УДК 543.07, 543.08 **DOI** 10.15622/sp.2019.18.4.1010-1036

В.В. Манойлов, А.Г. Бородинов, И.В. Заруцкий, А.И. Петров, В.Е. Курочкин

АЛГОРИТМЫ ОБРАБОТКИ СИГНАЛОВ ФЛУОРЕСЦЕНЦИИ МАССОВОГО ПАРАЛЛЕЛЬНОГО СЕКВЕНИРОВАНИЯ НУКЛЕИНОВЫХ КИСЛОТ

Манойлов В.В., Бородинов А.Д., Заруцкий И.В, Петров А.И., Курочкин В.Е. **Алгоритмы обработки сигналов флуоресценции массового параллельного секвенирования нуклеиновых кислот.**

Аннотация. Определение нуклеотидной последовательности ДНК или РНК, содержащих от нескольких сотен до сотен миллионов звеньев мономеров позволяет получить подробную информацию о геноме человека, животных и растений. Расшифровывать структуру нуклеиновых кислот научились достаточно давно, однако первоначально методы расшифровки были низко производительными, неэффективными и дорогими. Методы расшифровки нуклеотидной последовательности нуклеиновых кислот принято называть методами секвенирования. Приборы, предназначенные для реализации методов секвенирования, называются секвенаторами. Секвенирование нового поколения, массовое параллельное секвенирование — это родственные термины, описывающие технологию высокопроизводительного секвенирования ДНК, при котором весь человеческий геном можно секвенировать в течение одного-двух дней. Предыдущая технология, используемая для расшифровки генома человека, потребовала более десяти лет, чтобы получить окончательные результаты.

В Институте аналитического приборостроения РАН разрабатывается аппаратно-программный комплекс для расшифровки последовательности нуклеиновых кислот патогенных микроорганизмов методом массового параллельного секвенирования.

Программное обеспечение, входящее в состав аппаратно-программного комплекса играет существенную роль в решении задач расшифровки генома.

Цель статьи — показать необходимость создания алгоритмов программного обеспечения аппаратно-программного комплекса для обработки сигналов, получающихся в процессе генетического анализа при решении задач расшифровки генома, а также продемонстрировать возможности этих алгоритмов. В работе рассмотрены основные проблемы обработки сигналов и методы их решения. В их числе: автоматическая и полуавтоматическая фокусировка, коррекция изображения фона реакционной ячейки, обнаружение изображений кластеров, оценка координат их положений, создание шаблонов кластеров молекул нуклеиновых кислот на поверхности реакционной ячейки, коррекция влияния интенсивностей соседних оптических каналов и оценка достоверности результатов генетического анализа.

Ключевые слова: секвенирование нуклеиновых частот, алгоритмы обработки сигналов флуоресценции отдельных нуклеотидов нуклеиновых кислот, анализ параметров изображений, оценка достоверности результата генетического анализа.

1. Введение. Метод массового параллельного секвенирования относится к технологии секвенирования нового поколения (СНП) [1-3]. Технология методов СНП позволяет «прочитать» единовременно сразу множество участков генома, что является главным отличием от более ранних методов секвенирования. СНП осуществляется с помощью удлинения цепей фрагментов нуклеиновых кислот.

В разрабатываемом аппаратно-программном комплексе (АПК) молекулы флуорофора, которыми помечены нуклеотиды, возбуждаются под действием лазерного излучения. Сигнал флуоресценции регистрируется с помощью четырех видеокамер (по числу типов нуклеотидов). Регистрируемое излучение пропускается через различные светофильтры, соответствующие длинам волн флуоресценции каждого из четырех красителей, которыми специфично помечены нуклеотиды.

Таким образом, каждая из видеокамер регистрирует изображения кластеров молекул ДНК, на конце которых расположены нуклеотиды определенной «буквы». Требуемое количество генетической информации за один запуск АПК составляет 7.5 миллиардов пар нуклеотидов. Длина прочтения фрагмента ДНК должна быть не менее 250 пар нуклеотидов. Поэтому необходимое число кластеров, которое должно быть обработано, составляет примерно 30 миллионов.

Для применяемых в АПК видеокамер, имеющих разрешение 9 мегапикселей, на одной фотографии помещается примерно 800 тысяч кластеров. Таким образом, чтобы набрать 30 миллионов кластеров, для съема изображений на всей длине реакционной ячейки нам необходима серия из 38 таких фотографий. Каждая такая фотография называется полем зрения (ПЗ). Съем изображений на всей длине реакционной ячейки называется циклом.

После регистрации изображений флуоресцирующих кластеров определенного ПЗ, реакционная ячейка по команде компьютера сдвигается и начинается съемка изображения другого ПЗ. При смене ПЗ участка реакционной ячейки, качество фокусировки изображения может ухудшиться. Поэтому требуется постоянная коррекция фокусировки. После съемки сигналов флуоресценции на всей длине реакционной ячейки по команде компьютера производится переход к следующей стадии. В этой стадии через реакционную ячейку пропускают реагенты, отщепляющие флуорофор и останавливающие процесс синтеза. Затем в реакционную ячейку подаются другие реагенты, с помощью которых начинается новый процесс синтеза — новый цикл. К каждому кластеру, полученному в реакционной ячейке, добавляются новые нуклеотиды. В разрабатываемом АПК должно быть организовано не менее 250 циклов.

Ряд химических процессов, включенных в технологию СНП, искажает регистрируемые сигналы. Программное обеспечение, которое входит в состав АПК, должно вносить поправки в регистрируемые сигналы, искаженные этими процессами.

Под влиянием таких процессов в регистрируемых сигналах возможны следующие изменения, мешающие получить достоверную информацию о геноме:

1. Ферменты могут не сработать, в результате чего основание не включится в цепочку. Кроме того, рост нити ДНК может отставать или перегонять планируемый процесс синтеза по одному основанию, что приводит к неточным показаниям интенсивностей в цикле. Это явление называется фазированием.

2. Перефазирование происходит, когда небольшая часть цепей, забегая вперед, включает сразу два нуклеотида. Перефазирование имеет те же последствия, что и фазирование.

3. Во время процесса секвенирования реакционная ячейка промывается несколько раз, и возможно, что сам секвенируемый материал также будет смыт. Кроме того, неспособность ферментов влиять на синтез нуклеотидов приводит к неактивности в некоторых нитях секвенируемой ДНК. Такие потери вызывают снижение интенсивности сигнала и увеличение шума в процессе секвенирования. Это явление известно как затухание сигнала, и, очевидно, существует корреляция между длиной циклов секвенирования и количеством потерянного материала.

4. Частоты излучения используемых 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.

Для выполнения задач генетического анализа методами СНП в рассматриваемом АПК необходимы следующие алгоритмы и программы.

1. Ввода изображений с видеокамер.
2. Повышения качества фокусировки изображений автоматическими и полуавтоматическими методами.
3. Коррекции фона исходного изображения.
4. Обнаружения и оценки координат локальных объектов флуоресценции на реакционной ячейке.
5. Разделения изображений слипшихся объектов и оценка их координат.
6. Коррекции взаимовлияния флуоресценции в каналах.
7. Обработки бинарных изображений. Определения границ, координат и площадей объектов.
8. Построения таблицы из координат кластеров объектов флуоресценции. Иначе говоря, построения шаблона.
9. Оценки достоверности результатов генетического анализа с учетом влияния химических процессов в технологии СНП, изменяющих значения обрабатываемых сигналов: фазирование, перефазировние, затухание сигнала и другие.

Подобные алгоритмы и программы реализованы в ряде серийных зарубежных секвенаторов, которые частично описаны работах [4, 5]. Для построения первого отечественного секвенатора нового поколения эти алгоритмы должны быть проанализированы и модифицированы для приспособления к особенностям той аппаратуры, которая разрабатывается в настоящее время. При разработке описываемых ниже алгоритмов за основу брались известные. Они апробировались на изображениях, полученных с реальных приборов с программным обеспечением, которое было разработано в ходе выполнения настоящей работы. Ряд описываемых ниже алгоритмов по обработке сигналов, получаемых от секвенаторов, являются оригинальными. К ним относятся алгоритмы по коррекции размытых изображений изучаемых объектов флуоресценции (ОФ), алгоритмы оценки параметров «слипшихся» ОФ и частично алгоритмы по оценке достоверности генетического анализа.

2. Считывание изображений с видеокамер. Используются четыре черно-белых видеокамеры, по одной на каждый канал. Камеры позволяют регистрировать изображения с 4096 градациями серого. Изображения с камер поступают в компьютер в виде растров — массивов двоичных слов. Каждое слово содержит код яркости соответствующего пикселя.

