образца.

2.4. Вычисление длины прочтения последовательности секвенирующей системой

Полученные параметры системы секвенирования позволяют оптимизировать длину прочтения для любого данного K. На рис. 13 представлен процесс вычисления для двух случаев:

— оптимальной длины капилляра,

— фиксированной длины капилляра.

Для любого данного размера K фрагмента ДНК, воспользовавшись соотношением (20), вычисляем отношение высоты детектируемого пика $A_{\text{det }K}$ и высоты инжектированного пика $A_{\text{inj }K}$, зависящее от W_{inj} (верхняя часть рис. 13). Затем из (22) вычислим зависимость минимальной длины капилляра $L_{\text{det }K}$, необходимой для получения порогового разрешения R_{th}, от ширины зоны инжекции W_K (средняя часть рис. 13). И наконец, вычисляем зависимость разрешения R, определенного из (6), от ширины зоны инжекции для фиксированной длины капилляра (нижняя часть рис. 13, $L_{\text{det}} = 50$ cm). Выражения для $A_{\text{det }K}$ и $L_{\text{det }K}$ при использовании DPM модели не могут быть записаны аналитически, однако эти зависимости могут быть получены путем численной симуляции из исходной формы пика (12).

Затемненная область в верхней части рис. 13 соответствует условию $n_{\text{inj}} > n_{\text{det th}}$ (или $A_{\text{inj}} > A_{\text{det th}}$). Если это условие удовлетворяет для любого данного размера K фрагмента ДНК и любого данного n_{inj} (отрезок, отсекаемый на оси Y пунктирной линией, опущенной с кривой для $K = 600$), то можно определить соответствующую минимальную ширину зоны инжекции $W_{K\ \text{min}}$ (отрезок на оси X, отсекаемый штриховой линией, в верхней части рисунка), необходимую для поддержания амплитуды регистрируемого пика A_{det} выше порога детектирования $A_{\text{det th}}$. Получив $W_{K\ \text{min}}$, мы сможем вычислить минимальную длину капилляра $L_{\text{det min}}$, которая позволит разделить соседние группы фрагментов с пороговым разрешением R_{th} (средняя часть рисунка, отрезок на оси Y, отсекаемый пунктирной линией). Для того чтобы

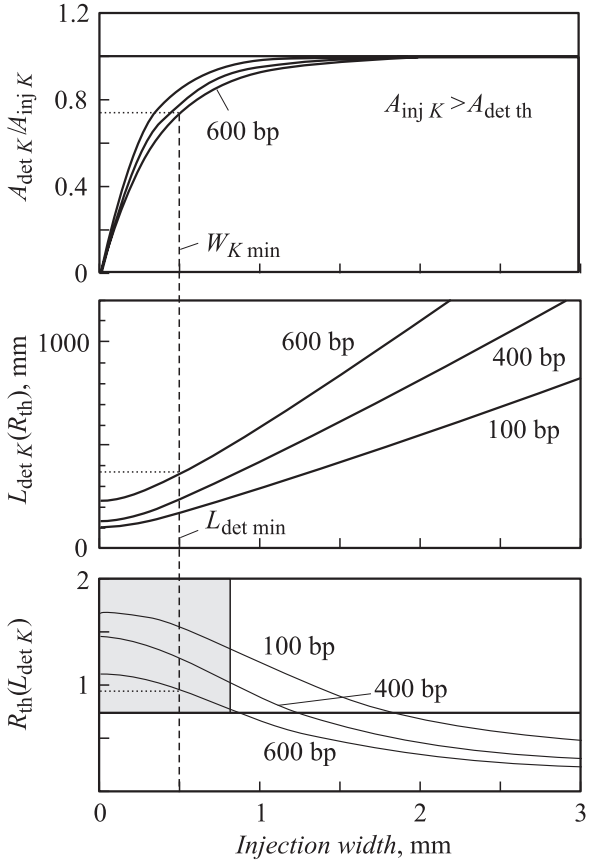

Рис. 13. Иллюстрация метода вычисления длины прочтения для фрагментов размером 600 bp: отношение высоты детектируемого и инжектированного пиков $A_{\text{det }K}/A_{\text{inj }K}$ (верхняя часть), минимальная длина капилляра $L_{\text{det }K}$ (средняя часть) и разрешение R (нижняя часть) в зависимости от ширины зоны инжекции.

ределить, может ли капилляр заданной длины разделить полосы, имеющие ширину $W_{K\ \min}$, следует проанализировать отрезок на оси Y, отсекаемый пунктирной линией, в нижней части рисунка. Если этот отрезок находится в пределах затемненной области графика, данная длина капилляра (по нашим расчетам, 50 cm) может обеспечить разрешение выше порогового R_{th}.

Заключение

Нами предложен метод экспериментального определения чувствительности и длины прочтения нуклеотидной последовательности для секвенаторов, основанных на капиллярном электрофорезе. Данный метод позволяет охарактеризовать разделяющую среду, определить ширину зоны инжекции и концентрацию фрагментов ДНК в ней, а также определить чувствительность системы секвенирования ДНК, используя полученные электрофореграммы. Основываясь на измеренных параметрах системы, предложенный метод позволяет вычислить длину прочтения с индексом качества Q20 и обеспечивает оптимизацию приготовления, инжекции и разделения образцов.

Список литературы

[1] *Tillib S.V., Mirzabekov A.D.* // Current Opinion in Biotechnology. 2001. Vol. 12. P. 53.

[2] *Shendure J., Mitra R.D., Varma C., Church G.M.* // Nat. Rev. Genet. 2004. Vol. 5. P. 335.

