

УДК 543.07

© В. В. Манойлов, А. Г. Бородин, А. Я. Логинов, А. И. Петров,
И. В. Заруцкий, В. Е. Курочкин, 2025

КОРРЕКЦИЯ ФЛУОРЕСЦЕНТНЫХ СИГНАЛОВ, ИСКАЖЕННЫХ НАЛОЖЕНИЕМ СПЕКТРАЛЬНЫХ ПОЛОС, В СЕКВЕНАТОРЕ НАНОФОР СПС

Важным этапом в технологии *массового параллельного секвенирования* нуклеиновых кислот является процесс измерения сигналов флуоресценции. Спектральный состав сигнала флуоресценции несет информацию о типе объекта флуоресценции (нуклеотида), входящего в состав образца ДНК, анализируемого секвенатором. Чем больше нуклеотидов в процессе чтения будет распознано верно, тем точнее будет результат прочтения всего генома. Явление наложения спектральных составляющих сигналов флуоресценции нуклеотидов А, С, G и Т (явление перекрестных помех, в англоязычной литературе: *cross-talk*) — один из негативных факторов, препятствующий достоверному прочтению генома. Для борьбы с этим явлением необходим алгоритм цветокоррекции. Этот алгоритм является решающим в определении типа нуклеотида, находящегося в данных координатах (кластере) в данном цикле чтения. В работе рассматривается алгоритм цветокоррекции, позволяющий с помощью метода линейной регрессии определить матрицу перекрестных помех. На основе этой матрицы определяется обратная матрица, использование которой позволяет существенно снизить влияние наложения спектральных полос на точность оценки измерений сигналов флуоресценции. Алгоритм апробирован на экспериментальных данных и показал высокую достоверность результатов.

Ключевые слова: секвенирование нуклеиновых кислот, математическая обработка данных, явление перекрестных помех, матрица перекрестных помех, линейная регрессия, метод наименьших квадратов

ВВЕДЕНИЕ

В Институте аналитического приборостроения РАН разработан первый отечественный аппаратно-программный комплекс (АПК) для расшифровки последовательности нуклеиновых кислот методом *массового параллельного секвенирования* Нанофор СПС [1, 2]. Метод *массового параллельного секвенирования* еще называют *технологией секвенирования методом синтеза*. В АПК Нанофор-СПС молекулы флуорофора (флуоресцентного красителя), которыми помечены нуклеотиды А, С, G и Т, возбуждаются под действием лазерного излучения. Сигналы флуоресценции каждого нуклеотида имеют длину волны, соответствующую типу нуклеотида. Сигналы флуоресценции регистрируются с помощью четырех видеокамер (по числу типов нуклеотидов) [3]. Регистрируемое излучение проходит через различные светофильтры, соответствующие длинам волн сигналов флуоресценции каждого из четырех красителей, которыми помечены нуклеотиды, — зеленый, оранжевый, красный и темно-красный. "Различные химические процессы, включенные в технологию секвенирования методом синтеза, вызывают смещения в значениях регистрируемых интенсивностей,

включая эффекты фазирования / префазирования (phasing / prephasing), затухания сигнала (signal decay) и перекрестные помехи (cross-talk)" [4]. В настоящей работе рассматриваются методы коррекции флуоресцентных сигналов, искаженных влиянием перекрестных помех (cross-talk).

Предполагается, что методы коррекции флуоресцентных сигналов, искаженных другими факторами, будут рассмотрены в другой работе авторов.

Частоты излучения (эмиссии) используемых флуоресцентных красителей частично перекрываются, что приводит к корреляции показаний интенсивностей. Это в свою очередь приводит к тому, что с ростом интенсивности сигнала, например в канале нуклеотида А, растет и интенсивность сигнала в канале нуклеотида С, и наоборот. Наибольшее взаимное влияние друг на друга попарно оказывают сигналы каналов А и С, G и Т соответственно (рис. 1). Также данное явление может быть названо как паразитная флуоресценция или перекрестные помехи: cross-talk.

На рис. 1 буквами G, T, A и C обозначены спектры высвечивания флуоресцентных красителей, используемых для окрашивания соответствующих нуклеотидов, а L1 и L2 — спектральные линии

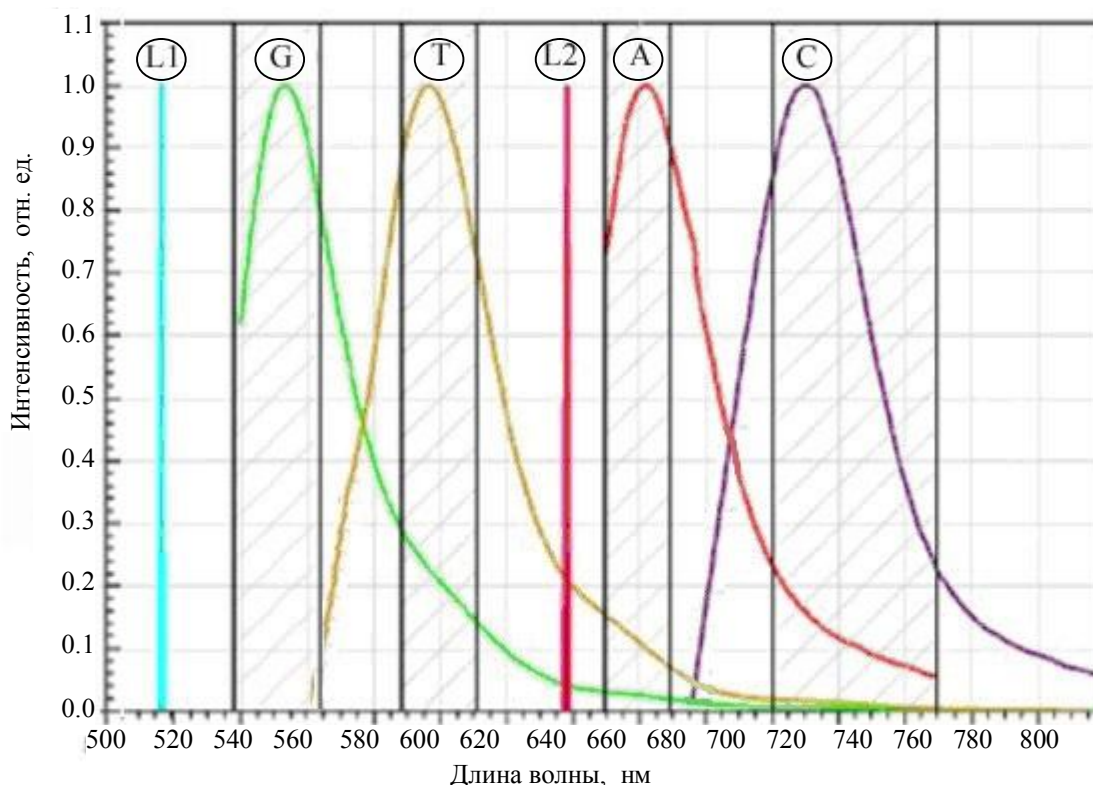


Рис. 1. Спектры эмиссии флуоресцентных меток и лазеров.

лазерных излучателей 515 и 638 нм оптической системы прибора Нанофор СПС. Штриховкой показаны полосы пропускания светофильтров в каналах регистрации. Длины волн лазерного возбуждения флуоресценции выбраны, исходя из обычных требований оптимальной квантовой эффективности каждой пары красителей. Можно ожидать эффективности возбуждения красителя G около 92% от эффективности на длине волны максимума экстинкции, красителя T — около 26%, красителя A — около 90%, красителя C — около 17%. Из данных, представленных на рис. 1, видно, что взаимное паразитное влияние присутствует для всех четырех каналов. Выбор полос пропускания светофильтров каналов регистрации основывается на соображениях охвата такого диапазона спектра полезного сигнала, интенсивность в котором превышает заданный порог и при этом паразитное влияние других каналов минимально. В данном случае эти два требования конфликтуют между собой, поэтому на практике реализовано компромиссное решение. В дальнейшем, после съема изображений с видеокамер прибора и ряда процедур цифровой обработки этих изображений, описанных в работах [3, 5] и других, явление наложения спектральных составляющих сигналов

флуоресценции корректируется в программном обеспечении. Программы, реализующие коррекцию данных, искаженных перекрестными помехами, являются обязательной составной частью программ первичного анализа данных, описанных в работах [6–9].

В настоящей работе описывается алгоритм, предназначенный для коррекции данных, искаженных в результате явления наложения спектральных составляющих оптических сигналов, регистрируемых прибором Нанофор СПС, и показывается, что отсутствие такой коррекции приводит к существенным ошибкам в проведении генетического анализа.

