ISSN 0868–5886 *НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2014, том 24, № 2, с. 79–85*

СИСТЕМНЫЙ АНАЛИЗ ПРИБОРОВ И ИЗМЕРИТЕЛЬНЫХ МЕТОДИК

УДК 543.426; 543.

© Я. И. Алексеев, Д. А. Белов, Ю. В. Белов, В. Е. Курочкин

ИССЛЕДОВАНИЕ ПОГРЕШНОСТЕЙ ОЦИФРОВКИ ПИКОВ ГЕНЕТИЧЕСКОГО АНАЛИЗАТОРА

Рассмотрены основные причины различия электрофоретической мобильности флуоресцентно-меченых фрагментов ДНК. Выполнен анализ результатов экспериментального разделения фрагментов секвенсной смеси. Предложен способ определения базового временно́го интервала. Отличия реальных временны́х интервалов от базового временно́го интервала предложено рассматривать как систематическую составляющую погрешности измерений при определении последовательности пиков.

Кл. сл.: ДНК, генетический анализатор, флуоресцентная детекция

ВВЕДЕНИЕ

Определение нуклеотидной последовательности (секвенирование ДНК) является одной из основных задач, решаемых с помощью генетического анализатора, основанного на принципе капиллярного электрофореза [1]. Нуклеотидная последовательность традиционно записывается с использованием букв A, C, G и T.

При подготовке пробы получается смесь фрагментов ДНК с шагом 1 нуклеотид. Для раздельного детектирования фрагментов на конец каждого фрагмента помещается соответствующая флуоресцентная метка. Максимальная длина фрагментов определяется разрешающей способностью генетического анализатора (450 и более нуклеотидов).

Во время разделения фрагментов ДНК в капилляре под действием электрического поля на выходе четырех цветовых каналов флуоресцентного детектора регистрируются цифровые последовательности, которые графически изображаются в виде последовательностей пиков, соответствующих нуклеотидам A, C, G, T. При этом каждому цифровому отсчету соответствует точка графика, имеющая по горизонтальной оси (оси времени) номер отсчета, а по вертикальной оси — значение интенсивности флуоресценции в относительных единицах (о.е.).

Задача определения нуклеотидной последовательности решается путем измерения положения во времени пиков в каждом канале флуоресцентного детектора, присвоения им буквенных обозначений и суммирования результатов в виде буквенной последовательности.

В генетическом анализаторе флуоресцентный детектор имеет пятый цветовой канал, на выходе которого могут быть получены сигналы калибровочной смеси фрагментов ДНК известной длины [2]. Однако при секвенировании ДНК сигналы калибровочной смеси не используются.

ПРИЧИНЫ РАЗЛИЧИЯ ЭЛЕКТРОФОРЕТИЧЕСКОЙ МОБИЛЬНОСТИ СОСЕДНИХ ФЛУОРЕСЦЕНТНО-МЕЧЕНЫХ ФРАГМЕНТОВ ДНК

Разделение фрагментов ДНК в капилляре под действием электрического поля происходит за счет различия их электрофоретической мобильности. Смесь фрагментов ДНК регистрируется в виде последовательности пиков, при этом в первом приближении наблюдается почти линейная зависимость времени выхода от длины фрагмента с шагом 1 нуклеотид. Однако если электрофоретическая мобильность фрагментов, отличающихся по длине на 1 нуклеотид и соответствующих разным нуклеотидам, не соответствует общей калибровочной зависимости, то на графиках наблюдается неравномерное распределение соседних пиков. При частичном наложении соседних пиков, значительных случайных ошибках определения положения вершин пиков и ухудшении разрешающей способности в конце эксперимента (при максимальной длине фрагментов) уменьшается достоверность полученной информации: возможны ошибки при определении истинной нуклеотидной последовательности.

Первая причина различия электрофоретической мобильности — различие молекулярных масс соседних нуклеотидов и флуоресцентных меток (красителей).