[3] *Sanger F., Nicklen S., Coulson A.R.* // Proc. Natl. Acad. Sci. USA. 1997. Vol. 74. P. 5463.

[4] *Lagally E.T., Emrich C.A., Mathies R.A.* // Lab Chip. 2001. Vol. 1. P. 102.

[5] *Koutny L., Schmalzing D., Salas-Solano O., El Difrawy S., Adourian A., Buonocore S., Abbey K., McEwan P., Matsudaira P., Ehrlich D.* // Anal. Chem. 2000. Vol. 72. P. 3388.

[6] *Khrapko K.R., Lysov Y., Khorlyn A.A., Shick V.V., Florentiev V.L., Mirzabekov A.D.* // FEBS Lett. 1989. Vol. 256. P. 118.

[7] *Lipshutz R.J., Morris D., Chee M., Hubbell E., Kozal M.J., Shah N., Shen N., Yang R., Fodor S.P.A.* // Biotechniques. 1995. Vol. 19. P. 442.

[8] *Drmanac S., Kita D., Labat I., Hauser B., Schmidt C., Burczak J.D., Drmanac R.* // Nature Biotechnology. 1998. Vol. 16. P. 54.

[9] *Patil N., Berno A.J., Hinds D.A., Barrett W.A., Doshi J.M., Hacker C.R., Kautzer C.R., Lee D.H., Marjoribanks C., McDonough D.P., Nguyen B.T., Norris M.C., Sheehan J.B., Shen N., Stern D., Stolowski R.P., Thomas D.J., Trulson M.O., Vyas K.R., Frazer K.A., Fodor S.P., Cox D.R.* // Science. 2001. Vol. 294. P. 1719.

[10] *Ronaghi M., Karamohamed S., Pettersson B., Uhlen M., Nyren P.* // Anal. Biochem. 1996. Vol. 242. P. 84.

[11] *Mitra R.D., Church G.M.* // Nucleic Acids Res. 1999. Vol. 27. P. 34.

[12] *Westin L., Xu X., Miller C., Wang L., Edman C.F., Nerenberg M.* // Nat. Biotechnol. 2000. Vol. 18. P. 199.

[13] *Ronaghi M.* // Genome Research. 2001. Vol. 11. P. 3.

[14] *Pourmand N., Elahi E., Davis R.W., Ronaghi M.* // Nucleic Acids Research. 2002. Vol. 30. P. 31.

[15] *Mitra R.D., Shendure J., Olejnik J., Edyta K.O., Church G.M.* // Anal. Biochem. 2003. Vol. 320. P. 55.

[16] *Dressman D., Yan H., Traverso G., Kinzler K.W., Vogelstein B.* // Proc. National Academy of Sci. 2003. Vol. 100. P. 8817.

[17] *Leamon J.H., Lee W.L., Tartaro K.R., Lanza J.R., Sarkis G.J., de Winter A.D., Berka J., Lohman K.L.* // Electrophoresis. 2003. Vol. 24. P. 3769.

[18] *Kartalov E.P., Quake S.R.* // Nucleic Acids Research. 2004. Vol. 32. P. 2873.

[19] *Korlach J., Levene M., Turner S.W., Larson D.R., Foquet M., Craighead H.G., Webb W.W.* // Biophys. J. 2001. Vol. 80. P. 147A.

[20] *Levene M.J., Korlach J., Turner S.W., Foquet M., Craighead H.G., Webb W.W.* // Science. 2003. Vol. 299. P. 682.

[21] *Braslavsky I., Hebert B., Kartalov E., Quake S.R.* // Proc. Natl. Acad. Sci. USA. 2003. Vol. 100. P. 3960.

[22] *Meller A., Nivon L., Brandin E., Golovchenko J., Branton D.* // Proc. Natl. Acad. Sci. USA. 2000. Vol. 97. P. 1079.

[23] *Deamer D.W., Akeson M.* // Trends in Biotechnology. 2000. Vol. 18. P. 147.

[24] *Winters-Hilt S., Vercoutere W., DeGurman V.S., Deamer D., Akeson M., Haussler D.* // Biophys. J. 2003. Vol. 84. P. 967.

[25] *Luckey J.A., Norris T.B., Smith L.M.* // J. Phys. Chem. 1993. Vol. 97. P. 3067.

[26] *Heller C.* // Electrophoresis. 2000. Vol. 21. P. 593.

[27] *Ewing B., Hillier L., Wendl M., Green P.* // Genome Research. 1998. Vol. 8. P. 175.

[28] *Ewing B., Green P.* // Genome Research. 1998. Vol. 8. P. 186.

[29] *Alaverdian L., Alaverdian S., Bilenko O., Bogdanov I., Filippova E., Gavrilov D., Gorbovitski B., Gouzman M., Gudkov G., Domratchev S., Kosobokova O., Lifshitz N., Luryi S., Rushovoloshin V., Stepoukhovitch A., Tcherevishnick M., Tyshko G., Gorfinkel V.* // Electrophoresis. 2002. Vol. 23. P. 2804.

[30] *Gavrilov D.N., Gorbovitski B., Gouzman M., Gudkov G., Stepoukhovitch A., Ruskovoloshin V., Tsuprik A., Tyshko G., Bilenko O., Kosobokova O., Luryi S., Gorfinkel V.* // Electrophoresis. 2003. Vol. 24. P. 1184.

[31] *Heller C.* // Electrophoresis. 1999. Vol. 20. P. 1962.

[32] *Cussler E.L.* // Diffusion: Mass Transfer in Fluid Systems. Cambridge: Cambridge University Press, 1984.