14

Алгоритмы обработки изображений в секвенаторе ДНК „Нанофор СПС“

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

Институт аналитического приборостроения РАН,
190103 Санкт-Петербург, Россия
e-mail: alex.niispb@yandex.ru

Поступило в Редакцию 16 декабря 2021 г.
В окончательной редакции 8 марта 2022 г.
Принято к публикации 24 марта 2022 г.

Успехи геномного секвенирования невозможны без развития информационных технологий и математических методов по обработке данных для установления различных особенностей в анализируемых объектах (нуклеиновых кислотах) и тенденций их изменений. Объем экспериментальных данных при исследовании генома существенно вырос, и для их обработки требуются новые методы и алгоритмы. Первичным этапом обработки данных приборов для геномного параллельного секвенирования является оценка параметров изображений, получающихся с видеокамер в виде электрических сигналов. Следующим этапом обработки является построение последовательности нуклеотидов по алгоритмам, зависящих от принципа действия прибора для секвенирования нуклеиновых кислот. При выполнении этого этапа важное значение имеют алгоритмы для оценки показателей качества для всех индивидуальных чтений (ридов). Одним из путей оценки качества является использование алгоритмов, основанных на методике анализа *k*-меров. Расчет числа встречаемости *k*-меров при проведении эксперимента на системе параллельного секвенирования дает возможность оценить достоверность проведенного анализа. Рассмотрены алгоритмы обработки данных генетического анализатора.

Ключевые слова: секвенирование, нуклеиновые кислоты, обработка изображений, достоверность генетического анализа.

DOI: 10.21883/JTF.2022.07.52655.318-21

Введение

Секвенирование ДНК — метод, позволяющий получить информацию, важную для различных областей деятельности человека, в том числе и для синтетической биологии, суть которого заключается в определении последовательного расположения нуклеотидов (A, T, G и C) в нуклеиновой кислоте. Изобретение технологий секвенирования, наряду с разработанными методами биоинформатики, внесло большой вклад в анализ генома организма. Выходной файл секвенатора предоставляет нам последовательность четырех нуклеотидов. Разработки Максама и Гилберта (Maxam and Gilbert) и Сэнгера (Sanger) стали знаковыми и открыли возможности для разработки более быстрой и эффективной технологии секвенирования [1,2]. Для современного секвенирования, также называемого технологией секвенирования следующего поколения (NGS), характерна очень высокая пропускная способность и гораздо более низкая стоимость запуска, нежели у технологий предыдущих лет [3].

Работа приборов для проведения массового параллельного секвенирования (МПС) основана на технологии секвенирования путем синтеза — Solexa, состоящей из нескольких этапов: фрагментация ДНК и присоединение адаптеров, пропускание библиотек ДНК через каналы реакционной ячейки, покрытой праймерами, комплементарными концам адаптеров, достраивание второй цепи ДНК методом Bridge-PCR (мостиковая амплификация с использованием полимеразно-цепной реакции) и последующая денатурация. В результате повторения двух последних действий образуются группы идентичных молекул — кластеры. Создание кластеров необходимо для усиления оптического сигнала.

В Институте аналитического приборостроения РАН (ИАП РАН) разрабатывается аппаратно-программный комплекс (АПК) для расшифровки последовательности нуклеиновых кислот (НК) методом массового параллельного секвенирования „Нанофор СПС“. Алгоритмы обработки информации, получающейся в ходе работы АПК, играют существенную роль в решении задач расшифровки генома.

Целью настоящей работы является анализ возможностей алгоритмов для обработки изображений, формируемых в процессе генетического анализа, а также оценки достоверности получаемой последовательности нуклеотидов при решении задач расшифровки генома. Для оценки достоверности получаемой генетической информации используется методика анализа распределения частот встречаемости *k*-меров. С помощью данной методики обработаны данные, полученные на отечественном секвенаторе „Нанофор СПС“.