3. Фокусировка размытых изображений. Для получения сфокусированных изображений объектов флуоресценции используются два метода:

1. Фокусировка математическими методами без перестройки фокуса объектива.
2. Фокусировка с программным управлением фокуса объектива.

3.1 Фокусировка математическими методами без перестройки фокуса объектива. С помощью функции ввода изображений загружаем в программу изображение объектов флуоресценции (ОФ). Загруженное изображение представлено на рисунке 1.

Размытое изображение возникает в результате недостаточной фокусировки объектива. Размытое изображение математически получается путем конволюции функции протяженности точки (PSF) [6-9] с исходным изображением. Такое размытое изображение представлено на рисунке 2. Для качественного восстановления исходного изображения важно иметь информацию о параметрах истинной функции протяженности точки.

Для восстановления размытого изображения (рисунок 2) были использованы методы решения обратных задач, изложенные в работах В. С. Сизикова и его соавторов [8-12]. Восстановленное изображение с использованием малоразмерной PSF представлено на рисунке 3.

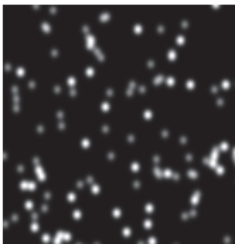

Рис. 1. Исходное изображения ОФ

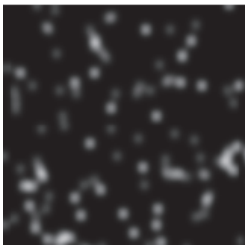

Рис. 2. Модель размытого изображения ОФ

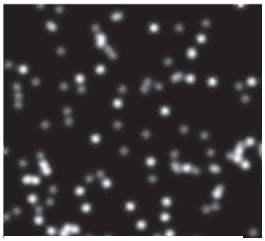

Рис. 3. Восстановленное изображение с использованием методов решения обратных задач

Представленная в данном подразделе информация показывает, что математические методы восстановления размытых изображений позволяют избежать механического управления фокусировкой объектива и снизить время анализа. Кроме указанных алгоритмов восстановления в АПК были реализованы алгоритмы фокусировки размытых изображений с помощью механического перемещения объектива.

3.2 Фокусировка размытых изображений с помощью программы управления объективом. Программное обеспечение рассматриваемого АПК имеет подпрограмму, которая обеспечивает фокусировку объектива по командам от компьютера. Важным параметром этой программы является критерий, с помощью которого имеется возможность установить, что фокусировка объектива выполнена корректно и ее дальнейшее улучшение не имеет смысла. Используется по крайней мере два таких критерия:

1. Ширина профиля одиночного ОФ.
2. Площадь изображения одиночного ОФ.

С помощью программы обнаружения и оценки параметров ОФ, которая будет описана ниже, из всех обнаруженных объектов обнаруживаются ОФ, не содержащие «залипаний». Такие ОФ будем называть одиночными. Пример таких одиночных ОФ показан на рисунке 4.

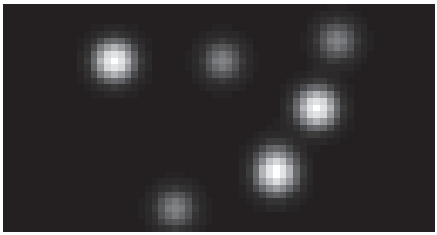

Рис. 4. Пример трех одиночных ОФ (яркие круги)

Рассмотрим теперь профиль одного из одиночных ОФ. Такой профиль представлен на рисунке 5. Ширина представленного профиля на полувысоте равна 5. Теперь рассмотрим профиль этого же ОФ для размытого изображения. Такой профиль представлен на рисунке 6. Ширина представленного профиля на полувысоте равна 8.

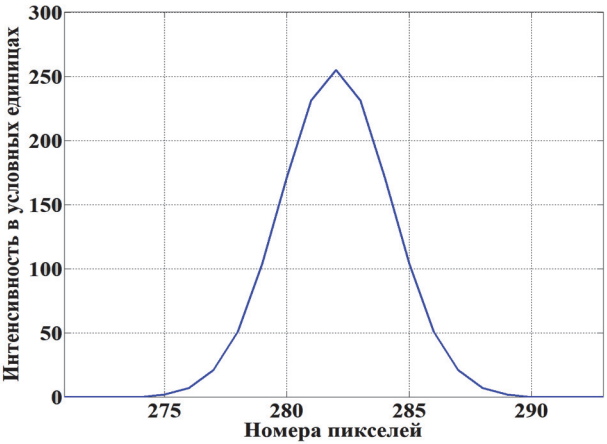

Рис. 5. Профиль одиночного ОФ, сделанный по горизонтальной оси

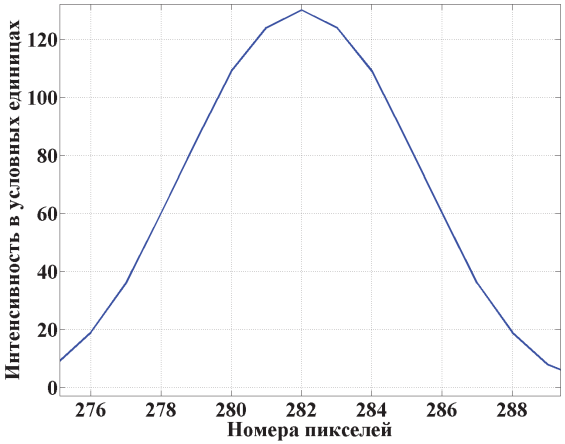

Рис. 6. Профиль одиночного ОФ для размытого изображения, сделанный по горизонтальной оси

Подпрограмма управления фокусировкой должна подавать управляющие команды на перемещения объектива до тех пор, пока ширина профиля одиночного объекта не станет равной ширине профиля одиночного ОФ для неискаженного плохой фокусировкой изображения. Для рассмотренного примера ширина профиля после окончания механического управления фокусом объектива должна быть равна 5 с заданной погрешностью. Выделим теперь из сфокусированного изображения ОФ одиночный объект. Изображение такого объекта в крупном масштабе представлено на рисунке 7, на рисунке 8 — изображение ОФ в размытом изображении.

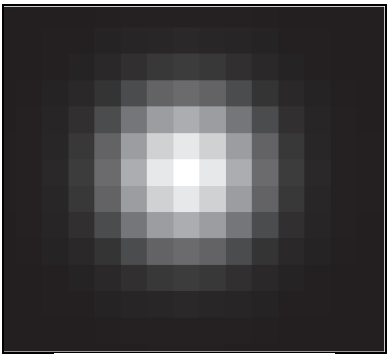

Рис. 7. Одиночный ОФ при сфокусированном изображении

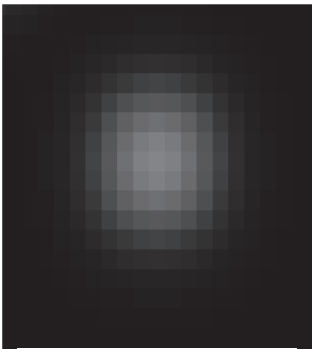

Рис. 8. Одиночный ОФ при размытом изображении

Преобразуем теперь изображения ОФ для сфокусированного и размытого изображений в бинарный формат. С помощью алгоритмов поиска границ и вычисления площадей объектов для бинарных изображений [6, 7, 10-14] найдем границы и вычислим площадь сфокусированного ОФ. Границы сфокусированного ОФ показаны белым цветом на рисунке 9. Площадь сфокусированного ОФ оказалась равной 185 пикселям. Найдем теперь границы и вычислим площадь для размытого ОФ. Границы размытого объекта флюоресценции ОФ показаны белым цветом на рисунке 10. Площадь размытого объекта флюоресценции ОФ оказалась равной 293 пикселя.

Подпрограмма управления фокусировкой должна подавать управляющие команды на перемещения объектива до тех пор, пока площадь одиночного объекта не станет равной с заданной погрешностью площади одиночного ОФ для неискаженного плохой фокусировкой изображения. Для рассмотренного примера площадь одиночного объекта после окончания механического управления фокусом объектива должна быть равна, например,185 пикселям.