Табл. 1. Соответствие красителей и нуклеотидов на конце фрагмента

Нуклеотид	Молекулярная масса нуклеотида	Второй краситель	Молекулярная масса второго красителя, г/моль	Суммарная молекулярная масса, без первого красителя	Длина волны излучения, нм
G	328	FAM	376	704	520
A	313	R6G	457	770	557
T	302	TAMRA	431	733	576
C	388	ROX	534	922	605

В генетическом анализаторе для получения флуоресцентных сигналов четырех красителей при использовании одного лазера с длиной волны излучения 488 нм используется эффект переноса энергии [3–5]. Для реализации этого эффекта на конец каждого фрагмента присоединяется соответствующая комбинация двух флуоресцентных красителей с помощью вставки линкерных молекул. В качестве первого красителя (донора) используется краситель FAM. В табл. 1 приведен возможный вариант сочетания нуклеотидов и вторых красителей (акцепторов) [6, 7].

На рис. 1 представлено взаимное расположение 2 соседних пиков во времени, относящихся к 2 ближайшим по массе фрагментам ДНК (с шагом 1 нуклеотид).

При допущениях, что мнимое положение первого "неокрашенного" пика принято равным нулю, а запаздывание "окрашенных" пиков пропорционально приращению молекулярной массы фрагментов, то расположение реальных "окрашенных" пиков можно описать следующими формулами:

$$t_1 = m \cdot M_{1\text{M}};$$
$$t_2 = m \cdot \left(M_2 + M_{2\text{M}}\right);$$
$$t_2 - t_1 = m \cdot \left(M_2 + M_{2\text{M}}\right) - m \cdot M_{1\text{M}} = m \cdot \left(M_2 + M_{2\text{M}} - M_{1\text{M}}\right),$$

где t_1 и t_2 — времена выхода 1-го и 2-го окрашенных пиков по отношению к неокрашенному 1-му пику; m — коэффициент электрофоретической мобильности; M_2 — приращение молекулярной массы фрагмента по отношению к первому; $M_{1\text{M}}$ и $M_{2\text{M}}$ — приращение молекулярной массы фрагментов за счет красителей.

Аналогичной формулой можно описать расположение всех последующих пиков с номерами (N–1) и N:

$$t_N - t_{(N-1)} = m \cdot \left(M_N + M_{N\text{M}} - M_{(N-1)\text{M}}\right).$$

Если принять в последней формуле допущение, что молекулярная масса нуклеотидов равна средней (базовой) $M_N = M_{\text{б}}$ и молекулярная масса красителей (с учетом первого красителя и линкера) равна средней (базовой) молекулярной массе красителей ($M_{N\text{M}} = M_{(N-1)\text{M}} = \text{M}_{\text{бК}}$), то временной интервал между соседними пиками будет равен

$$t_N - t_{(N-1)} = m \cdot M_{\text{б}} = T_{\text{б}}.$$

Временной интервал $T_{\text{б}}$ характеризует усредненную электрофоретическую мобильность, величину $T_{\text{б}}$ можно условно назвать как "базовый временной интервал". Последовательность пиков с интервалом $T_{\text{б}}$ носит равномерный характер, медленно изменяясь во времени от начала до конца эксперимента за счет изменения коэффициента m.

Отсюда следует:

$$m = T_{\text{б}} / M_{\text{б}},$$
$$\left(t_N - t_{(N-1)}\right) / T_{\text{б}} = T_N / T_{\text{б}} = \left(M_N + M_{N\text{M}} - M_{(N-1)\text{M}}\right) / M_{\text{б}}.$$

В последней формуле отношение временно́го интервала между соседними пиками T_N к базовому

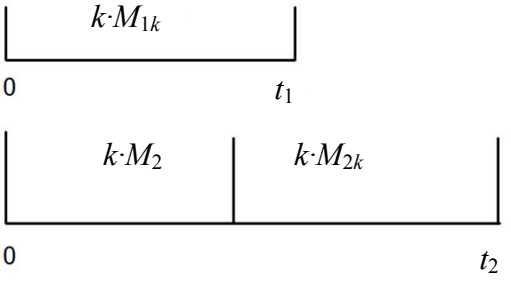

Рис. 1. Взаимное расположение во времени 1-го и 2-го соседних пиков (соответственно t_1 и t_2)

временно́му интервалу $T_б$ выражено в базовых относительных единицах (б.о.е.).