1. Основные операции при обработке изображений сигналов флуоресценции

В системе параллельного секвенирования прибора „Нанофор СПС“ используются четыре видеокамеры по числу типов нуклеотидов. Каждая из видеокамер настроена на регистрацию одного из типов нуклеотидов: A, C, G или T. Сигнал флуоресценции возбуждается двумя лазерами в определенном диапазоне излучения видимого света. 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“.

В силу конструктивных особенностей прибора имеются ряд технических трудностей в достижении полного геометрического совпадения изображений одного и того же поля зрения объектов флуоресценции. Параметры координат изображений одного и того же объекта флуоресценции в разных камерах могут быть сдвинуты на несколько пикселей. В ходе выполнения настоящей работы были разработаны алгоритмы и программы для коррекции этих сдвигов математическими методами.

При обработке изображений сигналов флуоресценции системы параллельного секвенирования выполняются следующие операции:

1. Обнаружение фоновой поверхности с помощью морфологических алгоритмов.

2. Фильтрация изображений с помощью свертки с Mexicanhat.

3. Алгоритмы обнаружения кластеров объектов флуоресценции.

4. Алгоритмы определения порогов обнаружения.

5. Определение средних значений интенсивностей флуоресценции кластеров.

6. Определение дисперсии шума в сигналах изображений.

7. Коррекция сдвигов изображений, вызванных конструктивными особенностями прибора математическими методами с использованием кросскорреляционных функций.

8. Фокусировка изображений с помощью перемещений объектива.

Важным параметром, влияющим на успешный исход секвенирования, является плотность кластеров. Слишком малое количество кластеров уменьшает объем выходных данных. Слишком большое — вызывает „слипание“ объектов, что влияет на качество данных и в некоторых случаях приводит к провалу эксперимента.

Алгоритмы выполнения операций 1–7 описаны в работах [4,5].

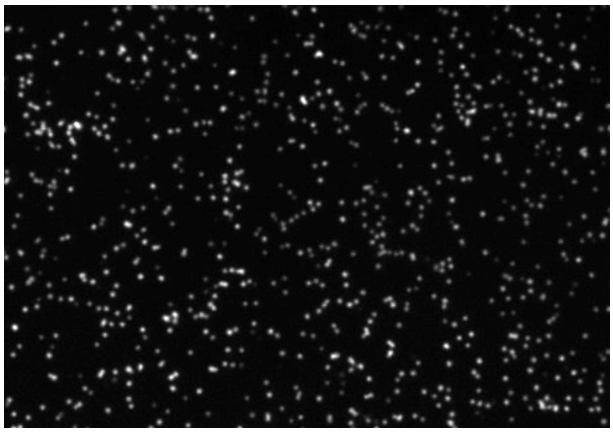

Рис. 1. Исходное изображение сигналов флуоресценции для канала A (аденин).

2. Считывание изображений с видеокамер

Используются четыре черно-белых видеокамеры, настроенных на регистрацию одного из типов нуклеотидов: A, C, G или T. Камеры позволяют регистрировать изображения с 4096 градациями яркости. Изображения с камер поступают в компьютер в виде растров — массивов двоичных слов. Каждое слово содержит код яркости соответствующего пикселя. На рис. 1 представлен фрагмент изображения сигналов флуоресценции для канала A — аденин.

3. Фокусировка размытых изображений

Для получения сфокусированных изображений объектов флуоресценции используются два метода: фокусировка математическими методами без перестройки положения объектива и фокусировка с помощью программы для механического перемещения объектива.

Фокусировка изображений с помощью механического перемещения объектива заключается в определение такого положения объектива, при котором качество фокусировки изображения было бы наиболее высоким, о котором будет сказано ниже.

На рис. 2 и 3 показаны фрагменты изображений до и после процедуры фокусировки.