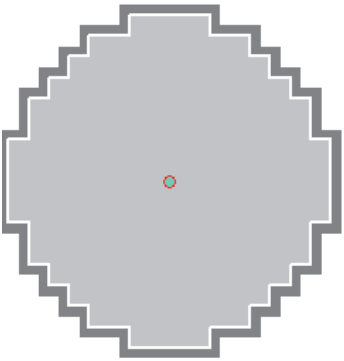

Рис. 9. Бинарное изображение сфокусированного ОФ

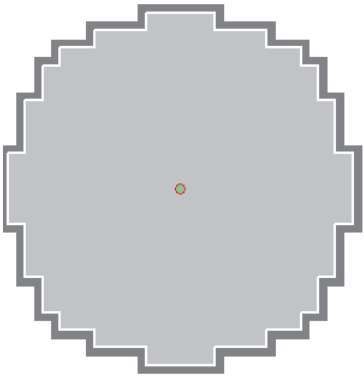

Рис. 10. Бинарное изображение размытого (ОФ) в крупном масштабе

4. Коррекция фона. В полученных с видеокамер изображениях значения интенсивности сигналов флуоресценции в отдельных пикселях зависят от координат конкретного пикселя. Интенсивности сигналов пикселей с координатами близкими к центру изображения более яркие по сравнению с интенсивностями пикселей с координатами далекими от центра. Для снижения влияния координат пикселя на погрешность оценки его интенсивности в программное обеспечение обработки сигналов флуоресценции введена подпрограмма коррекции фона исходного изображения. Под фоном исходного изображения будем понимать изображение, в каждом пикселе которого содержится информация об интенсивности сигнала при отсутствии флуоресценции. В алгоритме этой программы с использованием цифрового фильтра нижних частот создается матрица, каждый элемент которой содержит информацию об уровне интенсивности фона в каждом пикселе при отсутствии сигнала флуоресценции. Значения элементов такой матрицы затем вычитаются из значений интенсивностей сигнала каждого пикселя исходного изображения. На рисунке 11 представлено исходное изображение сигналов флуоресценции, полученное с видеокамеры для канала нуклеотида «A», в виде файла в формате *.tiff. Размер представленного кадра поля зрения 2866 x 2943 пикселей.

Из изображения, представленного на рисунке 11, видно, что интенсивности пикселей в центре изображения более яркие по сравнению с интенсивностями пикселей далеких от центра. Профиль строки изображения, отмеченной на рисунке 11 линией белого цвета, показан на рисунке 12. После обнаружения фона исходного изображения, а затем его

вычитания из исходного изображения получается изображение, в котором влияние фона существенно меньше. Профиль строки, указанной на рисунке 11 линией белого цвета после вычитания фона исходного изображения, показан на рисунке 13.

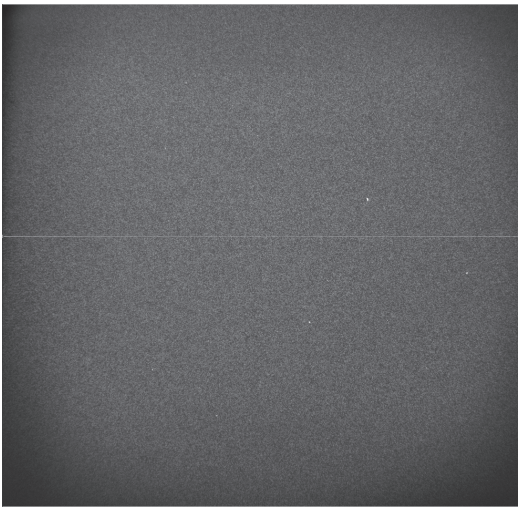

Рис. 11. Исходное изображение ОФ для канала нуклеотида «А»

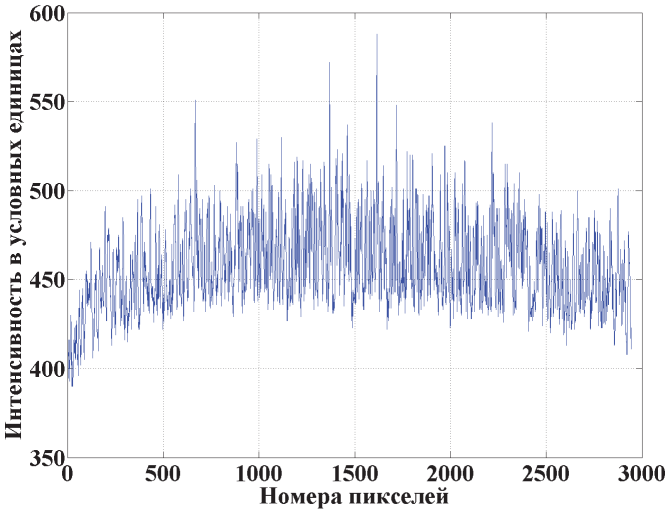

Рис. 12. Профиль строки изображения, показанной на рисунке 11 линией белого цвета

Труды СПИИРАН. 2019. Том 18 № 4. ISSN 2078-9181 (печ.),
ISSN 2078-9599 (онлайн), www.proceedings.spiiras.nw.ru

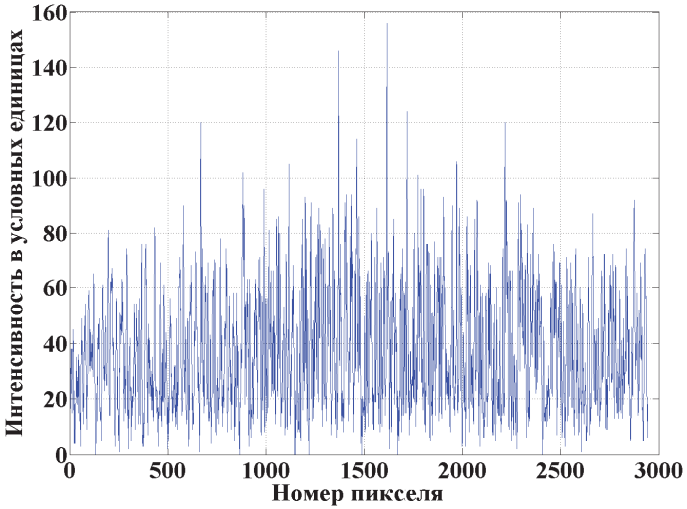

Рис. 13. Профиль строки изображения, показанной на рисунке 11 линией белого цвета после вычитания фона исходного изображения

Фильтр нижних частот для обнаружения фона исходного изображения реализован с использованием морфологической операции эрозии [6, 10, 14].

5. Обнаружение и оценка координат локальных объектов флуоресценции. Строгая формулировка задачи обнаружения звучит следующим образом: найти точки кадра изображения (пиксели), соответствующие центрам отыскиваемых объектов. Тогда зарегистрированное изображение можно трактовать как результат искажения исходного сигнала, представляющего собой совокупность δ-функций с координатами в центрах объектов, а форма зарегистрированного объекта представляет оператор искажения. В этом случае задача обнаружения является задачей восстановления сигнала. Применительно к объектам, подобным рассматриваемым ОФ, задачу восстановления еще называют обостряющей фильтрацией [7, 11, 16]. Типичными фильтрами, используемыми для подобной цели, являются фильтр Винера и фильтр Тихонова. Эти фильтры применяются для восстановления изображений, обусловленных, например, плохой фокусировкой и шумами [7, 9, 12, 15-17]. Фильтр Тихонова более удобен для практических целей, так как он требует для своей реализации меньшей априорной информации, а также он менее критичен к варьированию параметров. Математически фильтр Тихонова (как и фильтр Винера) формулируется в терминах преобразования Фурье, в данном случае двухмерного с независимыми переменными в виде пространственных частот [7, 8, 13, 16, 20].

Для объектов, представленных на рисунке 11, вполне удовлетворительные результаты дает использование в качестве геометрического образа функции, искажающей исходный сигнал, является трехмерный образ второй производной гауссовой функции (1) с шириной, равной примерно половине средней ширины объектов.

$$g(t) = \frac{4x^2}{\mu^4}\exp\left[-\left(\frac{t}{\mu^2}\right)^2\right] - \frac{2}{\mu^2}\exp\left[-\left(\frac{t}{\mu^2}\right)^2\right], \quad (1)$$

где t — независимая переменная, μ — параметр ширины.

Для получения трехмерного образа производиться вращение этой функции по вертикальной оси, проходящей через максимум, в результате которого получается двумерная функция. Затем с помощью преобразования Фурье получается Фурье образ в виде двумерной функции, которая используется в формуле алгоритма обостряющей фильтрации.

Следует добавить, что перед процедурой обостряющей фильтрации к рассматриваемым изображениям следует применить медианную фильтрацию [6, 9, 13, 16, 20]. Медианная фильтрация необходима для того, чтобы дефектные пиксели на видеокамерах не влияли бы на качество обостряющей фильтрации.