ПРЯМАЯ И ОБРАТНАЯ МАТРИЦЫ ПЕРЕКРЕСТНЫХ ПОМЕХ

Явление перекрестных помех заключается в том, что один и тот же кластер нуклеотида флуоресцирует одновременно во всех каналах. В отсутствие перекрестных помех данный кластер флуоресцировал бы только в одном канале, соответствующем красителю, которым мечен нуклеотид, находящийся в данный момент в данном кластере. Говоря иначе, сумма измеренных в каждом канале

значений интенсивности флуоресценции в данном кластере строго соответствовала бы значению интенсивности только в одном канале, являющемся в данный момент каналом полезного сигнала. То есть интенсивности сигнала флуоресценции в остальных каналах равны нулю. Математически интенсивность сигнала флуоресценции в случае отсутствия перекрестных помех можно представить соотношением (1)

$$I_i(x, y) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} I_A \\ I_C \\ I_G \\ I_T \end{bmatrix} + \begin{bmatrix} b_A \\ b_C \\ b_G \\ b_T \end{bmatrix}. \quad (1)$$

$$I_i(x, y) = \begin{bmatrix} 1 & w_{12} & w_{13} & w_{14} \\ w_{21} & 1 & w_{23} & w_{24} \\ w_{31} & w_{32} & 1 & w_{34} \\ w_{41} & w_{42} & w_{43} & 1 \end{bmatrix} \times \begin{bmatrix} I_A \\ I_C \\ I_G \\ I_T \end{bmatrix} + \begin{bmatrix} b_A \\ b_C \\ b_G \\ b_T \end{bmatrix} = \mathbf{M} \times \begin{bmatrix} I_A \\ I_C \\ I_G \\ I_T \end{bmatrix} + \begin{bmatrix} b_A \\ b_C \\ b_G \\ b_T \end{bmatrix}. \quad (2)$$

В соотношении (2): $\mathbf{M}$ — матрица перекрестных помех; w_{ij} — коэффициенты матрицы перекрестных помех, называемые также угловыми коэффициентами.

Задачей алгоритма цветокоррекции является сведение к минимуму влияния перекрестных помех. Математическое решение данной задачи можно представить в виде (3), как умножение правой части (2) на матрицу $\mathbf{M}^{-1}$, обратную матрице $\mathbf{M}$.

$$I_{i, \text{corr}}(x, y) = \mathbf{M}^{-1} \times \begin{bmatrix} I_{iA} \\ I_{iC} \\ I_{iG} \\ I_{iT} \end{bmatrix}. \quad (3)$$

В соотношении (3) $I_{i, \text{corr}}(x, y)$ — скорректированное суммарное значение интенсивности сигнала кластера с координатами x и y . Для решения данной задачи первоначально необходимо определить значения всех коэффициентов матрицы перекрестных помех.

ПОПАРНЫЕ ГРАФИКИ РАССЕЙВАНИЯ ИНТЕНСИВНОСТЕЙ

По данным измеренных значений сигналов флуоресценции во всех четырех каналах, форми-

В соотношении (1):

– $I_i(x, y)$ — сумма измеренных значений интенсивности во всех каналах в пикселе с координатами x и y ;

– I_A, I_C, I_G, I_T — измеренные чистые значения интенсивности в каждом канале;

– b_A, b_C, b_G, b_T — значения фона в ПЗС-матрице.

В случае наличия перекрестных помех суммарная померенная интенсивность уже не будет соответствовать интенсивности сигнала только в одном канале, при этом соотношение (1) примет вид (2)

руются так называемые попарные графики рассеивания интенсивности. На рис. 2 и рис. 3 представлены соответственно пары каналов С–А и каналов Т–Г. Графики пар каналов Г–А, Т–А, и других не представлены, т.к. искажения интенсивностей сигналов флуоресценции из-за наложения спектральных составляющих в этих каналах практически отсутствуют.

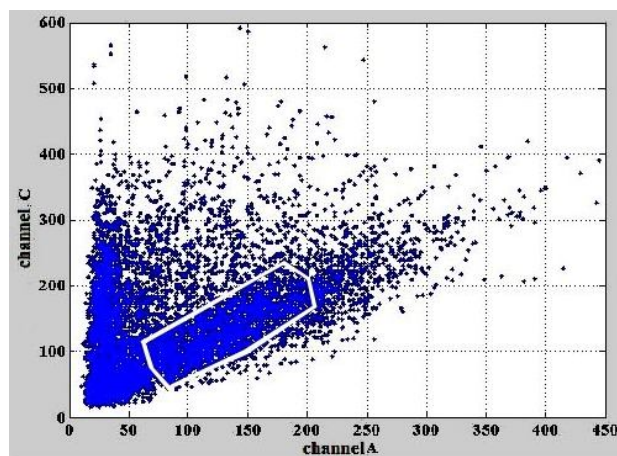


Рис. 2. Зависимость интенсивностей сигналов флуоресценции в канале С от интенсивностей сигналов флуоресценции в канале А.

Белым контуром выделена область данных, которая используется для определения параметров линии регрессии

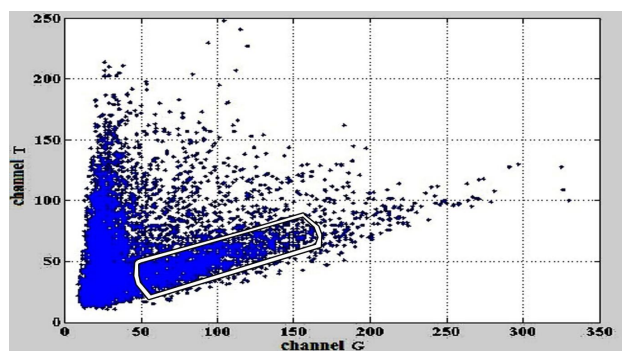


Рис. 3. Зависимость интенсивностей сигналов флуоресценции в канале Т от сигналов флуоресценции в канале G.

Белым контуром выделена область данных, которая используется для определения параметров линии регрессии

Каждый из двух графиков формируется следующим образом: по оси абсцисс откладывается значение интенсивности сигнала флуоресценции кластера с координатами x_i и y_i в одном канале, а по оси ординат откладывается значение интенсивности сигнала флуоресценции в этом же пикселе, но уже в другом канале;

Очевидно, что наибольшее значение интенсивности в кластере с координатами x_i и y_i будет наблюдаться в канале полезного сигнала. Таким образом, из всего набора данных, отображаемых на попарном графике рассеивания, отчетливо формируются две группы, для каждой из которых полезный сигнал относится только к одному из двух каналов.

АЛГОРИТМ ОЦЕНКИ КОЭФФИЦИЕНТОВ МАТРИЦЫ ПЕРЕКРЕСТНЫХ ПОМЕХ И ПОСЛЕДУЮЩЕЙ ЦВЕТКОРРЕКЦИИ

- Выделение области данных интенсивностей обнаруженных кластеров, которые искажены явлением перекрестных помех, шумом, выбросами и т.п. Данная область используется для определения параметров линии регрессии [6].
- Оценка матрицы перекрестных помех путем линейной регрессии данных попарных графиков.
- Вычисление обратной матрицы к матрице перекрестных помех.
- Применение найденной обратной матрицы к искаженным перекрестными помехами данным попарных графиков рассеивания.

ОСНОВНЫЕ ФУНКЦИИ ПРОГРАММЫ, РЕАЛИЗУЮЩЕЙ АЛГОРИТМ ЦВЕТКОРРЕКЦИИ

Общая характеристика

Перед реализацией алгоритма цветокоррекции в среде разработки программ необходимо сделать несколько уточнений, касающихся применения этого алгоритма к АПК Нанофор СПС. Входными данными алгоритма цветокоррекции является массив интенсивностей обнаруженных сигналов флуоресценции, полученных в кластерах с координатами, соответствующими координатам шаблона [3]. Для построения шаблона используется алгоритм, основанный на свертке со второй производной двумерного гауссового пика. В результате применения алгоритма свертки из обнаруженных сигналов флуоресценции практически удаляется фон [5]. В ходе описания алгоритма коррекции прежде всего будет рассмотрено взаимовлияние каналов А и С. Затем будет рассмотрено взаимовлияние каналов G и Т. В связи с этим окончательные входные данные представлены в виде двухстрочного массива экспериментально полученных значений интенсивностей в каналах А и С. При этом общую задачу цветокоррекции можно привести из вида (3) в вид (4):

$$I_{i, \text{corr AC}}(x, y) = \mathbf{M}_{\text{AC}}^{-1} \times \begin{bmatrix} I_{iA} \\ I_{iC} \end{bmatrix}, \quad (4)$$

$$\mathbf{M}_{\text{AC}} = \begin{bmatrix} 1 & w_{12} \\ w_{21} & 1 \end{bmatrix}. \quad (5)$$

Очевидным становится факт, что выходными данными алгоритма является двухстрочный массив откорректированных входных данных. Общая концепция алгоритма заключается в следующем.