Для проверки отношения сигнал/шум для изображений с низким и высоким качеством фокусировки была произведена оценка амплитуды сигнала для объекта, указанного на рис. 2 и 3 цифрой *1*. Амплитуда сигнала для этого объекта на изображении с низким качеством фокусировки оказалась равной 200 условных единиц, а на изображении с высоким качеством фокусировки амплитуда сигнала для этого объекта оказалась равной примерно 400 условных единиц. Величина шума определялась на основе среднеквадратичного значения (ско)

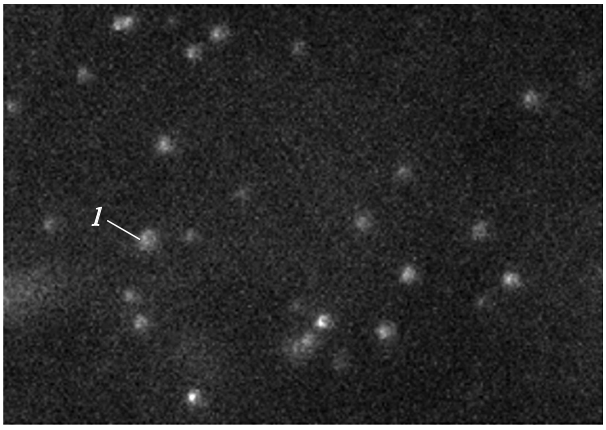

Рис. 2. Изображение фрагмента поля зрения при начальном положении объектива с низким значением параметра качества фокусировки до процедуры фокусировки.

Рис. 3. Изображение фрагмента поля зрения, соответствующее лучшему значению параметра качества фокусировки после процедуры фокусировки.

сигнала в том месте изображений, где отсутствовали объекты. Величины ско для изображений с низким и высоким качеством фокусировки оказались примерно равными 10 условным единицам. Таким образом, отношение сигнал/шум для изображений с низким и высоким качеством фокусировки оказались равными соответственно 20 и 40. Для изображений, представленных на рис. 2 и 3, нормировки на максимум не производилось. Для более наглядной визуальной оценки изображения с высоким качеством фокусировки рис. 3 представлен с более высокой яркостью.

Эффективность качества процедуры фокусировки исследовалась следующим образом. Имелись наборы фотографий различных областей реакционной ячейки, полученных при различном положении объектива, т. е. с разной степенью сфокусированности. Каждый набор содержал фотографии, сделанные при одинаковых условиях (экспозиция и освещение). Затем для каждой из сделанных фотографий производилась оценка качества фокусировки.

Методы для оценки качества фокусировки разрабатываются давно, придумано их много, и они неплохо исследованы, например [6–10]. При выполнении работы, описываемой в настоящей работе, было произведено сравнение 30 методов [6] применительно к изображениям, получаемым в приборе „Нанофор СПС“. При выборе наилучшего метода оценки качества строились функции, в которых по горизонтальной оси откладывался номер изображения с разной фокусировкой, а по вертикальной оси откладывалась оценка качества по соответствующему методу. Полученные функции сравнивались по трем критериям:

1. Функция должна иметь один экстремум.

2. Функция должна давать экстремум для изображения с наилучшей фокусировкой.

3. Робастность, т. е. величина градиента зависимости значения функции от степени сфокусированности изображения.

Абсолютные величины функций фокусировки не важны, так как имеют значения только положение экстремума и величина градиента. Для сравнения величины всех функций нормировались к максимуму функции в точке экстремума, т. е. максимумы были приведены к единице. Для оценки качества фокусировки изображения используется все изображение целиком, какой-либо части (строки, столбца или области) не выделяется.

Лучшие показатели по перечисленным критериям показали метод Vollath′s correlation [7] и дисперсия [10]. В рабочей программе прибора „Нанофор СПС“ для оценки качества фокусировки используется дисперсия.