Рассмотрим теперь профиль строки на рисунке 11 с номером 1433 до и после процедуры обостряющей фильтрации для столбцов с номерами от 1000 до 1250. К этому профилю сначала была применена процедура вычитания фоновой составляющей, а потом процедура обостряющей фильтрации. На рисунке 14 исходный сигнал показан сплошной линией, а сигнал после обработки по алгоритму обостряющей фильтрации показан пунктирной линией. По оси абсцисс номера пикселей. По оси ординат амплитуда сигнала после нормализации на максимальную интенсивность сигнала в изображении. Из рисунка 14 видно, что большая часть «сдвоенных» (слипшихся) пиков разделились.

После операций вычитания фона и обостряющей фильтрации, решаются задачи поиска и оценка координат положений ОФ. В задачу поиска входит обнаружение объектов и определение координат их центров. Обнаружение объектов представляет собой операцию выделения областей изображения, принадлежащих отыскиваемым объектам [14-17]. Эта операция является принципиально пороговой, поэтому важной задачей является определение порога, который бы позволил надежно отделить «сигнал» (объект) от помех. Для поиска порога используется метод гистограмм распределений интенсивностей сигналов [6, 7].

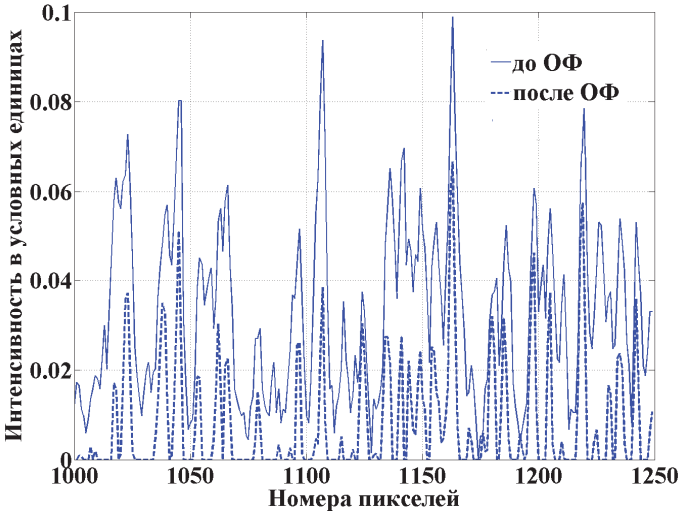

Рис. 14. Сравнение профилей сигналов до и после обостряющей фильтрации

6. Разделение изображений слипшихся объектов и оценка их координат. Теперь рассмотрим результат работы программы обостряющей фильтрации на примере двумерных объектов. Возьмем объекты флуоресценции, расположенные с одинаковыми координатами по вертикальной оси, а по горизонтальной оси их координаты отличаются менее чем на удвоенную ширину объекта на полувысоте. Пример изображений таких слипшихся объектов показан на рисунке 15. После работы обостряющегося фильтра, слипшиеся объекты разделились и их отдельные составляющие стали уже, как показано на рисунке 16.

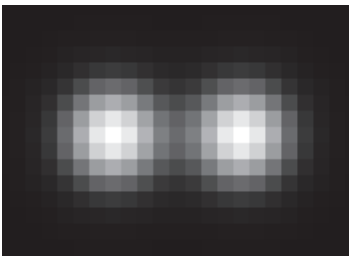

Рис. 15. Пример двух слипшихся объектов

Рис. 16. Пример разделения слипшихся объектов после работы программы обостряющей фильтрации

После операции обостряющей фильтрации не все слипшиеся объекты удается разделить. Для поиска и дальнейшего разделения слипшихся объектов применяются повторная операция обостряющей фильтрации, но с более узким ядром, а затем ряд операций морфологической обработки. Для примера на рисунке 17 показаны сильно слипшиеся объекты, а результат обработки их изображений по алгоритму обостряющей фильтрации показан на рисунке 18.

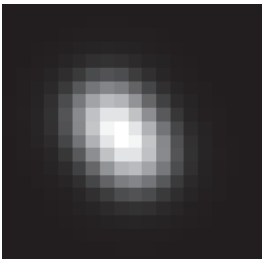

Рис. 17. Пример двух сильно слипшихся объектов

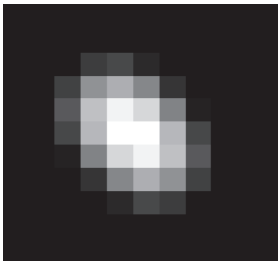

Рис. 18. Пример двух сильно слипшихся объектов после обостряющей фильтрации

В качестве «контрпримера» алгоритмов обнаружения ОФ и разделения слипшихся объектов кратко рассмотрим алгоритм обработки этих данных по алгоритму свертки с ядром в виде многочленов Эрмита:

$$H_n(t) = (-1)^n \exp(t^2) \frac{d^n \exp(-t^2)}{dt^n}, \qquad (2)$$

где n — порядок многочлена Эрмита, t — независимая переменная.

При четном n многочлен $H_n(t)$ симметричен (содержит только четные степени t), при нечетном — многочлен $H_n(t)$ антисимметричен (содержит только нечетные степени x). В настоящей работе использовались многочлены Эрмита 2 порядка. Для получения трехмерного образа так же, как и для второй производной гауссовой функции, производилось вращение функции, которая представлена формулой (2) по вертикальной оси, проходящей через максимум. В результате получалась двумерная функция. Сравнения алгоритмов на основе сверток с ядром в виде многочлена Эрмита (2) и второй производной гауссовой функции (1) показало, что применение таких алгоритмов позволяет увеличить отношение сигнал/шум от 8 до 10 раз в зависимости от диаметра ОФ в исходном изображении. Применение в качестве ядра свертки в виде многочленов Эрмита позволяет увеличить отношение сигнал шум

на 10-20% больше по сравнению с ядром в виде второй производной гауссовой функции. Для задачи разделения слипшихся объектов алгоритм на основе свертки со второй производной гауссовой функции, по сравнению с алгоритмом с использованием многочленов Эрмита, более чувствителен к шумам. Точность оценки координат центров слипшихся ОФ для этих двух алгоритмов примерно одинакова, но алгоритм с использованием многочленов Эрмита имеет большую погрешность в оценке интенсивностей ОФ. В разрабатываемом программном обеспечении алгоритм на основе второй производной гауссовой функции используется для обнаружения и оценки параметров ОФ в тех циклах процесса секвенирования, где отношение сигнал/шум высокое, а алгоритм на основе многочленов Эрмита когда отношение сигнал/шум низкое, что имеет место в циклах процесса секвенирования с высоким порядковым номером. Как было отмечено выше, с увеличением порядкового номера цикла сигналы флуоресценции уменьшаются (затухают), а шум увеличивается.

7. Обработка бинарных изображений. Определение границ, координат и площадей объектов. Как было сказано выше, в результате пороговой обработки формируется битовая карта, которая является бинарным изображением, показанным на рисунке 19. Пиксели, принадлежащие объектам на этом изображении, равны нулю и представлены черным цветом, а фону единице и представлены белым цветом.

Для определения границ, координат и площадей были использованы алгоритмы, описанные в работах [6, 7], а также в работах [9-11]. Для представленного на рисунке 11 исходного изображения канала A размер массива обнаруженных ОФ равен 225056. Объекты являются мечеными, а их границы записаны в элементах матрицы, количество строк которой равно количеству обнаруженных ОФ, а в двух столбцах h и v записываются координаты центройд обнаруженных объектов. Фрагмент обнаруженных объектов представлен на рисунке 20.

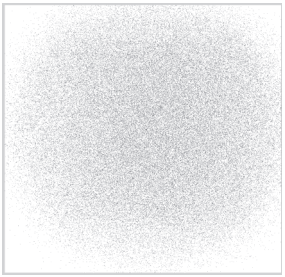

Рис. 19. Бинарное изображение, полученное в результате пороговой обработки исходного изображения канала «A»

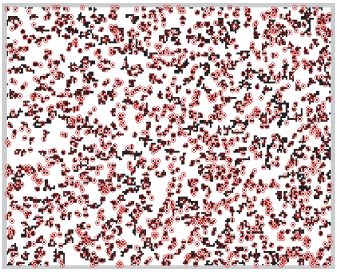

Рис. 20. Фрагмент обнаруженных объектов. Центры кружков соответствуют их координатам по горизонтальной и вертикальной осям

Центры кружков соответствуют их координатам по горизонтали: h и по вертикали: v. Как видно из рисунка 20, часть объектов оказалась не разделенными, то есть слипшимися объектами. Для их разделения используется дополнительная программа «разделения», основанная на методах водораздела (watershed) по алгоритмам, описанным в работах [6, 7, 18-20].