- Создание подпрограммы, выделяющей области данных интенсивностей обнаруженных кластеров, которые искажены явлением перекрестных помех и устраняющей влияние шумов и базовой линии — **"BuildPrc"**.
- Создание подпрограммы для оценки угла наклона полинома линейной регрессии входных данных — **"SlopeEstimateN"**.
- Реализация полного алгоритма (совместной работы **"SlopeEstimateN"** и **"BuildPrc"**).
- Нахождение корректирующей матрицы **M** и ее преобразование в обратную матрицу **M**⁻¹.
- Итоговое тестирование алгоритма с использованием референтного генома бактериофага Phix 174.

Подпрограмма "BuildPrc"

Основное назначение данной подпрограммы — формирование вектора данных, которые могут быть использованы для вычисления углового коэффициента наклона линии регрессии.

Этот вектор формируется из вектора входных данных, искаженных перекрестными помехами, шумом и другими мешающими полезной информации факторами. Согласно данным работы [6], кластеры (пиксели), значение интенсивности которых ниже 60-го и выше 90-го перцентилей, не могут быть использованы для корректной оценки углового коэффициента. Поэтому, чтобы оценить чистое влияние перекрестных помех без постороннего шума, кластеры с подходящими под критерий шума и других мешающих факторов значениями интенсивности должны быть удалены из вектора входных данных. Это и реализует подпрограмма "BuildPrc". После удаления "плохих" кластеров участок значений интенсивности в канале полезного сигнала разбивается на несколько равных интервалов (бинов). В каждом интервале (бине) берется среднее значение интенсивности по оси канала полезного сигнала и минимальное по оси канала паразитного сигнала [6]. Это и есть значения интенсивностей в векторе, который далее подвергается обработке подпрограммой "SlopeEstimateN".

На рис. 2 и рис. 3 белым контуром выделены данные, которые используются для определения параметров линии регрессии.

На рис. 4 представлена линия регрессии (пунктирная линия), построенная после процедуры оценки перцентилей, разбиения на интервалы (би-

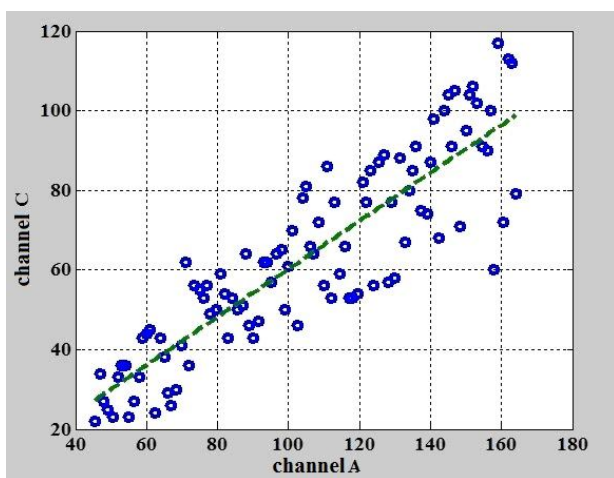


Рис. 4. Линия регрессии (пунктирная линия), построенная после выделения данных с помощью подпрограммы "BuildPrc" для данных, представленных на рис. 2.

ны) и нахождения минимумов в этих интервалах с помощью подпрограммы "BuildPrc" для данных, представленных на рис. 2. Линия регрессии была построена по 100 точкам, количество которых соответствовало количеству интервалов (бинов).

Угловой коэффициент — тангенс угла наклона линии регрессии, представленной на рис. 4, равен 0.6032, и, соответственно, угол наклона α равен 0.5428 рад. Далее в подпрограмме "SlopeEstimateN" с помощью матрицы поворота **MrRt** (6) производится поворот вектора данных, полученных в подпрограмме "BuildPrc", на угол наклона α по часовой стрелке:

$$\mathbf{MrRt} = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix}. \quad (6)$$

Результат поворота представлен на рис. 5.

Подпрограмма "SlopeEstimateN"

Данная подпрограмма является по сути ядром алгоритма. Ее основное назначение — поиск углового коэффициента вектора входных данных, к которому применяется линейная регрессия (7):

$$\sum_{i=1}^N (y_i - ax_i - b)^2 \rightarrow \min_{a,b}. \quad (7)$$

Поиск углового коэффициента представляет собой итерационную процедуру приближения угла наклона линии регрессии к нулевому значению. В каждой итерации вычисляется матрица поворота на угол α — **MrRt** (6), которая обновляется в каждой итерации; α — угол между линией регрессии и горизонтальной осью.

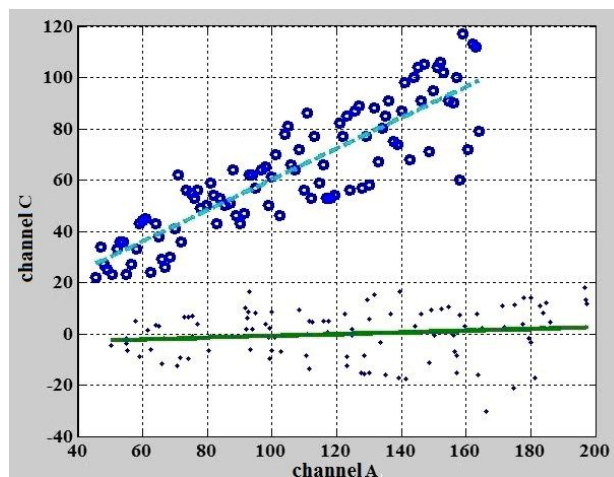


Рис. 5. Данные (o) и линия регрессии (пунктирная линия), показанные на рис. 4, и данные (точки), линия регрессии (сплошная линия), полученные после поворота на угол α .

Поворот на угол α является первым шагом итерационного процесса, который выполняет подпрограмма "SlopeEstimateN". Следующим шагом этого итерационного процесса является оценка угла наклона α_1 полученной линии регрессии. Например, для линии регрессии, показанной внизу рис. 5 (сплошная линия) угловой коэффициент — тангенс угла наклона равен 0.0342 и, соответственно, угол наклона α_1 равен 0.0342 рад. Затем углы α и α_1 складываются и осуществляется поворот на суммарный угол, значение которого присваивается переменной α . Далее подпрограмма "SlopeEstimateN" продолжает итерационную процедуру, осуществляя повороты, вычисления тангенсов углов наклона, значений самих углов и суммы значения угла, полученного на очередном шаге итерации, со значением угла, полученного на предыдущем шаге с учетом знака. Итерационная процедура заканчивается после того, как абсолютное значение тангенса угла наклона не станет меньше заданного порогового значения (th). Другим параметром окончания итерационного процесса является количество выполненных итераций N . Для рассматриваемого примера $th = 5 \cdot 10^{-5}$, а $N = 8$.

Полученный после итерационной процедуры тангенс угла наклона является элементом w_{21} искомой матрицы M_{AC} .

Для нахождения элемента w_{12} искомой матрицы M_{AC} (5), т.е. для оценки перекрестных помех из канала С в канал А должна строиться еще одна регрессия, для которой также используются рассмотренные подпрограммы "BuildPrc" и "SlopeEstimateN". В качестве входных данных используется зависимость интенсивностей в канале А от интенсивностей в канале С, а не зависимость интенсивностей в канале С от интенсивностей в канале А, как представлено на рис. 2. То есть то, что было функцией, стало аргументом, а то, что было аргументом, стало функцией. Зависимость интенсивностей в канале А от интенсивностей в канале С представлена на рис. 6. Белым контуром показана область данных, которая используется для оценки параметров линии регрессии. С помощью подпрограмм "BuildPrc" и "SlopeEstimateN" был найден угловой коэффициент линии регрессии для данных, выделенных белым цветом на рис. 6, значение которого является значением коэффициента w_{12} . Для рассматриваемого примера $w_{12} = 0.0421$.

Таким образом, теперь по известным элементам матрицы M_{AC} можно найти обратную матрицу M_{AC}^{-1} и рассчитать откорректированные значения интенсивностей в каналах А и С по формуле (4).