На рис. 4 и 5 показаны профили одного из объектов флуоресценции до и после проведения процедуры фокусировки соответственно. После процедуры фокусировки

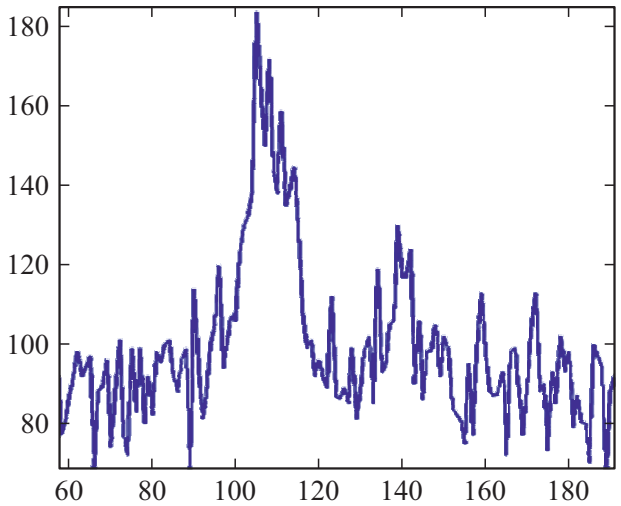

Рис. 4. Профиль одного из расфокусированных объектов флуоресценции, представленных на рис. 2. По горизонтальной оси — номер пикселя в изображении, по вертикальной оси — интенсивность флуоресценции в условных единицах.

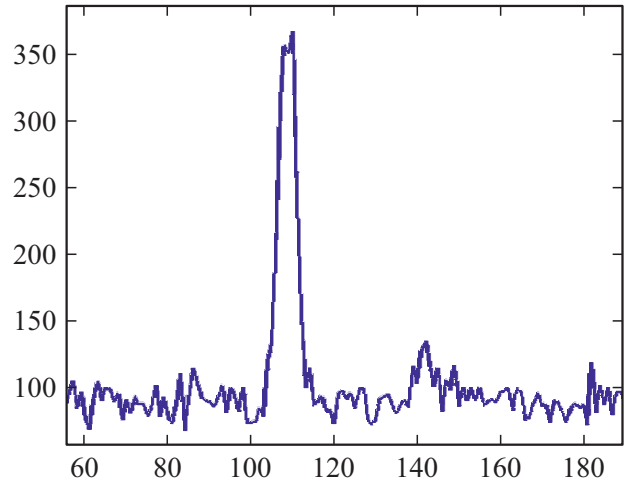

Рис. 5. Профиль объекта флуоресценции, представленного на рис. 3 после проведения процедуры фокусировки.

увеличилась амплитуда сигнала и уменьшилась ширина профиля.

Для дополнительной коррекции фокуса изображений и разделения частично перекрывающихся сигналов объектов флуоресценции была использована обостряющая фильтрация на основе обратной задачи конволюции функции протяженности точки.

Размытое (расфокусированное) изображение математически получается путем конволюции функции протяженности точки (PSF) [11–14] с исходным изображением. Для качественного восстановления исходного изображения важно иметь информацию о параметрах истинной функции протяженности точки. Для восстановления размытого изображения были использованы методы решения обратных задач, изложенные в работах В.С. Сизикова и его соавторов [12–14]. Математические методы восстановления размытых изображений позволяют избежать механического управления фокусировкой объектива и снизить время анализа.

4. Определение сдвигов изображений в различных каналах и в различных полях зрения с помощью кросскорреляционных функций

В ходе выполнения настоящей работы были разработаны алгоритмы и программы для коррекции сдвигов изображений, вызванных конструктивными особенностями прибора. Эти алгоритмы основаны на вычислении кросскорреляционных функций между изображениями, полученными разными видеокамерами, или между изображениями одной видеокамеры, но полученными в разных циклах (сканах) эксперимента. Координаты максимального значения корреляционной функции соответствуют координатам сдвига анализируемых изображений.