8. Коррекция взаимовлияния флуоресцирующих каналов. Коррекция взаимовлияния каналов осуществляется путем вычисления корреляционных коэффициентов и использования этих коэффициентов в формулах компенсации взаимовлияния данных в различных каналах, которые приводятся в работе [21].

Корреляция сигнала, компенсация и разделение применяются к изображениям для устранения перекрестных помех между флуоресцентными сигналами. Основная часть перекрестных помех объясняется значительной корреляцией между каналами A и C и каналами G и T соответственно.

Другой проблемой является корреляция данных между циклами. По мере выполнения циклов, корреляция между циклами возрастает для всех каналов. Это явление, вероятно, происходит из-за накопления флуорофоров.

По ходу выполнения циклов интенсивность фоновой флуоресценции возрастает, а интенсивность ОФ пропорционально уменьшается. Чтобы восстановить ОФ высокой интенсивности, рассчитываются коэффициенты корреляции Пирсона, которые используются в фазовой коррекции.

Чтобы устранить влияние одного сигнала на другой (перекрестные помехи), корреляция определяется между двумя изображениями по формуле (3), также из сигнала вычитается средняя величина по ПЗ.

Затем вычитаются пропорциональные вклады (корреляционное усредненное изображение), что приводит к коррекции спектрального перекрытия.

$$r_{\mathrm{Q,R}} = \frac{\sum_{i=1}^{m} Q_i R_i - n\bar{Q}\bar{R}}{(n-1)S_Q S_R}, \qquad (3)$$

где Q и R представляют разные типы нуклеотидов, i — соответствующие местоположения пикселей, S_Q и S_R представляют стандартное отклонение Q и R, n — размер выборки.

Чтобы сравнить сигналы в стандартизованном формате, среднее значение интенсивности должно быть вычтено из каждого значения пикселя.

Корреляция рассчитывается для каждой пары каналов и применяется обратно к изображениям, так что коэффициенты оставшихся каналов вычитаются пропорционально значениям каждого изображения соответственно. Значения каждой пары каналов корректируются в соответствии с уравнениями (4) при незначительной корреляции между другими комбинациями каналов после устранения основных эффектов корреляции (A / C, G / T).

Формулы, используемые для коррекции сигналов:

$$\begin{aligned} A_{corrected} &= A - \overline{A} - r_{A,\mathrm{C}}C, \\ C_{corrected} &= C - \overline{C} - r_{\mathrm{C,A}}A, \\ G_{corrected} &= G - \overline{G} - r_{G,T}T, \\ T_{corrected} &= T - \overline{T} - r_{T,G}G. \end{aligned} \tag{4}$$

Применение этого метода компенсации взаимной засветки для всех циклов позволяет получить максимальное разделение сигналов и минимизацию влияния перекрестных помех.

9. Построение таблицы координат кластеров ОФ (Построение Шаблона). Шаблоном будем называть матрицу размером N x 2, где N — количество объектов в одном поле зрения во всех четырех каналах: A, C, G, T для первых 4-х циклов. В каждом цикле 4 изображения, соответствующие каждому из каналов. Каждая строка матрицы состоит из двух чисел: 1) горизонтальная координата объекта h и 2) вертикальная координата объекта v. Для формирования шаблона используются бинарные изображения, в которых координаты объектов были найдены по алгоритмам, описанным выше. Каждый отдельный объект на бинарном изображении отображается одним пикселем черного цвета, его горизонтальная и вертикальная координаты записываются в матрицу.

Шаблон определяет позиции каждого кластера в ПЗ, и используется в качестве эталона для последующей регистрации интенсивностей кластеров. Алгоритм генерации шаблона и его структурная схема приведены в [22].

В полученный шаблон кроме информации о координатах, обнаруженных ОФ, еще записывается информация об интенсивностях ОФ. Таким образом, создается файл с расширением *.cif [22], в котором каждой точке шаблона заданного цикла приписывается информация о том, какому нуклеотиду соответствует конец фрагмента ДНК, «прикрепившейся» к определенной точке реакционной ячейки.

Решение о том, нуклеотид какого канала соответствует данной точке шаблона, принимается по результатам сравнения интенсивностей обнаруженных ОФ и выбирается такой канал (A, C, G или T),

интенсивность сигнала в котором больше в определенной точке шаблона. Интенсивности других каналов также записываются в файл с расширением *.cif [22].

Данные об интенсивностях ОФ в других каналах нужны в дальнейшем, чтобы оценить вероятность правильного решения о том, что данной точке шаблона соответствует выбранный нуклеотид.

Результатом работы программ обнаружения, оценки параметров центров ОФ и построения шаблона является таблица, в которой для каждой пары координат h и v обнаруженных кластеров ОФ записываются «буквы» нуклеотидов, определенных в каждом цикле по результатам сравнения интенсивностей.

Предположим, что были выявлены координаты кластеров всех нуклеотидов в определенном шаблоне в 28 циклах. Предположим, что координаты кластеров всех нуклеотидов известны. Рассмотрим пример таблицы, в которой каждому кластеру будет соответствовать определенная последовательность нуклеотидов, полученная в 28 циклах.

Для 20 пар координат, найденных из изображений всех четырех каналов, ниже представлена таблица 1, в четвертом столбце которой приведена последовательность нуклеотидов, полученных в 28 циклах.

Таблица 1. Координаты кластеров и последовательности нуклеотидов

Номер кластера	h	v	Последовательность нуклеотидов
1	336	403	GACTGGTATTCCGCACCAGGTCTGGCCA
2	216	262	TTGTCCATTAGGCCCCACAAGGGCGGG
3	128	385	CCGTCGTCGTTACGGCCCCCGATAGTCG
4	5	16	GCTATGGATGCCCGGTCGCCGGCCCCA
5	266	398	AAGAGGGGTCTGGTCTCTTCACGGGCCT
6	140	183	AGTCACGGCTAGAGGGCCCGGGAGGCC
7	473	16	AGGAGCCGGGTGCATCGGGGCGCGCCC
8	453	336	AGGCGCCCCACGGAGCGGCTGTGAAAC
9	196	70	TACAAACGTGTCCCGTCGGGTCTAGTCC
10	12	65	GGCAACTTCTATGTTACACCCTCGGGCG
11	336	273	TTTGGCACGGGGTGTAGTCTGGGGGGG
12	419	115	TGGTCAGCGTCCGGGGACCACTTGAGA
13	486	146	GTGCTTCGGTCTTCGTTAAGACAGCTGG
14	28	181	ACCGCCACCGGAAAGGTACAGCGAGAA
15	291	174	ACCGCCACCGGAAAGGTACAGCGAGAA
16	58	105	CAGCCCTCCGAAGTGGGTCTAGACGCG
17	29	333	GAACGCCCCCCGGGCTCGCTCGCTGCCC
18	225	183	CGATGTGCGCGGCCCTACCCGTAAACA
19	291	174	GGGCTAGTCCATTCCATACCAAGCCCCC
20	343	274	ACCCGTGGGGGGCGCCACACAGGAGCA

Данная таблица приведена здесь для примера. В реальной работе количество строк в подобной таблице соответствует количеству обнаруженных кластеров в шаблоне для определенного ПЗ. Количество строк в такой таблице может находиться в диапазоне от 600000 до 900000.

10. Оценка достоверности результатов генетического анализа. Процесс секвенирования всегда сопряжен с различного рода ошибками. В секвенаторах СНП и подобных им они происходят во время фазы, когда необходимо распознать «помеченные» нуклеотиды, то есть понять каким цветом и с какой интенсивностью светятся кластеры из многократного клонирования сегментов ДНК. Проблема в том, что из-за несовершенства остальных этапов процесса секвенирования кластеры никогда не светятся одним светом, это всегда смесь всех четырех цветов с различной интенсивностью.

Задача выбора наиболее интенсивного компонента ДНК и оценка принятия решения о включении этого компонента в геномную последовательность называется **base calling** (распознание нуклеотидов). В результате секвенирования каждому нуклеотиду каждой последовательности определенной точки шаблона (рида) программа обработки определяет вероятность того, что этот нуклеотид распознан правильно.