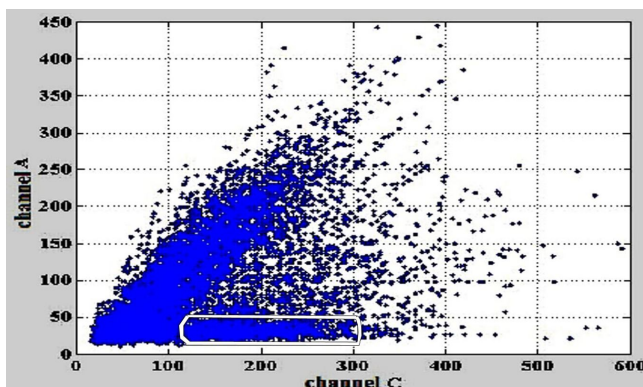


Рис. 6. Зависимость интенсивностей сигналов флуоресценции в канале А от интенсивностей сигналов флуоресценции в канале С.

Белым контуром выделена область данных, которая используется для определения параметров линии регрессии

ОТОБРАЖЕНИЕ РЕЗУЛЬТАТОВ ЦВЕТКОРРЕКЦИИ ДЛЯ КАНАЛОВ А–С И G–Т

Откорректированные данные для каналов А и С представлены на рис. 7.

Мы рассмотрели коррекцию влияния перекрестных помех на интенсивности сигналов флуоресценции в каналах А и С.

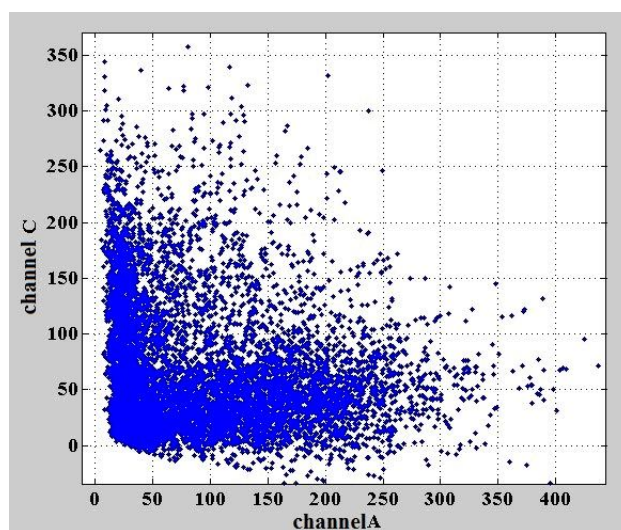


Рис. 7. Зависимость интенсивностей сигналов флуоресценции в канале С от интенсивностей сигналов флуоресценции в канале А после коррекции влияния перекрестных помех.

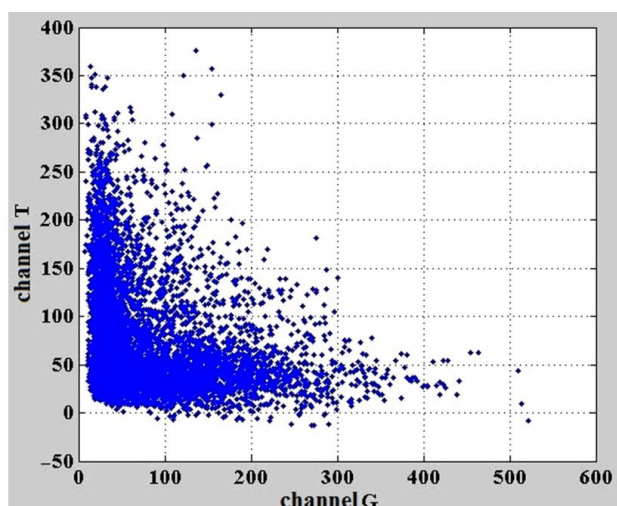


Рис. 8. Зависимость интенсивностей сигналов флуоресценции в канале Т от интенсивностей сигналов флуоресценции в канале G после коррекции влияния перекрестных помех.

Коррекция влияния перекрестных помех на интенсивности сигналов флуоресценции в каналах G и Т производится аналогично с помощью рассмотренных выше подпрограмм **"BuildPrc"** и **"SlopeEstimateN"**. Откорректированные данные для каналов G и Т представлены на рис. 8.

Мы рассмотрели процедуру коррекции влияния перекрестных помех на значения интенсивностей сигналов флуоресценции в каналах А–С и G–Т для результатов измерений в одном цикле. Аналогично производятся коррекции влияния перекрестных помех последовательно во всех циклах эксперимента.

По откорректированным значениям интенсивностей сигналов флуоресценции с помощью процедуры **"BaseCalling"** и программного обеспечения сборки генома строится последовательность букв нуклеотидов для исследуемой пробы.

Сравним результаты сборки генома для исходных и откорректированных на влияние перекрестных помех измеренных интенсивностей сигналов флуоресценции.

СРАВНЕНИЕ РЕЗУЛЬТАТОВ СБОРКИ ГЕНОМА ДЛЯ ОТКОРРЕКТИРОВАННЫХ И НЕОТКОРРЕКТИРОВАННЫХ ДАННЫХ

Сборка генома и оценка ее характеристик для исходных и откорректированных на влияние перекрестных помех измеренных интенсивностей сигналов флуоресценции проводилась с помощью программ **"Ugene"** и **"Qualimap"**. С помощью этих программ был определен процентный состав содержания нуклеотидов А, С, G и Т в пробе, содержащей геном бактериофага Phix174. Этот геном содержит 5386 нуклеотидов с известным процентным содержанием каждого из нуклеотидов. Этот геном является стандартом для проверки получаемой в секвенаторах генетической последовательности. Содержание отдельных нуклеотидов в этом геноме следующее: А — 24.0%, С — 21.5%, G — 23.3%, Т — 31.2%. [10]. В результате эксперимента оценки процентного содержания нуклеотидов по интенсивности сигналов флуоресценции, для которых была выполнена корректировка влияния перекрестных помех по рассмотренному выше алгоритму, процентное содержание нуклеотидов получилось следующее: А — 23.5%, С — 21.5%, G — 23.6%, Т — 31.4%, что соответствует стандартному геному Phix 174.

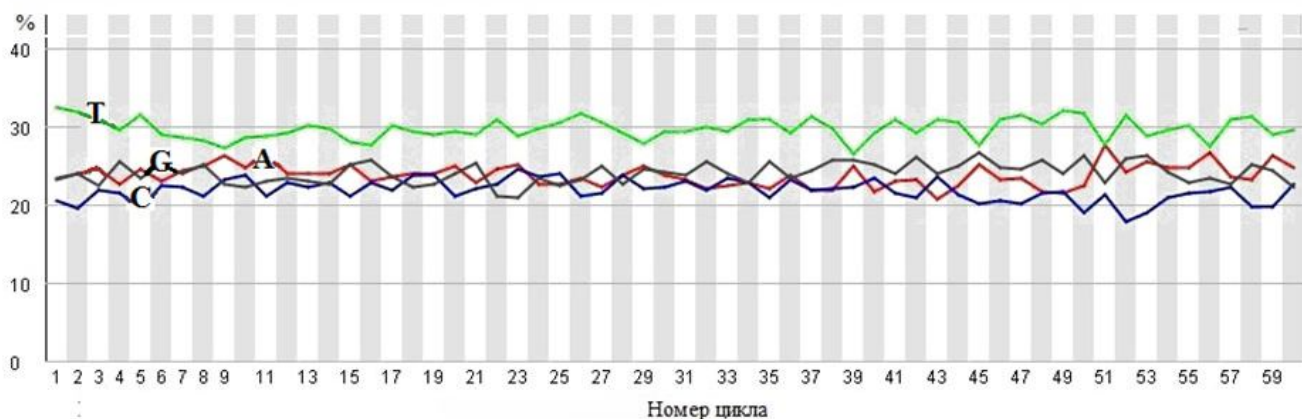


Рис. 9. Процентное содержание нуклеотидов А, С, G, и Т в зависимости от номера цикла измерений в секвенаторе при применении алгоритма коррекции наложения спектральных полос.