Сигналы флуоресценции под действием красителей генерируются для каждого из нуклеотидов в определенном диапазоне длин волн (полос) видимого света. Однако существует такое явление, как перекрытие полос между сигналами различных нуклеотидов. Данное явление приводит к тому, что кластеры, которые, например, „светились“ в канале A будут „светиться“ и в канале C. Это свойство используется для оценки геометрических смещений изображений различных каналов друг от друга. Смещение изображений между различными каналами определяется с помощью кросскорреляционной функции.

Кросскорреляционная функция между двумя изображениями вычислялась по формуле (1) [15]:

$$Y(u,v) = \frac{\sum\limits_{x,y}[f(x,y)-\bar{f}_{u,v}][g(x-y,y-u)-\bar{g}_{u,v}]}{\left\{\sum\limits_{x,y}[f(x,y)-\bar{f}_{u,v}]^2\sum\limits_{x,y}[g(x-y,y-u)-\bar{g}_{u,v}]^2\right\}^{0.5}}, \quad (1)$$

где $f(x,y)$ — двумерная функция первого изображения; $g(x,y)$ — двумерная функция второго изображения; x,y — координаты пикселей изображений; u,v — координаты кросскорреляционной функции, $\bar{f}_{u,v}$, $\bar{g}_{u,v}$ — средние значения функций f и g соответственно.

Определим начало координат двумерного изображения в точке максимума кросс-корреляционной функции двух одинаковых изображений $u=0$, $v=0$. Теперь рассмотрим координаты максимального значения кросскорреляционной функции изображений. Максимальное значение кросскорреляционной функции изображений разных каналов, например, каналов флуоресценции нуклеотидов „a“ и „c“, имеет координаты, соответствующие искомому сдвигу. Например, координаты максимального значения кросскорреляционной функции могут быть равны $x=-2$, $y=8$.

Аналогичным образом рассчитываются координаты сдвига для других каналов и для сдвига изображений одного канала, но регистрируемых в разных циклах эксперимента.

Кроме вычислений кросскорреляционной функции по формуле (1) в программах обработки изображений используется алгоритм более быстрых вычислений на основе прямого и обратного быстрого двумерного преобразования Фурье. Сначала вычисляются двумерные преобразования фурье-функций $f(x,y)$ и $g(x,y)$. Затем фурье-образы этих функций перемножаются, и в соответствии с теоремой Планшереля [16] в результате получается кросскорреляционная функция.

5. Коррекция фона

Под фоном исходного изображения понимается изображение, в каждом пикселе которого содержится инфор-

мация об интенсивности сигнала при отсутствии полезного сигнала. В алгоритме коррекции фона используется цифровой фильтр, основанный на свертке исходного изображение со второй производной двумерной гауссовой функции. Этот алгоритм проще алгоритма коррекции фона с использованием цифрового фильтра нижних частот, основанный на морфологических операциях эрозии и делатации, описанных в работе [4]. Алгоритм на основе свертки со второй производной гауссовой функции (Mexicanhat) уменьшает значение фона практически до нуля, а алгоритм на основе морфологических операций его уменьшает примерно в 5–7 раз. Ширина гауссовой функции, на основании которой вычисляется вторая производная, составляет 0.7 средней ширины анализируемых объектов изображений.

На рис. 6 показаны профили сигнала флуоресценции до проведения операции по коррекции фона, после операции по коррекции фона на основе морфологических операций и с помощью алгоритма свертки. Можно видеть, что профиль пика находится на уровне фона, составляющим примерно 120–140 условных единиц. После проведения операции по коррекции фона с помощью алгоритма свертки уровень фона практически равен нулю. Отрицательные значения сигнала, получающиеся после операции свертки не влияют на точность оценки координат центров объектов флуоресценции и их интенсивностей, но могут ухудшить результаты вторичной обработки информации на этапе получения последовательностей нуклеотидов. Такие значения удаляются при вторичной обработке, примерно также как это делается в программе Swift [17].