В качестве примера можно привести два частных случая, используя данные первой и четвертой строк табл. 1 (без учета 1 красителя и линкера):

$$(t_{3C} - t_{2G}) / T_б = (922 - 376) / 333 = 1.64 \text{ б.о.е.},$$
$$(t_{3G} - t_{2C}) / T_б = (704 - 534) / 333 = 0.51 \text{ б.о.е.}$$

Этот пример показывает, что при упрощенном допущении, когда запаздывание пиков пропорционально приращению молекулярной массы фрагментов, временны́е интервалы между соседними пиками значительно отличаются от базового временно́го интервала. Это приводит к неравномерному распределению пиков в последовательности, по этой причине в приведенном примере интервалы между пиками отличаются более чем в 3 раза.

Величину относительной неравномерности следования временны́х интервалов между соседними пиками можно вычислить следующим образом:

$$D_N = T_N / T_б - 1 = (M_N + M_{NM} - M_{(N-1)M}) / M_б - 1.$$

Для приведенных выше примеров величины D_N соответственно равны +0.64 и –0.49 б.о.е. Максимальная величина D_N наблюдается в случае, когда масса N-го нуклеотида значительно отличается от средней молекулярной массы, а массы красителей соседних пиков значительно отличаются между собой.

Большие разбросы временны́х интервалов между пиками относительно базового временно́го интервала в приведенных ниже экспериментальных данных могут быть объяснены второй причиной, которая приводит к изменениям электрофоретической мобильности, а именно отличиями пространственной конфигурации нуклеотидов и красителей.

ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ РАЗДЕЛЕНИЯ ФРАГМЕНТОВ СЕКВЕНСНОЙ СМЕСИ

В качестве примера на рис. 2, а, приведен небольшой участок исходных (сырых) данных, отражающих результаты разделения фрагментов секвенсной смеси плазмидной ДНК, образованной с помощью набора для секвенирования ABI PRISM BigDye TerMinator v3.1 Cycle Sequencing Mit (Life Technologies, США). Эти результаты получены в ЗАО "СИНТОЛ", Москва, при испытании опытного образца генетического анализатора НАНОФОР-05, производства Института аналитического приборостроения РАН, Санкт-Петербург.

На графиках рис. 2, а, некоторые соседние пики (нумерация пиков — на б): 1 и 2, 8 и 9, 10 и 11, 12 и 13, 18 и 19 от фрагментов ДНК, окрашенных разными красителями, — в значительной степени взаимно перекрываются.

Для обработки данных и определения последовательности ДНК использован программный модуль анализатора — ДНК АЛ. Полученная последовательность содержит более 450 нуклеотидов.

ИССЛЕДОВАНИЕ ЭЛЕКТРОФОРЕТИЧЕСКОЙ МОБИЛЬНОСТИ ФЛУОРЕСЦЕНТНО-МЕЧЕНЫХ ФРАГМЕНТОВ ДНК

Для оценки различия электрофоретической мобильности при использовании экспериментальных данных предлагается применить модифицированный способ аппроксимации табличной функции степенны́м полиномом в среде Excel (табл. 2). Этот способ отличается от использованного нами ранее при фрагментном анализе [4] тем, что в качестве критерия при аппроксимации табличной функции используется не временно́е положение пиков, а временны́е интервалы между ними. Результатом такой аппроксимации является монотонная функция, представляющая зависимость

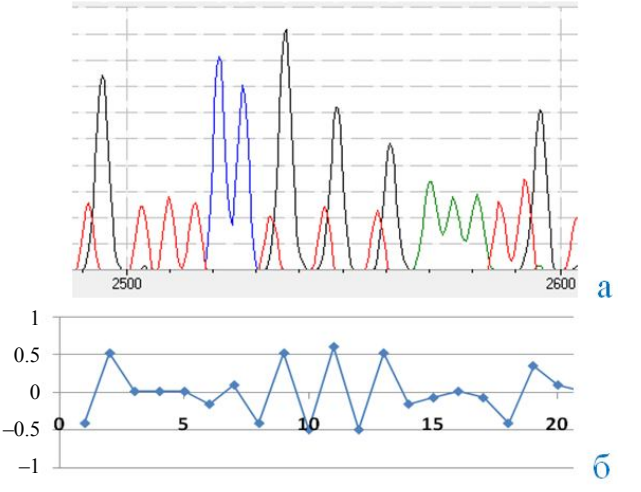

Рис. 2. Разделение фрагментов ДНК.
а — участок исходных данных, отражающих результаты разделения фрагментов секвенсной смеси; по горизонтали — текущее время в секундах, по вертикали — значение интенсивности флуоресценции в относительных единицах (о.е.). б — отличие временны́х интервалов между соседними пиками на рис. 1, а, от базового временно́го интервала (б.о.е.); горизонтальная ось — номер пика равномерной базовой последовательности (точки этого графика для наглядности соединены линиями)