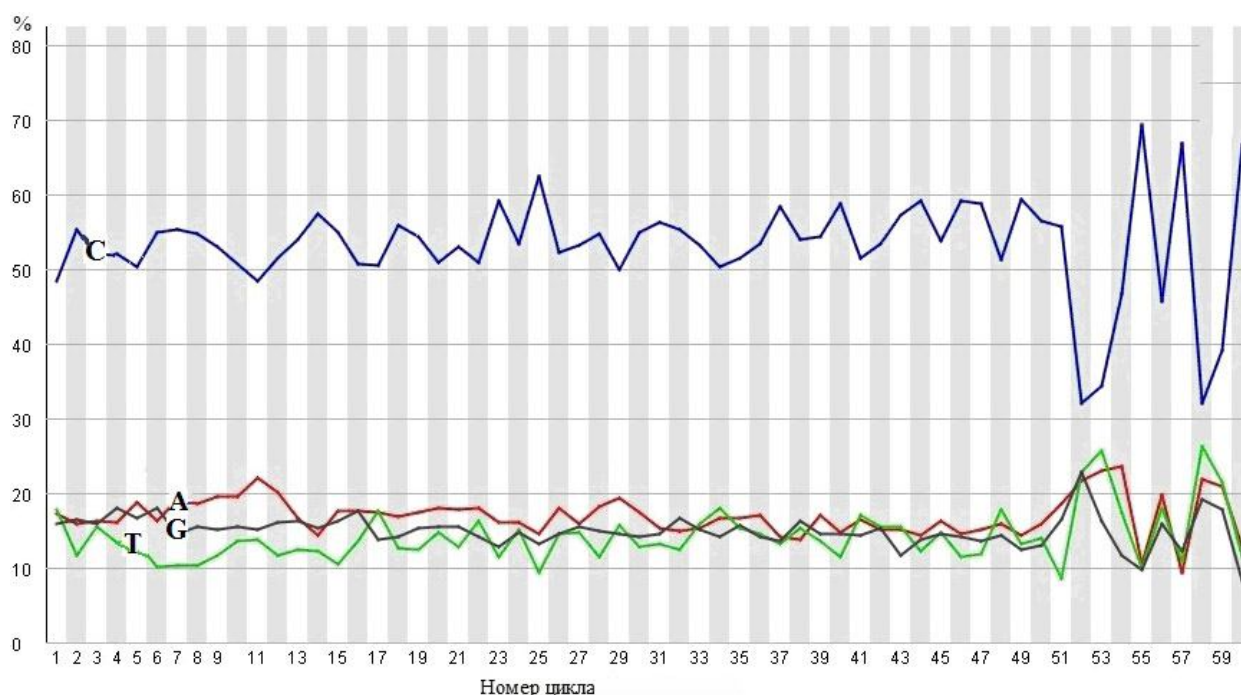


Рис. 10. Процентное содержание нуклеотидов А, С, G, и Т в зависимости от номера цикла измерений в секвенаторе без применения алгоритма коррекции наложения спектральных полос

На рис. 9 представлен результат процентного содержания отдельных нуклеотидов для исходных данных интенсивностей сигналов флуоресценции, которые подвергались корректировке на влияние перекрестных помех. На рис. 9 показано процентное содержание нуклеотидов А, С, G и Т в зависимости от номера цикла измерений в секвенаторе.

Как видно из данных, представленных на рисунке, процентное суммарное содержание нуклеотидов G и С практически во всех циклах приблизительно составляет 44%, что соответствует содержанию этих нуклеотидов в стандарте Phix174.

На рис. 10 представлен результат определения процентного содержания отдельных нуклеотидов для данных интенсивностей сигналов флуоресценции, для которых не была проведена корректировка на влияние перекрестных помех по рассмотренному выше алгоритму.

Как видно из данных, представленных на рисунке, процентное суммарное содержание нуклеотидов G и С практически во всех циклах превышает 65%. Этот факт свидетельствует о том, что интенсивности сигналов флуоресценции в канале С были искажены интенсивностями сигналов флуоресценции в канале А.

Сравнение результатов, представленных на рис. 9 и 10, показывает эффективность применения рассмотренного алгоритма.

Следует отметить, что использование коррекции перекрестных помех облегчает нам формирование обучающей выборки при применении машинного обучения в программном обеспечении секвенатора Нанофор СПС [4, 11]. В этом случае параметры моделей становятся менее коррелированными. С другой стороны, мы можем возложить саму коррекцию на методы машинного обучения.

ВЫВОДЫ

1. Явление перекрестных помех при измерениях сигналов флуоресценции приводит к ошибочным оценкам величин интенсивностей сигналов в соседних каналах секвенатора, искажающих последующую идентификацию буквенного кода нуклеотида.

2. Для коррекции ошибочных величин интенсивностей флуоресцентных сигналов необходимо определить матрицу перекрестных помех. На основе этой матрицы определяется обратная матрица, использование которой позволяет существенно снизить влияние наложения спектральных полос на точность оценки измерений сигналов флуоресценции.

3. Алгоритм нахождения элементов матрицы перекрестных помех включает следующие операции:

- а) с помощью оценки 60 и 90 перцентилей определяется диапазон величин интенсивностей

сигналов флуоресценции в канале, которые искажают величины интенсивностей в соседнем канале из-за явления наложения спектральных полос;

б) полученный диапазон величин интенсивностей сигналов разбивается на ряд равных интервалов;

в) в каждом интервале находится минимальная интенсивность;

г) полученные значения минимумов в каждом интервале используются для построения линии регрессии в итерационной процедуре;

д) находится угловой коэффициент, который является элементом матрицы перекрестных помех.

4. Явление перекрестных помех в секвенаторе Нанофор СПС искажает величины интенсивности только в двух парах каналов: А–С и G–Т. В других парах каналах, таких как А–G, А–Т, С–G, С–Т и т.п., искажения величин интенсивностей из-за явления перекрестных помех практически отсутствует. Это обстоятельство позволяет для коррекции влияния явления перекрестных помех упростить матрицу перекрестных помех и использовать две матрицы размером 2×2 . Одну матрицу для коррекции искажений в каналах А–С, а другую для каналов G–Т.

5. Результат генетического анализа стандарта Phix174 при использовании интенсивностей сигналов флуоресценции, не скорректированных на влияние перекрестных помех, показал, что процентное содержание нуклеотидов С и G существенно искажено, тогда как для скорректированных интенсивностей процентное содержание нуклеотидов С и G соответствует стандарту.

6. Методы коррекции флуоресцентных сигналов, искаженных эффектами фазирования / префазирования (phasing/prephasing), в настоящей работе не были рассмотрены, хотя эти методы также основаны на оценке параметров данных с помощью линейной регрессии. Предполагается рассмотрение этих методов в следующей работе.

Работа выполнена в ИАП РАН в рамках Государственного задания 075-01157-23-00 Министерства науки и высшего образования.

СПИСОК ИСТОЧНИКОВ

1. Курочкин В.Е., Алексеев Я.И., Петров Д.Г., Евстафьев А.А. Отечественные приборы для молекулярно-генетического анализа: разработки ИАП РАН и ООО "Синтол" // Известия Российской Военно-медицинской академии. 2021. Т. 40, № 3. С. 69–74. DOI: 10.17816/rmmar76918
2. Полногеномный секвенатор Нанофор СПС [Электронный ресурс] URL: <https://lab.scienceid.net/catalog/9824/28806/> (дата обращения 10.12.2023).
3. Манойлов В.В., Бородинов А.Г., Заруцкий И.В., Петров А.И., Курочкин В.Е. Алгоритмы обработки сигналов флуоресценции массового параллельного секвенирования нуклеиновых кислот // Труды СПИИРАН. 2019. Т. 18, № 4. С. 1010–1036. DOI: 10.15622/sp.2019.18.4
4. Бородинов А.Г., Манойлов В.В., Сараев А.С., Петров А.И., Заруцкий И.В., Курочкин В.Е., Сараев А.С. Машинное обучение в задачах base-calling для методов секвенирования нового поколения // Информатика и автоматизация. 2022. Т. 21, № 3. С. 572–603. DOI: 10.15622/ia.21.3.5
5. Манойлов В.В., Бородинов А.Г., Сараев А.С., Петров А.И., Заруцкий И.В., Курочкин В.Е. Алгоритмы обработки изображений в секвенаторе ДНК "Нанофор СПС" // Журнал технической физики. 2022. Т. 92, № 7. С. 985–992. DOI: 10.21883/JTF.2022.07.52655.318-2
6. Li L., Speed T.P. An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing // Electrophoresis. 1999. Vol. 20, iss. 7. P. 1433–1442. DOI: 10.1002/(SICI)1522-2683(19990601)20:7<1433::AID-ELPS1433>3.0.CO;2-0
7. Whiteford N., Skelly T., Curtis Ch., Ritchie M.E., et al. Swift: primary data analysis for the Illumina Solexa sequencing platform // Bioinformatics. 2009. Vol. 25, iss. 17. P. 2194–2199. DOI: 10.1093/bioinformatics/btp383
8. Ye Ch., Hsiao Ch., Bravo H.C. BlindCall: ultra-fast base-calling of high-throughput sequencing data by blind deconvolution // Bioinformatics. 2014. Vol. 30, iss. 9. P. 1214–1219. DOI: 10.1093/bioinformatics/btu010
9. Cacho A. Base-Calling of High-Throughput Sequencing Data Using a Random Effects Mixture Model (A Dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy UNIVERSITY OF CALIFORNIA RIVERSIDE). <https://escholarship.org>
10. https://en.wikipedia.org/wiki/Chargaff%27s_rules
11. Манойлов В.В., Бородинов А.Г., Сараев А.С., Петров А.И., Заруцкий И.В., Курочкин В.Е. Алгоритм с использованием метода главных компонент для машинного обучения в задаче построения последовательности нуклеотидов в секвенаторе "Нанофор СПС" // Научное приборостроение. 2023. Т. 33, № 2, С. 35–48. URL: <http://iairas.ru/mag/2023/abst2.php#abst4>