Точность определения каждого нуклеотида (base-call) представлена в виде показателя качества (**quality score**) и соотносится с каждым определенным нуклеотидом. Показатели качества являются логарифмическими функциями вероятности неверного определения нуклеотидного основания. Эти статистические показатели далее используются в процессе выравнивания (alignment) полученных последовательностей нуклеотидов (ридов) с известным эталонным геномом, например в процессе таких видов анализа, как анализ ChIP-seq [23] или RNA-seq [24-26].

Следовательно, точность и качество процедуры base calling могут напрямую влиять на результаты дальнейшего анализа последовательностей ДНК.

Определим математические обозначения для модели base calling. В таблице 2 даны обозначения, которые будут использоваться для описания модели base calling. В таблице 3 приведен пример того, как четверки интенсивностей представлены в математических обозначениях.

Все методы base calling работают со значениями интенсивности, полученными во время секвенирования. После рассмотрения различных методов base calling [23, 24] становится ясно, что в методах, используемых для моделирования интенсивностей, есть немало общего.

Таблица 2. Обозначения, которые будут использоваться для описания единой статистической модели base calling

Параметр	Размерность	Описание
i		Индекс прочтений (рид индекс) $i = 1,2,\ldots,N$
j		Индекс циклов $j = 1,2,\ldots,J$
k		Индекс каналов $k = A,C,G,T$
Z_i	$4 \times J$	Интенсивности после коррекции
B	$4 \times J$	Коррекция фона
Y_i	$4 \times J$	Наблюдаемые интенсивности
M	4×4	Матрица взаимовлияния каналов
X_i	$4 \times J$	Истинные интенсивности
Ph	$J \times J$	Фазинг-префазинг матрица
D	$J \times J$	Матрица затухания сигнала
E_i	$4 \times J$	Матрица ошибок

Таблица 3. Пример представления наблюдаемых интенсивностей согласно введенным обозначениям

Нуклео-тиды	Цикл						
	1	2	3	4	5	6	7
A	Yi1A	Yi2A	Yi3A	Yi4A	Yi5A	Yi6A	Yi7A
C	Yi1C	Yi2C	Yi3C	Yi4C	Yi5C	Yi6C	Yi7C
G	Yi1G	Yi2G	Yi3G	Yi4G	Yi5G	Yi6G	Yi7G
T	Yi1T	Yi2T	Yi3T	Yi4T	Yi5T	Yi6T	Yi7T
A	-17.7	16.5	847.7	1077.6	1044.7	1039.9	17.4
C	9.2	34.5	651.8	835.4	754.6	708.4	38.1
G	1121	956	-6.4	15.4	9.9	3.9	37.2
T	588.9	495	14.8	3.6	5.4	25.6	639.2

Методы, которые используются в алгоритмах base calling, варьируются от параметрических до непараметрических и статистических моделей, основанных на полностью эмпирических методах машинного обучения. Попытаемся определить общую модель, которая объединяет подавляющее большинство методов base calling. Определим следующую обобщенную модель (5) для base calling как:

$$Z_i - B = Y_i = MX_iPhD + E_i. \quad (5)$$

Для решения подобной матричной системы предполагается итеративно применять взвешенный метод наименьших квадратов. При этом критически важными являются выбор начального приближения и критерии останова итеративного процесса. Далее предполагается реализация для оценки Phred Quality Score [21] подхода на основе логистической регрессии [25, 26].

В работе [26] предлагаются и сравниваются три алгоритма оценки quality core, основанные на логистической регрессии. Для проверки этих алгоритмов используется обучающее множество, для которого авторы взяли информацию о нуклеотидных последовательностей, полученную на первом ПЗ. В качестве обучающей последовательности была использована информация от 30000 кластеров, которые были образованы в 101 цикле. Таким образом, для создания обучающего множества была использована последовательность примерно из 3 миллионов нуклеотидов.

Обозначим обучающее множество $S_i = S_1, S_2, \ldots, S_n$, где S_i принимают значения A, C, G, T. Для рассмотренного примера n=101.

Обозначим I_i индикатор ошибки в определении $S_i (i = 1,2,\ldots n)$

$$I_i = \begin{cases} 1 & \text{если последовательнсть } S_i \text{ определена корректно} \\ 0 & \text{в другом случае} \end{cases}.$$

Пусть q_i — показатель качества (quality core) S_i:

$$\begin{cases} q_i = -10 \log_{10} \varepsilon_i \\ \varepsilon_i = \Pr(I_i = 0 \mid F_i = X_i) \end{cases}$$

где ε_i — вероятность ошибки процедуры base-calling и F_i — вектор характеристик, влияющих на ошибку определения нуклеотида, X_i — истинные интенсивности, полученные в результате решения системы на основе матричной системы (5). Задача определения вектора F_i — отдельная проблема, критичная для правильности оценки ошибки. Мы используем вектор характеристик F_i, предложенный в работе [26]. Модель логистической регрессии служит для оценки вероятности правильности определения S_i.

$p(x_i; \beta) = 1 - \varepsilon_i = \Pr(I_i = 1 \mid F_i = X_i; \beta)$, где β — параметр, подлежащий определению. В качестве модели $p(X_i; \beta)$ предполагается использовать логистическую функцию:

$$\log\left(\frac{p(X_i;\beta)}{1-p(X_i;\beta)}\right) = X_i^T\beta;$$

$$p(X_i;\beta) = \frac{1}{1+\exp(-X_i^T\beta)}.$$

Логарифмическая функция правдоподобия в процедуре base calling записывается в следующем виде:

$$L(\beta;X_1,X_2,\cdots,X_n) = \sum_{i=1}^{n}(y_i \log p(X_i;\beta) + (1-y_i)\log(1-p(X_i;\beta)),$$

где y_i представляют значения I_i (0 или 1), а β — вектор неизвестных параметров. Этот вектор β подлежит оцениванию путем максимизации логарифмической функции правдоподобия. При введении регуляризующего фактора задача сводится к L-regularized logistic regression [26].

$$\min_{\beta} -L(\beta;X_1,X_2,\cdots X_n) + \lambda\|\beta\|_1,$$

где $\|\beta\|_1$ — сумма абсолютного значения каждого элемента в β, λ определяется на основе процедуры перекрестной проверки [26].

Задачу предполагается решать итерационным методом Ньютона — Рафсона.

11. Заключение. Рассмотренные алгоритмы представляют собой результаты исследований, проведенных на этапе разработки макета АПК. Для оценки возможностей алгоритмов были использованы изображения, полученные на зарубежных приборах и математические модели изображений совокупностей кластеров ДНК. По результатам испытаний макета АПК в рассмотренный комплекс алгоритмов необходимо будет внести ряд корректив как для задач точного определения координат обнаруживаемых ОФ, так и для задач окончательного построения генома и оценки его достоверности. Для уточнения алгоритмов обнаружения ОФ и оценки их координат необходимы еще точные знания об отношении полезного сигнала к шуму в реальной электронно-оптической системе. Кроме того, для точной оценки координат ОФ необходимо еще учитывать смещения (сдвиг) ПЗ при перемещениях по длине реакционной ячейки, а также смещения в изображения, получаемых в отдельных каналах (A,C,G,T). Для окончательного построения генома необходимы еще выбор и исследование алгоритма выравнивания (alignment) полученных последовательностей нуклеотидов с известным эталонным геномом, а также выработка стратегии тестирования АПК при анализе ДНК биологических объектов различных типов.

Рассмотренные алгоритмы в большей части используют известные математические методы и приемы. Объединение этих методов в единый комплекс позволяет решить ряд важных практических и научных задач по построению последовательностей нуклеотидов анализируемого генома различных объектов. Полученные результаты являются полезными для генетических исследований в различных областях науки и практики. Программное обеспечение, основанное на рассмотренных алгоритмах и применяемое для обработки и анализа результатов геномных исследований, позволит получить достаточно низкую цену в расчете на один нуклеотид.