Табл. 2. Вычисление отличия временны́х интервалов между соседними пиками от базового временно́го интервала

A	B	C	D	E	F	G	H	I	J	M
1906	1	G	1894.94	1900.80	5.85	9.5	3.65	13.30	806.23	0.62
1915.5	2	C	5.86081	1906.65	5.85	7	1.15	1.31	0.28	0.20
1922.5	3	T	0.00064	1912.50	5.86	4.5	–1.36	1.84		–0.23
1927	4	C	–2.3E–06	1918.36	5.86	2	–3.86	14.87		–0.66
1929	5	G		1924.21	5.86	6.5	0.64	0.41		0.11
1935.5	6	G		1930.07	5.86	11	5.14	26.43		0.88
1946.5	7	T		1935.93	5.86	5.5	–0.36	0.13		–0.06
1952	8	A		1941.79	5.86	3.5	–2.36	5.57		–0.40

базового временно́го интервала $T_б$ от времени и позволяющая выразить неравномерность интервалов между пиками в базовых относительных единицах.

При предварительной обработке сигналов флуоресцентного детектора программой ДНК АЛ были автоматически определены положения центров пиков (время выхода в секундах — столбец A в табл. 2) и присвоены номера пиков (столбец B). Каждому пику автоматически присвоено буквенное обозначение, соответствующее нуклеотиду и присоединенному красителю A, C, G или T (столбец C).

Аппроксимирующая функция выражена в виде полинома третьей степени в столбце E в следующем виде:

```
E1=$D$1+$D$2*B1+$D$3*(B1)^2+
   +$D$4*(B1)^3,
```

где величина $D$1 определяет начальное значение этой функции (сдвиг) и задается в первом приближении равной значению A1; величина $D$2 определяет единичное приращение этой функции (наклон) и задается в первом приближении равной значению (A2 – A1); величины $D$3 (кривизна) и $E$4 (изменение кривизны) задаются в первом приближении равными нулю.

В столбце F вычислены интервалы между соседними пиками аппроксимирующей функции в секундах (базовый временной интервал $T_б$):

```
F1=E2–E1.
```

В столбце G вычислены интервалы между реальными соседними пиками в секундах:

```
G1=A2–A1.
```

В столбце H вычислены разности интервалов реальных соседних пиков и соседних пиков аппроксимирующей функции (ошибки) в секундах:

```
H1=G2–F1.
```

В столбце I вычислены квадраты ошибок I1=H1^2, а столбце J — сумма квадратов ошибок J1=СУММ(I1:I400).

Величины $D$1, $D$2, $D$3, $D$4 уточняются с помощью метода наименьших квадратов (минимум величины J1) и метода последовательного приближения в меню Данные\Анализ\Поиск решения.

В столбце M вычислены отличия временны́х интервалов между соседними пиками от базового временно́го интервала M1=H1/F1, где F1 — базовый временной интервал.

Для демонстрации на рис. 3 результатов (столбец M) выбран участок от 51 до 350 нуклеотидов, на котором случайные составляющие погрешности определения положения пиков носят равномерный характер.

В ячейке J2 вычислено стандартное отклонение (величина, близкая к СКО):

J2=СТАНДОТКЛОН(M51:M350) = 0.28 б.о.е.