**Институт аналитического приборостроения РАН,
Санкт-Петербург**

Контакты: Манойлов Владимир Владимирович,
manoilov_vv@mail.ru

Материал поступил в редакцию 06.09.2024

CORRECTION OF FLUORESCENT SIGNALS DISTORTED BY OVERLAPPING SPECTRAL BANDS IN THE "NANOFOR SPS" SEQUENCER

**V. V. Manoilov, A. G. Borodinov, A. Ya. Loginov,
A. I. Petrov, I. V. Zarutsky, V. E. Kurochkin**

Institute for Analytical Instrumentation of RAS, Saint Petersburg, Russia

An important stage in the technology of massively parallel sequencing of nucleic acids is the process of measuring fluorescence signals. The spectral composition of the fluorescence signal carries information about the type of the fluorescence object (nucleotide) included in the DNA sample analyzed by the sequencer. The more nucleotides are correctly recognized during the reading process, the more accurate the result of reading the entire genome will be. The phenomenon of overlapping spectral components of the fluorescence signals of nucleotides A, C, G and T (cross-talk) is one of the negative factors that prevents reliable reading of the genome. To combat this phenomenon, a color correction algorithm is needed. This algorithm is decisive in determining the type of nucleotide located in these coordinates (cluster) in a given reading cycle. The paper considers a color correction algorithm that allows one to determine a crosstalk matrix using the linear regression method. Based on this matrix, an inverse matrix is determined, the use of which allows one to significantly reduce the influence of spectral band overlap on the accuracy of fluorescence signal measurement evaluation. The algorithm has been tested on experimental data and has shown high reliability of the results.

Keywords: nucleic acid sequencing, mathematical data processing, crosstalk phenomenon, crosstalk matrix, linear regression, least squares method nucleic acid sequencing, mathematical processing and classification of multivariate data, principal component analysis, machine learning

REFERENCES

1. Kurochkin V.E., Alekseev Y.I., Petrov D.G., Evstratov A.A. [Domestic devices for molecular genetic analysis: developments of the IAP RAS and SINTOL LLC]. *Izvestiya Rossiiskoi Voenno-meditsinskoi akademii* [Russian Military Medical Academy Reports], 2021, vol. 40, no. 3, pp. 69–74. DOI: 10.17816/rmmar76918 (In Russ.).
2. *Polnogenomnyi sekvenator Nanofor SPS* [Nanofor SPS Full Genome Sequencer]. URL: <https://lab.scienceid.net/catalog/9824/28806/> (accessed 10.12.2023). (In Russ.).
3. Manoilov V.V., Borodinov A.G., Zarutsky I.V., Petrov A.I., Kurochkin V.E. [Algorithms of Processing Fluorescence Signals for Mass Parallel Sequencing of Nucleic Acids]. *Trudy SPIIRAN* [SPIIRAS Proceedings], 2019, vol. 18, no. 4, pp. 1010–1036. (In Russ.). DOI: 10.15622/sp.2019.18.4
4. Borodinov A.G., Manoilov V.V., Petrov A.I., Zarutsky I.V., Kurochkin V.E., Saraev A.S. [Machine Learning in Base-Calling for Next-Generation Sequencing Methods]. *Informatika i avtomatizatsiya* [Informatics and Automation], 2022, vol. 21, no. 3, pp. 572–603. DOI: 10.15622/ia.21.3.5 (In Russ.).
5. Manoilov V.V., Borodinov A.G., Saraev A.S., Petrov A.I., Zarutsky I.V., Kurochkin V.E. [Image processing algorithms in the DNA Sequencer "Nanofor SPS"]. *Zhurnal tekhnicheskoi fiziki* [Journal of Technical Physics], 2022, vol. 92, no. 7, pp. 985–992. DOI: 10.21883/JTF.2022.07.52655.318-2 (In Russ.).
6. Li L., Speed T.P. An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis*, 1999, vol. 20, no. 7, pp. 1433–1442. DOI: 10.1002/(SICI)1522-2683(19990601)20:7<1433::AID-ELPS1433>3.0.CO;2-0
7. Whiteford N., Skelly T., Curtis Ch., Ritchie M.E., et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics*, 2009, vol. 25, no. 17, pp. 2194–2199. DOI: 10.1093/bioinformatics/btp383
8. Ye Ch., Hsiao Ch., Bravo H.C. BlindCall: ultra-fast base-calling of high-throughput sequencing data by blind deconvolution. *Bioinformatics*, 2014, vol. 30, no. 9, pp. 1214–1219. DOI: 10.1093/bioinformatics/btu010
9. Cacho A. *Base-Calling of High-Throughput Sequencing Data Using a Random Effects Mixture Model* (A Dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy UNIVERSITY OF CALIFORNIA RIVERSIDE). URL: <https://escholarship.org>
10. Chargaff E. *Chargaff's rules*. URL: https://en.wikipedia.org/wiki/Chargaff%27s_rules
11. Manoilov V.V., Borodinov A.G., Saraev A.S., Petrov A.I., Zarutsky I.V., Kurochkin V.E. [Machine learning algorithm for the construction of a nucleotide sequence in the "Nanofor SPS" sequencer using the principal component analysis]. *Nauchnoe Priborostroenie* [Scientific Instrumentation], 2023, vol. 33, no. 2, pp. 35–48. URL: <http://iairas.ru/mag/2023/abst2.php#abst4> (In Russ.).

Contacts: *Manoilov Vladimir Vladimirovich*,
manoilov_vv@mail.ru

Article received by the editorial office on 06.09.2024

INTRODUCTION

The Institute of Analytical Instrumentation of the Russian Academy of Sciences has developed the first domestic software-hardware complex (SHC) for decoding nucleic acid sequences using the massively parallel sequencing method (Nanofor SPS) [1, 2]. The massively parallel sequencing method is also called sequencing technology by synthesis. In the SHC Nanofor-SPS, fluorophore molecules (fluorescent dyes) that label the nucleotides A, C, G, and T are excited by laser radiation. The fluorescence signals of each nucleotide have a wavelength corresponding to the nucleotide type. The fluorescence signals are recorded using four video cameras (according to the number of nucleotide types) [3]. The recorded radiation passes through various light filters corresponding to the wavelengths of the fluorescence signals of each of the four dyes that label the nucleotides — green, orange, red, and dark red. "Various chemical processes involved in sequencing by synthesis technology cause shifts in the values of the recorded intensities, including phasing/prephasing, signal decay, and cross-talk effects" [4]. In this paper, we consider methods for correcting fluorescent signals distorted by the influence of cross-talk.

It is assumed that methods for correcting fluorescent signals distorted by other factors are to be considered in the authors' other paper.

The emission frequencies of the fluorescent dyes used partially overlap, which leads to a correlation of the intensity readings. This in turn leads to the fact that an increase in the signal intensity, for example, in the channel of the nucleotide A, results in the signal intensity growth in the channel of the nucleotide C, and vice versa. The signals of channels A and C, G and T exert the greatest mutual influence on each other in pairs (Fig. 1). This phenomenon can also be called parasitic fluorescence or cross-talk.

Fig. 1. Emission spectra of fluorescent labels and lasers.

In Fig. 1, the letters G, T, A, and C denote the emission spectra of the fluorescent dyes used to stain the corresponding nucleotides, and L1 and L2 are the spectral lines of the 515 and 638 nm laser emitters of the optical system of the Nanofor SPS device. The passbands of the light filters in the recording channels

are shown by shading. The wavelengths of laser excitation of fluorescence were selected based on the usual requirements for the optimal quantum efficiency of each pair of dyes. The excitation efficiency of dye G is expected to be about 92% of the efficiency at the wavelength of the extinction maximum, dye T — about 26%, dye A — about 90%, dye C — about 17%. From the data presented in Fig. 1, it is evident that mutual parasitic influence is present for all four channels. The choice of the passbands of the recording channels' filters is based on ideas of covering such a range of the useful signal spectrum, which intensity exceeds a given threshold, whereas the parasitic influence of other channels is minimal. In this case, these two requirements conflict, so a compromise solution is implemented. Subsequently, after downloading images from the device's video cameras and realizing a number of digital processing procedures for these images, described in works [3, 5] and others, the phenomenon of superposition of the spectral components of fluorescence signals undergoes software correction. Programs implementing the correction of data distorted by crosstalk are a mandatory component of the primary data analysis programs described in works [6–9].