Как было показано в работе [4], алгоритм на основе свертки позволяет фильтровать шум и „обострять“ пики, что необходимо для повышения точности оценок

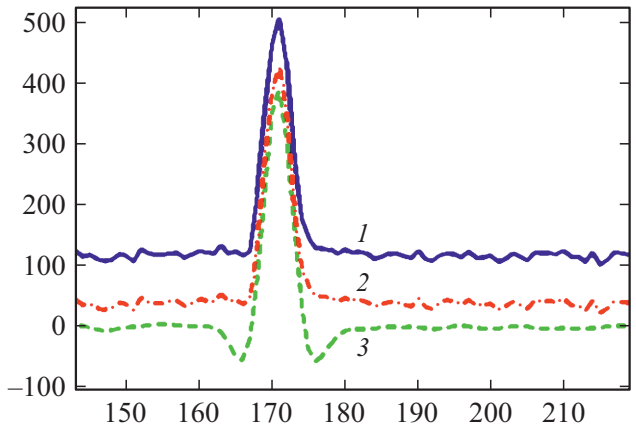

Рис. 6. Профиль объекта флуоресценции до и после операций коррекции фона: *1* сплошная линия — профиль исходного сигнала, *2* линия с точками — профиль сигнала после коррекции фона по алгоритмам с морфологическими операциями, *3* пунктирная линия — профиль сигнала после коррекции фона с помощью алгоритма свертки. По горизонтальной оси — номер пикселя в изображении, по вертикальной оси — интенсивность флуоресценции в условных единицах.

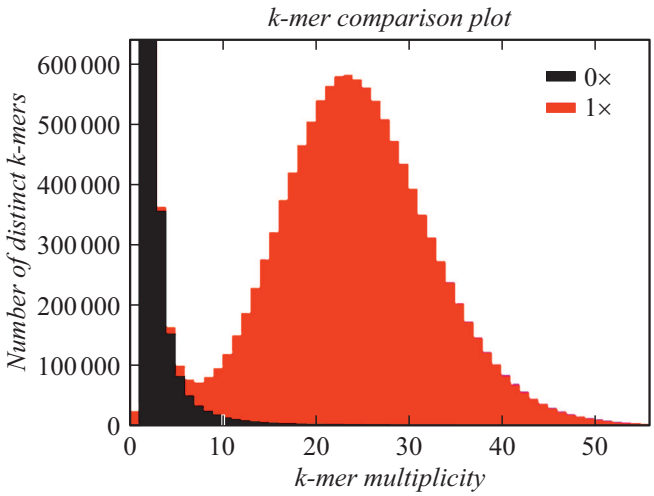

Рис. 7. Близкий к идеальному пример использования *K*-mer, полученный на приборе Нанофор СПС.

координат обнаруженных кластеров. Уменьшение шума и „обострение“ пика алгоритмом свертки показано на рис. 7.

Метод конволюции со второй производной гауссовых пиков используется во многих программах обработки изображений секвенаторов подобного „Нанофор СПС“ типа.

6. Обнаружение и оценка координат локальных объектов флуоресценции

После операций вычитания фона применяется „обостряющая“ фильтрация изображения на основе свертки со второй производной двумерной гауссовой функции. Обнаружение объектов представляет собой операцию выделения областей изображения, принадлежащих отыскиваемым объектам. Эта операция является принципиально пороговой, поэтому важной задачей является определение порога, который позволил бы надежно отделить „сигнал“ (объект) от помех. Для поиска порога используется метод гистограмм распределений интенсивностей сигналов.

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.

В результате пороговой обработки формируется битовая карта, которая является бинарным изображением. Пиксели, принадлежащие объектам на этом изображении, равны нулю, а фон — единице. Координаты объекта определяются на основе максимума интенсивности в обнаруженном объекте. Площадь определяется по количеству пикселей, превышающих половину максимальной интенсивности.