Литература

1. *Ansorge W.J.* Next-generation DNA sequencing techniques // New Biotechnology. 2009. vol. 25. no. 4. pp. 195–203.
2. *Bentley R.D. et al* Accurate whole human genome sequencing using reversible terminator chemistry // Nature. 2008. vol. 456. no. 7218. pp. 53–59
3. *Shendure J. et al.* DNA sequencing at 40: past, present and future // Nature. 2017. vol. 550. no. 7676. pp. 345.
4. *Nava W.* The Solexa pipline. URL: http//41j.com/blog/wp-content/uploads/2012/04/pipeline.pdf (дата обращения: 13.07.2019).
5. *Dena L.* Introduction to Deep-Sequencing Data Analysis Illumina Primary Analysis Pipeline & Quality Control URL: http://dors.weizmann.ac.il/course/course2017/Dena_IlluminaPrimaryAnalysisPipeline-course2017.pdf (дата обращения: 13.07.2019).
6. *Журавель И.М.* Краткий курс теории обработки изображений. URL: http://matlab.exponenta.ru/imageprocess/book2/49.php (дата обращения: 06.06.2019).
7. *Гонсалес Р., Вудс Р.* Цифровая обработка изображений // М.: Техносфера. 2012. 1104 с.
8. *Сизиков В.С.* Прямые и обратные задачи в восстановления изображений, спектроскопии и томографии с Матлаб // СПб.: Лань. 2017. 412 с.
9. *Sizikov V.S.* Spectral method for estimating the point-spread function in the task of eliminating image distortions // Journal of Optical Technology. 2017. vol. 84. no. 2. pp. 95–101.
10. *Sizikov V.S. et al.* Determining image-distortion parameters by spectral means when processing pictures of the earth's surface obtained from satellites and aircraft // Journal of Optical Technology. 2018. vol. 85. no. 4. pp. 203–210.
11. *Сизиков В.С., Экземпляров Р.А.* Предшествующая и последующая фильтрация шумов в алгоритмах восстановления // Научно-технический вестник информационных технологий механики и оптики. 2014. № 1(89). С. 112–122.
12. *Сизиков В.С., Лавров А.В.* Устойчивые методы математико-компьютерной обработки изображений и спектров // СПб.: Университет ИТМО. 2018. 70 с
13. *Fu G, Shen D., Sabunci M.R.* Machine Learning and Medical Imaging Book // Academic Press. 2016. 512 p.
14. *Живрин Я.Э., Алкзир Н. Б.* Методы определения объектов на изображении // Молодой учёный. 2018. № 7. С. 8–19.
15. *Кулакович А.Ю., Венцов Н.Н.* Краткий обзор и программная реализация избранных методов для деконволюции // Инженерный вестник Дона. 2017. № 4(47). 11 р.
16. *Бардин Б.В., Чубинский-Надеждин И.В.* Обнаружение локальных объектов на цифровых микроскопических изображениях// Научное приборостроение. 2009. Т. 19. № 4. С. 96–102.

17. *Szeliski R.* Concise Computer Vision. An Introduction into Theory and Algorithms // Springer-Verlag. 2014 p. 441.
18. *Najman L., Schmitt M.* Watershed of a Continuous Function // Signal Processing. 1994. vol. 38. no. 1. pp. 99–112.
19. *Roerdink J.B., Meijster A.* Watershed Transform: Definitions, Algorithms and Parallelization Strategies // Fundamenta Informaticae. 2001. vol. 41. no. 1,2. pp. 187–228
20. *Старовойтов В.В., Голуб Ю.И.* Цифровые изображения от получения до обработки // ОИПИ НАН Беларуси. 2014. 202 с.
21. *Kriseman J., Busick C., Szelinger S., Dinu V.* Bing: Biomedical informatics pipeline for Next Generation Sequencing // Journal of Biomedical Informatics. 2010. vol. 43. no. 3. pp. 428–434.
22. On-Instrument Primary Analysis for HiSeq Theory // Operation manual ILLUMINA PROPRIETARY Pub. no. 770-2009-020. 2011.
23. *Cacho A, Smirnova E, Huzurbazar S, Cui X.* A Comparison of Base-calling Algorithms for Illumina Sequencing Technology // Briefings in Bioinformatics. 2015. vol. 17. no. 5. pp. 786–795.
24. *Ledergerber C, Dessimoz C* Base-calling for next-generation sequencing platforms // Briefings in bioinformatics. 2011. vol. 12(5). pp. 489–497.
25. *Mitra A., Skrzypczak M., Ginalski K., Rowicka M.* Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform // PlOS one. 2015. vol. 10. no. 4. pp. e0120520.
26. *Zhang et al.* Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling // BMC Bioinformatics. 2017. vol. 18. no. 1. pp. 335.

Манойлов Владимир Владимирович — д-р техн. наук, доцент, заведующий лабораторией, лаборатория автоматизации измерений и цифровой обработки сигналов, Институт аналитического приборостроения Российской академии наук (ИАП РАН). Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 71. manoilov-vv@mail.ru; Рижский проспект, 26, 190103, Санкт-Петербург, Российская Федерация; р.т.: +7(812)363-0750; факс: +7(812)363-0720.

Бородинов Андрей Геннадьевич — канд. физ.-мат. наук, начальник сектора информационных проектов, АО "Научные приборы". Область научных интересов: математическая статистика, проблемы анализа, обработки и представления данных, искусственный интеллект. Число научных публикаций — 10. borodinov@sinstr.ru; ул. Маршала Говорова, 52, 198095, Санкт-Петербург, Российская Федерация; р.т.: +7 (812) 313-1-555, доб. т.: 407.

Заруцкий Игорь Вячеславович — канд. техн. наук, старший научный сотрудник, лаборатория автоматизации измерений и цифровой обработки сигналов, Институт аналитического приборостроения Российской академии наук (ИАП РАН). Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 41. igorzv@yandex.ru; Рижский проспект, 26, 190103, Санкт-Петербург, Российская Федерация; р.т.: +7(812)363-0750; факс: +7(812)363-0720.

Петров Александр Иванович — канд. техн. наук, заведующий сектором электроники и программного обеспечения, лаборатория методов и приборов иммунного и генетического анализа, Институт аналитического приборостроения Российской академии наук (ИАП РАН). Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 21. fataip@mail.ru; Рижский проспект, 26, 190103, Санкт-Петербург, Российская Федерация; р.т.: +7(812)363-0765; факс: +7(812)363-0720.

Курочкин Владимир Ефимович — д-р техн. наук, профессор, директор, заведующий лабораторией, лаборатория методов и приборов иммунного и генетического анализа, Институт аналитического приборостроения Российской академии наук (ИАП РАН). Область научных интересов: исследования и оптимизация электромиграционных методов анализа, развитие аналитических методик для капиллярного электрофореза, исследование оптических методов детектирования, разработка методов и приборов для ДНК анализа, разработка методик подготовки проб и специализированных реактивов. Число научных публикаций — 200. lavrovas@yandex.ru; Рижский проспект, 26, 190103, Санкт-Петербург, Российская Федерация; р.т.: +7(812)363-0719; факс: +7(812)363-0720.

Поддержка исследований. Работа выполнена в рамках Государственного задания 075-00780-19-00 по теме № 0074-2019-0013 Министерства науки и высшего образования РФ.

DOI 10.15622/sp.2019.18.4.1010-1036

V.V. MANOILOV, A.G. BORODINOV, A.I. PETROV, I.V. ZARUTSKY, V.E. KUROCHKIN

ALGORITHMS OF PROCESSING FLUORESCENCE SIGNALS FOR MASS PARALLEL SEQUENCING OF NUCLEIC ACIDS

Manoilov V.V., Borodinov A.G., Petrov A.I., ZarutskyI.V., KurochkinV.E. **Algorithms of Processing Fluorescence Signals for Mass Parallel Sequencing of Nucleic Acids.**

Abstract. Determination of the nucleotide sequence of DNA or RNA containing from several hundred to hundreds of millions of monomers units allows to obtain detailed information about the genome of humans, animals and plants. The deciphering of nucleic acids' structure was learned quite a long time ago, but initially the decoding methods were low-performing, inefficient and expensive. Methods for decoding nucleotide nucleic acid sequences are usually called sequencing methods. Instruments designed to implement sequencing methods are called sequencers.

Sequencing new generation (SNP), mass parallel sequencing are related terms that describe the technology of high-performance DNA sequencing in which the entire human genome can be sequenced within a day or two. The previous technology used to decipher the human genome required more than ten years to get final results.

A hardware-software complex (HSC) is being developed to decipher the nucleic acid sequence (NA) of pathogenic microorganisms using the method of NGS in the Institute for Analytical Instrumentation of the Russian Academy of Sciences.

The software included in the HSC plays an essential role in solving genome deciphering problems. The purpose of this article is to show the need to create algorithms for the software of the HSC for processing signals obtained in the process of genetic analysis when solving genome deciphering problems, and also to demonstrate the capabilities of these algorithms.

The paper discusses the main problems of signal processing and methods for solving them, including: automatic and semi-automatic focusing, background correction, detection of cluster images, estimation of the coordinates of their positions, creation of templates of clusters of NA molecules on the surface of the reaction cell, correction of influence neighboring optical channels for intensities of signals and the assessment of the reliability of the results of genetic analysis.