This paper describes an algorithm designed to correct data distorted as a result of the phenomenon of superposition of the spectral components of optical signals recorded by the Nanofor SPS device and shows that the absence of such correction leads to significant errors in the genetic analysis.

DIRECT AND INVERSE CROSSTALK MATRICES

The phenomenon of crosstalk is that the same nucleotide cluster fluoresces simultaneously in all channels. In the absence of crosstalk, the cluster would fluoresce only in one channel, corresponding to the dye that labels the nucleotide currently in the cluster. In other words, the sum of the fluorescence intensity values measured in each channel in a given cluster corresponds strictly to the intensity value in only one channel, which is currently the channel of the useful signal. That is, the fluorescence signal intensities in the remaining channels are zero. Mathematically, the fluorescence signal intensity in the absence of cross-talk can be represented by the relation (1)

$$I_i(x, y) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} I_A \\ I_C \\ I_G \\ I_T \end{bmatrix} + \begin{bmatrix} b_A \\ b_C \\ b_G \\ b_T \end{bmatrix}. \quad (1)$$

In the relation (1):

– $I_i(x, y)$ is the sum of the measured intensity values in all channels by the pixel with coordinates x and y ;

– I_A, I_C, I_G, I_T are the measured net intensity values in each channel;

– b_A, b_C, b_G, b_T are the background values in the CCD matrix.

In the case of crosstalk, the total measured intensity does not any longer correspond to the signal intensity in only one channel, and the relation (1) takes the form (2).

$$I_i(x, y) = \begin{bmatrix} 1 & w_{12} & w_{13} & w_{14} \\ w_{21} & 1 & w_{23} & w_{24} \\ w_{31} & w_{32} & 1 & w_{34} \\ w_{41} & w_{42} & w_{43} & 1 \end{bmatrix} \times \begin{bmatrix} I_A \\ I_C \\ I_G \\ I_T \end{bmatrix} + \begin{bmatrix} b_A \\ b_C \\ b_G \\ b_T \end{bmatrix} = \mathbf{M} \times \begin{bmatrix} I_A \\ I_C \\ I_G \\ I_T \end{bmatrix} + \begin{bmatrix} b_A \\ b_C \\ b_G \\ b_T \end{bmatrix}. \quad (2)$$

In the relation (2): $\mathbf{M}$ is the crosstalk matrix; w_{ij} are the coefficients of the crosstalk matrix, called angular coefficients.

The task of the color correction algorithm is to minimize the influence of crosstalk. The mathematical solution to this problem can be represented in the form (3), as the multiplication of the right-hand side of (2) by the matrix $\mathbf{M}^{-1}$, the inverse of the matrix $\mathbf{M}$.

$$I_{i, \text{corr}}(x, y) = \mathbf{M}^{-1} \times \begin{bmatrix} I_{iA} \\ I_{iC} \\ I_{iG} \\ I_{iT} \end{bmatrix}. \quad (3)$$

In the ratio (3), $I_{i, \text{corr}}(x, y)$ is the corrected total value of the cluster signal intensity with coordinates x and y . To solve this problem, it is initially necessary to determine the values of all the coefficients of the crosstalk matrix.

PAIRWISE INTENSITY SCATTER PLOTS

According to the measured values of fluorescence signals in all four channels, so-called pairwise intensity scatter plots are formed. Figures 2 and 3 show pairs of channels C–A and channels T–G, respectively. Pairwise plots of channels G–A, T–A, and others are not shown, since distortions of fluorescence signal intensities due to the overlap of spectral components in these channels are practically absent.

Each of the two graphs is formed as follows: the abscissa axis shows the fluorescence signal intensity value of the cluster with coordinates x_i and y_i in one

channel, and the ordinate axis shows the fluorescence signal intensity value by the same pixel, but in another channel.

Fig. 2. Dependence of fluorescence signal intensities in channel C on fluorescence signal intensities in channel A. The white outline highlights the data region that is used to determine the parameters of the regression line

Fig. 3. Dependence of fluorescence signal intensities in channel T on fluorescence signals in channel G. The white outline highlights the data region that is used to determine the parameters of the regression line

Obviously, the highest intensity value in the cluster with coordinates x_i and y_i is observed in the useful signal channel. Thus, from the entire set of data displayed on the pairwise scatter plot, two groups are clearly formed; for each of them, the useful signal belongs to only one of the two channels.

ALGORITHM FOR ESTIMATING THE COEFFICIENTS OF THE CROSSTALK MATRIX AND SUBSEQUENT COLOR CORRECTION

- In the detected clusters, selection of the data areas of the intensities, which are distorted by the phenomenon of crosstalk, noise, outliers, etc. This region is used to determine the parameters of the regression line [6].

- Estimation of the cross-talk matrix by linear regression of pairwise graph data.
- Calculation of the inverse matrix for the cross-talk matrix.
- Application of the found inverse matrix to pairwise scatter plot data distorted by cross-talk.

MAIN FUNCTIONS OF THE PROGRAM IMPLEMENTING THE COLOR CORRECTION ALGORITHM

General characteristic

Before implementing the color correction algorithm in the software development environment, it is necessary to make several clarifications regarding the application of this algorithm to the Nanofor SPS SHC. The input data of the color correction algorithm is an array of intensities of the detected fluorescence signals, which are obtained in clusters with coordinates corresponding to the template coordinates [3]. To construct the template, an algorithm based on convolution starting with the second derivative of a two-dimensional Gaussian peak is used. As a result of applying the convolution algorithm, the background practically removes from the detected fluorescence signals [5]. In the course of describing the correction algorithm, the mutual influence of channels A and C is primarily considered. Then the mutual influence of channels G and T will be considered. In this regard, the final input data are presented as a two-line array of experimentally obtained intensity values in channels A and C. In this case, the general color correction problem can be brought from form (3) to form (4).

$$I_{i,\text{corr AC}}(x, y) = \mathbf{M}_{AC}^{-1} \times \begin{bmatrix} I_{IA} \\ I_{IC} \end{bmatrix}, \quad (4)$$

$$\mathbf{M}_{AC} = \begin{bmatrix} 1 & w_{12} \\ w_{21} & 1 \end{bmatrix}. \quad (5)$$

It becomes obvious that the output data of the algorithm is a two-line array of corrected input data. The general concept of the algorithm is as follows.

- Creating a subroutine that selects the areas of a given intensities in the detected clusters that are distorted by the phenomenon of crosstalk and that eliminates the influence of noise and the baseline — **BuildPrc**.
- Creating a subroutine for estimating the slope angle of the linear regression polynomial of the input data — **SlopeEstimateN**.
- Implementation of the full algorithm (joint work of **SlopeEstimateN** and **BuildPrc**).
- Finding the correction matrix $\mathbf{M}$ and transforming it into an inverse matrix $\mathbf{M}^{-1}$.

- Final testing of the algorithm using the reference genome of the Phix174 bacteriophage.

Subroutine BuildPrc

The main purpose of this subroutine is to form a data vector that can be used to calculate the slope coefficient of the regression line.

This vector is formed from the input data vector, distorted by crosstalk, noise, and other factors interfering with useful information. According to the data in [6], clusters (pixels) with an intensity value below the 60th and above the 90th percentiles cannot be used to correctly estimate the slope. Therefore, in order to estimate the pure influence of crosstalk without extraneous noise, clusters with intensity values that meet the criterion of noise and other interfering factors must be removed from the input data vector. This is what the subroutine **BuildPrc** performs. After removing the "bad" clusters, the section of intensity values in the useful signal channel is divided into several equal intervals (bins). In each interval (bin), the average intensity value is taken along the axis of the useful signal channel, and the minimum along the axis of the parasitic signal channel [6]. These are the intensity values in the vector, which is then processed by the subroutine **SlopeEstimateN**. In Figs. 2 and 3, the data used to determine the regression line parameters are highlighted in white.

Fig. 4 shows the regression line (dotted line) plotted based on the procedure of estimating percentiles, dividing into intervals (bins) and finding the minima in these intervals using the subroutine "**BuildPrc**" for the data presented in Fig. 2. The regression line was constructed using 100 points, the number of which corresponded to the number of intervals (bins).

Fig. 4. Regression line (dotted line) plotted after extracting data using the BuildPrc subroutine for the data shown in Fig. 2.

The angular coefficient is the slope angle tangent of the regression line shown in Fig. 4, which is equal to 0.6032, and accordingly, the slope angle α is equal to 0.5428 radians. Then, using the **SlopeEstimateN** subroutine and the rotation matrix **MrRt** we carry out the rotation of the vector of data, obtained in the **BuildPrc** subroutine, which is rotated by the clockwise slope angle α

$$\mathbf{MrRt} = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix}. \quad (6)$$

The result of the rotation is shown in Fig. 5.