На рис. 2, б, в увеличенном масштабе изобра-

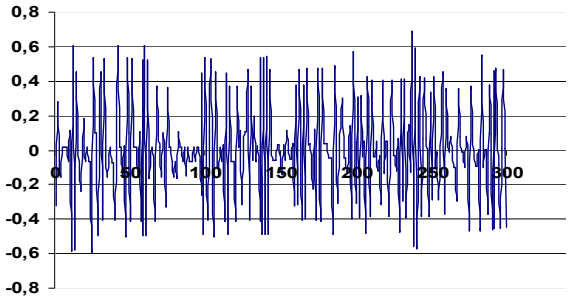

Рис. 3. Отличия временны́х интервалов между соседними пиками от базового временно́го интервала.
D_N (б.о.е.) — вертикальная ось. Горизонтальная ось — номер пика. Точки графика для наглядности соединены линиями

жена часть рис. 3. Точка 1 на рис. 2, б, соответствует точке 101 на рис. 3 и т. д.

При рассмотрении графика рис. 2, б, можно отметить, что максимальные отрицательные величины $D_N \approx -0.5$ б.о.е. относятся к нуклеотидам T (пики 1, 8, 10, 12, 18), а максимальные положительные величины $D_N \approx 0.5$ б.о.е. относятся к нуклеотидам с обозначением G (пики 2, 9, 11, 13, 19). Этот эффект наблюдается в случаях, когда нуклеотиды с обозначением T и G чередуются. На основании этого наблюдения можно сделать вывод о том, что фрагменты ДНК, которые оканчиваются нуклеотидами с обозначениями T и G, значительно различаются по электрофоретической мобильности (из-за отличия по массе и пространственной конфигурации нуклеотидов и красителей).

Отличия реальных временны́х интервалов от базового временно́го интервала носят систематический характер, их предлагается рассматривать как систематическую составляющую погрешности измерений при определении положения и последовательности пиков.

ОБСУЖДЕНИЕ ТРЕБОВАНИЙ К КОМБИНАЦИИ НУКЛЕОТИДОВ И КРАСИТЕЛЕЙ

При секвенировании ДНК четыре комбинации нуклеотидов на концах фрагментов и соответствующих красителей должны обеспечивать несколько требований:

1) минимальное расстояние в пространстве между красителями (донором и акцептором), поскольку степень переноса энергии обратно пропорциональна 6-й степени от этого расстояния [8];

2) близкие расстояния между этими красителями для получения одинаковых интенсивностей 4 флуоресцентных сигналов;

3) близкие характеристики электрофоретической мобильности.

Для оценки неравномерности положения соседних пиков во времени и уточнения границы допустимого различия характеристик электрофоретической мобильности рассмотрим еще раз положение во времени двух соседних пиков.

Для этого вернемся к начальной формуле

$$t_2 - t_1 = m_2 \cdot (M_2 + M_{2\text{M}}) - m_1 \cdot M_{1\text{M}},$$

где m_1 и m_2 — коэффициенты, характеризующие электрофоретическую мобильность окрашенных фрагментов ДНК, которые регистрируются в виде соседних пиков.

Сдвиг по времени между пиками будет близок к нулю или даже может стать отрицательным, если коэффициент электрофоретической мобильности первого фрагмента будет значительно больше второго. Предельно допустимая неравномерность следования пиков включает в себя еще случайную погрешность определения пиков d_C, которая увеличивается при уширении пиков в конце эксперимента (при максимальной длине фрагментов).

В случае выполнения равенства

$$m_2 \cdot (M_2 + M_{2\text{M}}) + d_{2\text{C}} = m_1 \cdot M_{1\text{M}} + d_{1\text{C}}$$

соседние пики совпадут, и их последовательность не будет определена.

Характеристики электрофоретической мобильности комбинаций конечных нуклеотидов и красителей, которые зависят от молекулярных масс нуклеотидов и красителей, а также их пространственной конфигурации могут быть частично выровнены путем выбора соответствующих линкеров. Анализируя экспериментальные результаты разделения фрагментов в приведенном примере, можно сделать вывод о том, что фрагменты ДНК, окрашенные разными красителями, значительно различаются по электрофоретической мобильности, т. е. полного выравнивания с помощью линкеров не наблюдается. Поэтому необходимо найти способ устранения этого недостатка при вторичной обработке результатов разделения фрагментов секвенсной смеси.