Keywords: Sequencing of Nucleic Acids, Algorithms for Processing Fluorescence Signals of Individual Nucleic Acid Nucleotides, Analysis of Image Parameters, Assessment of the Reliability of the Result of Genetic Analysis.

Manoilov Vladimir Vladimirivich — Ph.D., Dr.Sci., Associate Professor, Head of Laboratory, Laboratory of Automation of Measurements and Digital Signal Processing, Institute for Analytical Instrumentation Russian Academy of Sciences (IAI RAS). Research interests: the Representation and Processing of Signals and Images in Analytical Devices. The number of publications — 71. manoilov-vv@mail.ru; 26, Rizhskij prospekt, 190103, St. Petersburg, Russian Federation; office phone: +7(812)363-0750; fax: +7(812)363-0720.

Borodinov Andrew Gennad'evich — Ph.D., Head of Sector, Sector of Information Projects, Scientific Instruments Joint Stock Company. Research interests: Mathematical Statistics, Problems of Analysis, Processing and Presentation of Data, Artificial Intelligence. The number of publications — 10. borodinov@sinstr.ru; 52, Marshala Govorova str., 198095, St. Petersburg, Russian Federation; office phone: +7 (812) 313-1-555, ext: 407.

Zarutsky Igor Viacheslavovich — Ph.D., Senior Researcher, Laboratory of Automation of measurements and digital signal processing, Institute for Analytical Instrumentation Russian

Academy of Sciences (IAI RAS). Research interests: the Representation and Processing of Signals and Images in Analytical Devices. The number of publications — 41. igorzv@yandex.ru; 26, Rizhskij prospekt, 190103, St. Petersburg, Russian Federation; office phone: +7(812)363-0750; fax: +7(812)363-0720.

Petrov Alexander Ivanovich — Ph.D., Head of Sector of Electronics and Software, Laboratory of Methods and Instruments for Immune and Genetic Analysis, Institute for Analytical Instrumentation Russian Academy of Sciences (IAI RAS). Research interests: the representation and processing of signals and images in analytical devices. The number of publications — 21. fataip@mail.ru; 26, Rizhskij prospekt, 190103, St. Petersburg, Russian Federation; office phone: +7(812)363-0765; fax: +7(812)363-0720.

Kurochkin Vladimir Ephimovich — Ph.D., Dr.Sci., Professor, Director, Head of the Laboratory, Laboratory of Methods and Instruments for Immune and Genetic Analysis, Institute for Analytical Instrumentation Russian Academy of Sciences (IAI RAS). Research interests: research and optimization of electromigration analysis methods, the development of analytical methods for capillary electrophoresis, the study of optical methods of detec-tion, the development of methods and instruments for DNA analysis, the development of methods for preparing samples and specialized reagents. The number of publications — 200. lavrovas@yandex.ru; 26, Rizhskij prospekt, 190103, St. Petersburg, Russian Federation; office phone: +7(812)363-0719; fax: +7(812)363-0720.

Acknowledgements. This research was performed within the framework of the state assignment 075-00780-19-00 on the topic number 0074-2019-0013 Ministry of Science and Higher Education of the Russian Federation.

References

1. Ansorge W.J. Next-generation DNA sequencing techniques. *New Biotechnology.* 2009. vol. 25. no. 4. pp. 195–203.
2. Bentley R.D. et al Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 2008. vol. 456. no. 7218. pp. 53–59
3. Shendure J. et al. DNA sequencing at 40: past, present and future. *Nature.* 2017. vol. 550. no. 7676. pp. 345.
4. Nava W. The Solexa pipline. Available at: http//41j.com/blog/wp-content/uploads/2012/04/pipeline.pdf (accessed: 13.07.2019).
5. Dena L. Introduction to Deep-Sequencing Data Analysis Illumina Primary Analysis Pipeline & Quality Control Available at: http://dors.weizmann.ac.il/course/course2017/Dena_IlluminaPrimaryAnalysisPipeline-course2017.pdf (accessed: 13.07.2019).
6. Zhuravel I.M. *Kratkiy kurs teorii obrabotki izobrazheniy* [Short course of image processing theory]. Available at: http://matlab.exponenta.ru/imageprocess/book2/49.php (accessed: 06.06.2019). (In Russ.).
7. Gonsales R., Vuds R. *Tsifrovaia obrabotka izobrazhenii* [Digital Image Processing]. M.: Tekhnosfera Publ. 2012. 1072 p. (In Russ.).
8. Sizikov V.S. Spectral method for estimating the point-spread function in the task of eliminating image distortions. *Journal of Optical Technology*. 2017. vol. 84. no. 2. pp. 95–101.
9. Sizikov V.S. *Pryamyye i obratnyye zadachi v vosstanovleniya izobrazheniy* [Direct and inverse problems in image reconstruction, spectroscopy and tomography with Matlab]. SPb.: "Lan'". 2017. 412 p. (In Russ.).
10. Sizikov V.S. et al. Determining image-distortion parameters by spectral means when processing pictures of the earth's surface obtained from satellites and aircraft. *Journal of Optical Technology.* 2018. vol. 85. no. 4. pp. 203–210.

11. Sizikov V.S., Ékzemplyarov R.A. [Pre and post noise filtering in recovery algorithms] *Nauchno-tekhnicheskiy vestnik informatsionnyh tekhnologiy mekhaniki i optiki – Scientific and Technical Journal of Information Technologies, Mechanics and Optics.* 2014. vol. 1(89). pp. 112–122. (In Russ.).
12. Sizikov V.S., Lavrov A.V. *Ustoychivyye metody matematiko-komp'yuternoy obrabotki izobrazheniy i spektro*v [Stable methods of computer-computer processing of images and spectra]. SPb.:Universitet ITMO. 2018. 70p. (In Russ.).
13. Fu G, Shen D., Sabuncu M.R. Machine Learning and Medical Imaging Book. Academic Press. 2016. 512 p.
14. Zhivrin Ya.E., Alkzir N. B. [Methods for determining objects in an image]. *Molodoj uchonyj – Young Scientist.* 2018. vol. 7. pp. 8–19. (In Russ.).
15. Kulakovich A.Y., Ventsov N.N. [Overview and software implementation of selected methods for deconvolution]. *Inzhenernyj vestnik Dona – Engineering Journal of Don.* 2017. vol. 4(47). pp 1–19. (In Russ.).
16. Bardin B.V., Chubinsky-Nadezhdin I.V. [Detection of local objects on digital microscopic images]. *Nauchnoe priborostroyeniye – Nauchnoe priborostroyeniye.* 2009. Issue 19. vol. 4. pp. 96–102. (In Russ.).
17. Szeliski R. Concise Computer Vision. An Introduction into Theory and Algorithms. Springer-Verlag. 2014 p. 441.
18. Najman L., Schmitt M. Watershed of a Continuous Function. *Signal Processing.* 1994. vol. 38. no. 1. pp. 99–112.
19. Roerdink J.B., Meijster A. Watershed Transform: Definitions, Algorithms and Parallelization Strategies. *Fundamenta Informaticae*. 2001. vol. 41. no. 1,2. pp. 187–228.
20. Starovojtov V.V., Golub Yu.I. *Tsifrovyye izobrazheniya ot polucheniya do obrabotki* [Digital images from receipt to processing]. OIPI NAN Belarusi. 2014. 202 p. (In Russ.).
21. Kriseman J., Busick C., Szelinger S., Dinu V. Bing: Biomedical informatics pipeline for Next Generation Sequencing. *Journal of Biomedical Informatics*. 2010. vol. 43. no. 3. pp. 428–434.
22. On-Instrument Primary Analysis for HiSeq Theory. Operation manual ILLUMINA PROPRIETARY Pub. no. 770-2009-020. 2011.
23. Cacho A, Smirnova E, Huzurbazar S, Cui X. A Comparison of Base-calling Algorithms for Illumina Sequencing Technology. *Briefings in Bioinformatics*. 2015. vol. 17. no. 5. pp. 786–795.
24. Ledergerber C, Dessimoz C Base-calling for next-generation sequencing platforms. *Briefings in bioinformatics.* 2011. vol. 12(5). pp. 489–497.
25. Mitra A., Skrzypczak M., Ginalski K., Rowicka M. Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform. *PlOS one.* 2015. vol. 10. no. 4. pp. e0120520.
26. Zhang et al. Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling. *BMC Bioinformatics*. 2017. vol. 18. no. 1. pp. 335.