Fig. 5. Data (o) and regression line (dashed line) shown in Fig. 4, and data (dots), regression line (solid line) are obtained after rotation by an angle α .

Subroutine SlopeEstimateN

This subroutine is, in fact, the core of the algorithm. Its main purpose is to find the slope of the input data vector to which the linear regression (7) is applied

$$\sum_{i=1}^N (y_i - ax_i - b)^2 \rightarrow \min_{a,b}. \quad (7)$$

Searching for the slope is an iterative procedure for approaching the slope of the regression line to a zero value. In each iteration, the rotation matrix by an angle α — **MrRt** (6) is calculated, which is updated in each iteration; α is the angle between the regression line and the horizontal axis.

Rotation by an angle α is the first step of the iterative process, which is performed by the subroutine **SlopeEstimateN**. The next step of this iteration process is to estimate the slope angle α_1 of the obtained regression line. For example, for the regression line shown at the bottom in Fig. 5 (solid line), the slope coefficient — the tangent of the slope angle is 0.0342 and, accordingly, the slope angle α_1 is 0.0342 rad. Then the angles α and α_1 are added together and a rotation is performed by the total angle, the value of which is assigned to the variable α . Then the subroutine **SlopeEstimateN** continues the iteration procedure, performing rotations, calculating the tangents of the slope angles, the values of the angles themselves, and the sum of the angle value obtained at the next iteration step with the angle value obtained at the previous step, taking into account the sign. The iteration procedure ends after the absolute value of the tangent of the slope angle becomes less than the specified threshold value (th). Another parameter for the end of the iteration process is the number of iterations N performed. For the example under consideration, $th = 5 \cdot 10^{-5}$, and $N = 8$.

The tangent of the slope obtained after the iteration procedure is an element w_{21} of the desired matrix **M**_{AC}.

To find the element w_{12} of the sought matrix **M**_{AC} (5), i.e., to estimate the crosstalk from channel C to channel A, another regression must be built, for which the considered subroutines **BuildPrc** and **SlopeEstimateN** are also used. The input data is the dependence of the intensities in channel A on the intensities in channel C but not the dependence of the intensities in channel C on the intensities in channel A, as shown in Fig. 2. That is, what was a function became an argu-

ment, and what was an argument became a function. The dependence of the intensities in channel A on the intensities in channel C is shown in Fig. 6. The white contour shows the data area that is used to estimate the parameters of the regression line.

Fig. 6. Dependence of fluorescence signal intensities in channel A on fluorescence signal intensities in channel C.

The white contour highlights the data area used to determine the regression line parameters

Using the subroutines **BuildPrc** and **SlopeEstimateN**, the regression line slope coefficient was found for the data highlighted in white in Fig. 6. The slope coefficient value is the value of the coefficient w_{12} . For the example under consideration $w_{12} = 0.0421$.

Thus, now, based on the known **M**_{AC} matrix elements, we can find the inverse matrix **M**_{AC}⁻¹ and calculate the corrected intensity values in channels A and C using formula (4).

DISPLAYING THE COLOR CORRECTION RESULTS FOR CHANNELS A–C AND G–T

The corrected data for channels A and C are shown in Fig. 7.

Fig. 7. Dependence of fluorescence signal intensities in channel C on fluorescence signal intensities in channel A after correction for the influence of crosstalk.

We have considered the correction of the effect of crosstalk on the fluorescence signal intensities in channels A and C.

The correction of the effect of crosstalk on the fluorescence signal intensities in channels G and T is performed similarly using the above-described subroutines **BuildPrc** and **SlopeEstimateN**. The corrected data for channels G and T are presented in Fig. 8.

Fig. 8. Dependence of fluorescence signal intensities in channel T on fluorescence signal intensities in channel G after correction of the effect of crosstalk.

We have considered the procedure for correcting the effect of crosstalk on the fluorescence signal intensities in channels A–C and G–T for the measurement results in one cycle. The correction of the effect of crosstalk is performed similarly and sequentially in all cycles of the experiment.

A sequence of nucleotide letters for the studied sample is constructed based on the corrected values of fluorescence signal intensities using the **BaseCalling** procedure and genome assembly software. Let us compare the results of genome assembly for the original and cross-talk-corrected measured fluorescence signal intensities.

COMPARISON OF THE RESULTS OF GENOME ASSEMBLY FOR THE CORRECTED AND UNCORRECTED DATA

The genome assembly and its characteristics were assessed for the original and cross-talk-corrected measured fluorescence signal intensities using the **Ugene** and **Qualimap** software programs. Using these programs, the percentage composition of the A, C, G, and T nucleotides in the sample containing the Phix174 bacteriophage genome was determined. This genome contains 5386 nucleotides with a known percentage content of each nucleotide. This genome is a standard for checking the genetic sequence obtained in sequencers. The content of individual nucleotides in this genome is as follows: A — 24.0%, C — 21.5%, G — 23.3%, T — 31.2%. [10]. As a result of the experiment, an assessment of the nucleotide percentage for the fluorescence signal intensities corrected for the effect of crosstalk was performed according to the algorithm discussed above revealed that the percentage of nucleotides was as follows: A — 23.5%, C — 21.5%, G — 23.6%, T — 31.4%, which corresponds to the standard Phix174 genome.

Fig. 9 shows the result of the percentage ratio of individual nucleotides for the original data on fluorescence signal intensities, which were corrected for the effect of crosstalk. Fig. 9 shows the percentage ratio of nucleotides A, C, G and T depending on the measurement cycle number in the sequencer.

Fig. 9. Percentage of A, C, G, and T nucleotides depending on the measurement cycle number in the sequencer when using the spectral band overlap correction algorithm.

As can be seen from the data presented in the figure, the percentage total content of G and C nucleotides in almost all cycles is approximately 44%, which corresponds to the content of these nucleotides in the Phix174 standard.

Fig. 10 shows the result of the percentage content of individual nucleotides for the fluorescence signal intensities for which no correction for the effect of crosstalk using the algorithm discussed above was performed.

Fig. 10. Percentage of A, C, G, and T nucleotides depending on the measurement cycle number in the sequencer when using the spectral band overlap correction algorithm.

As can be seen from the data presented in the figure, the percentage total content of G and C nucleotides in almost all cycles exceeds 65%. This fact indicates that the fluorescence signal intensities in channel C were distorted by the fluorescence signal intensities in channel A.

A comparison of the results presented in Fig. 9 and Fig. 10 shows the efficiency of the considered algorithm.

It should be noted that the use of crosstalk correction makes it easier for us to form a training sample when applying machine learning in the Nanofor SPS sequencer software [4, 11]. In this case, the parameters of the models become less correlated. On the other hand, we can assign the correction itself to machine learning methods.

CONCLUSIONS

1. The phenomenon of crosstalk in fluorescence signal measurements leads to erroneous estimates of signal intensities in adjacent sequencer channels, distorting subsequent identification of the nucleotide letter code.

2. To correct erroneous fluorescence signal intensities, it is necessary to determine the crosstalk matrix. Based on this matrix, an inverse matrix is determined, making it possible to significantly reduce the effect of spectral overlapping bands on the accuracy of fluorescence signal measurement assessment.

3. The algorithm for finding the elements of the crosstalk matrix includes the following operations:

- a) using an estimate of the 60th and 90th percentiles, the range of fluorescence signal intensities in the channel is determined, which distorts the intensity values in the adjacent channel due to the phenomenon of spectral band overlap;
- b) the resulting range of signal intensities is divided into a number of equal intervals;
- c) the minimum intensity is determined for each interval;
- d) the obtained values of the minima in each interval are used to construct a regression line in the iteration procedure;
- e) the angular coefficient, which is an element of the crosstalk matrix, is defined.

4. The phenomenon of crosstalk in the Nanofor SHS sequencer distorts the intensity values only in two pairs of channels: A–C and G–T. In other pairs of channels, such as A–G, A–T, C–G, C–T, etc., there is

practically no distortion of the intensity values caused by the crosstalk phenomenon. This circumstance allows us to simplify the crosstalk matrix and use two 2×2 matrices to correct the influence of the phenomenon of crosstalk: one matrix for correcting distortions in channels A–C, and the another for channels G–T.

5. The result of the genetic analysis of the Phix174 standard using fluorescence signal intensities uncorrected for the influence of crosstalk showed that the

percentage of C and G nucleotides is significantly distorted, while, for the corrected intensities, the percentage of C and G nucleotides corresponds to the standard.

6. Methods for correcting fluorescence signals distorted by phasing / prephasing effects were not considered in this work, although these methods are also based on the estimation of data parameters using linear regression. It is planned to consider these methods in the next work.