ЗАКЛЮЧЕНИЕ

1 Рассмотрены основные причины различия электрофоретической мобильности флуоресцентно-меченых фрагментов ДНК: различие молекулярных масс соседних нуклеотидов, флуоресцентных меток (красителей) и пространственной конфигурации нуклеотидов.

2 Выполнен анализ результатов экспериментального разделения фрагментов секвенсной смеси. Предложен способ определения базового

временно́го интервала $T_б$. Показано, что последовательность пиков с интервалом $T_б$ носит равномерный характер. Отличия реальных временны́х интервалов от базового временно́го интервала носят систематический характер, их предлагается рассматривать как систематическую составляющую погрешности измерений при определении положения и последовательности пиков. Способ компенсации таких погрешностей будет предложен в следующей статье.

3 На конкретном примере выполнена оценка характерных отличий временны́х интервалов между соседними пиками. Показано, что чередующиеся пики, которые в значительной степени взаимно перекрываются, относятся к фрагментам ДНК, отличающимся по электрофоретической мобильности.

4 Рассмотрены требования к комбинации нуклеотидов и красителей.

СПИСОК ЛИТЕРАТУРЫ

1. *Алексеев Я.И., Белов Ю.В., Малюченко О.П. и др.* Генетический анализатор для фрагментного анализа ДНК // Научное приборостроение. 2012. Т. 22, № 4. С. 17–22.
2. *Белов Ю.В., Петров А.И., Лавров В.В., Курочкин В.Е.* Построение калибровочной линии при фрагментном анализе ДНК // Научное приборостроение. 2013. Т. 23, № 3. С. 26–31.
3. Методы расшифровки нуклеотидной последовательности фрагментов ДНК. URL: (http://molbiol.ru/protocol/13_03.html).
4. *Tu O., Mnott T., Marsh M. et al.* The influence of fluorescent dye structure on the electrophoretic mobility of end-labeled DNA // Nucleic Acids Research. 1998. Vol. 26, nu. 11. P. 2797–2802.
5. *Jingyue J., Alexander N., Glazer I. and Mathies A.R.* Cassette labeling for facile construction of energy transfer fluorescent primers // Nucleic Acids Research. 1996. Vol. 24, nu. 6. P. 1144–1148.
6. Система обработки нуклеотидных последовательностей HEID. URL: (http://www.impb.ru/pdf/NL_1984_1r.pdf).
7. Флуоресцентные красители. URL: (http://www.syntol.ru/infoflu.htm).
8. Перенос энергии между двумя хромофорами. URL: (http://en.wikipedia.org/wiki/Förster_resonance_energy_transfer).

Институт аналитического приборостроения РАН, г. Санкт-Петербург (*Белов Д.А., Белов Ю.В., Курочкин В.Е.*)

ЗАО "СИНТОЛ", г. Москва (*Алексеев Я.И.*)

Контакты: *Белов Юрий Васильевич,* bel3838@mail.ru

Материал поступил в редакцию: 7.03.2014

RESEARCH OF THE GENETIC ANALYZER DIGITIZATION PEAKS ERRORS

Ya. I. Alekseev[1], D. A. Belov[2], Yu. V. Belov[2], V. E. Kurochkin[2]

[1]*JSC Syntol, Moscow, RF*
[2]*Institute for Analytical Instrumentation of RAS, Saint-Petersburg, RF*

Basic reasons of distinction of electrophoretic mobility of fluorescently marked DNA fragments were considered. The analysis of results of the experimental sequence mixture fragments division was made. The method of basic time interval determination was offered. It were suggested to consider the differences between real and base time intervals as a systematic component of an error of measurements in case of peaks sequence determination.

Keywords: DNA, genetic analyzer, fluorescent detection

REFERENCES

1. Tu O., Mnott T., Marsh M. et al. The influence of fluorescent dye structure on the electrophoretic mobility of end-labeled DNA. *Nucleic Acids Research*, 1998, vol. 26, nu. 11, pp. 2797–2802.

2. Jingyue J., Alexander N., Glazer I. and Mathies A.R. Cassette labeling for facile construction of energy transfer fluorescent primers. *Nucleic Acids Research*, 1996, vol. 24, nu. 6, pp. 1144–1148.

Contacts: *Belov Yurii Vasilyevich*,
bel3838@mail.ru

Article arrived in edition: 7.03.2014