Для определения границ, координат и площадей были использованы алгоритмы, описанные в работах [4,5].

7. Отбраковка „ложных“ кластеров

Одной из основных операций при построении шаблона [4] является совмещение бинарных изображений отдельных каналов сигналов флуоресценции, например, каналов нуклеотидов „a“ и „c“ с помощью двоичной операции ИЛИ. Такое совмещение производится после операции коррекции сдвига изображений, рассмотренной выше. После проведения операции коррекции сдвига в некоторых точках изображения возможны ошибки в определении координат кластеров по x и y из-за того, что полученные значения величин смещения оказываются неодинаковыми для всех точек изображения. Величины этих ошибок не превышают 1 или 2-х пикселей, однако приводят к появлению в шаблоне координат „ложных“ кластеров и увеличивается ошибка правильности построения последовательности нуклеотидов. Для уменьшения такой ошибки в программное обеспечение обработки изображений сигналов флуоресценции введена подпрограмма, которая производит отбраковку „ложных“ кластеров. Если для какой-то точки изображения с координатами x и y в квадратной области, с координатами в четырех углах соответственно $(x-2, y-2)$, $(x+2, y-2)$, $(x-2, y+2)$, $(x+2, y-2)$ обнаруживается более одного кластера, то все кластеры, кроме одного, считаются „ложными“. Программа отбраковки „ложных“ кластеров среди всех кластеров, обнаруженных в указанном выше квадрате, оставляет в шаблоне только один кластер, который дает максимальную интенсивность сигнала флуоресценции из всех обнаруженных в указанном квадрате кластеров.

8. Конечный результат работы программы обработки изображений сигналов флуоресценции

Конечным результатом работы программы обработки сигналов объектов флюоресценции является таблица (фрагмент представлен ниже), в которой содержатся координаты изображений объектов кластеров фрагментов исследуемых нуклеиновых кислот. Каждой объект обнаруженного кластера содержит две координаты: y — горизонтальную координату и x — вертикальную координату на исследуемых изображениях. Каждому кластеру должна соответствовать обнаруженная для него последовательность нуклеотидов в виде латинских букв: A, C, G, T (см. таблицу).

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

9. Алгоритмы оценки достоверности получаемой последовательности нуклеотидов, основанные на оценке встречаемости k-меров

K-мер — это просто последовательность из k-символов в строке (или нуклеотидов в последовательности ДНК в задаче секвенирования). Например, для получения всех k-мер из последовательности нужно получить первые k-символов, затем сместиться на один символ для начала, следующего k-мера и так далее.

Разложение последовательности на ее k-меры для анализа позволяет анализировать этот набор фрагментов фиксированного размера, а не последовательность целиком и это может быть более эффективным подходом. Операции над множествами k-меров выполняются быстрее и проще. Простой пример: чтобы проверить, происходит ли последовательность S из организма A или из организма B, предполагая, что геномы A и B известны и достаточно разные, мы можем проверить, содержит ли S больше k-меров, присутствующих в A или в B.

Практически любой геном содержит повторяющиеся области, однако, начиная с определенного значения k, k-меры определенным образом однозначно идентифицируют его. Если мы посчитаем количество появлений k-мер для достаточно большого k (ограниченного сверху длиной чтения), оказывается, что большинство из них находятся в геноме в единственном экземпляре. Например, если порядок длины генома сравнима с человеческим, вероятность встретить случайную подстроку длины 14 хотя бы один раз составляет 0.975893 [18]. Для $k = 20$ эта же вероятность составляет 0.000909. Для геномов меньших размеров, например, бактерий или грибов, можно выбрать меньшее k, чтобы добиться схожей малой вероятности многократной встречаемости строки.

Анализ распределения частот встречаемости k-меров позволяет находить ошибки сборки в уже сформированных контигах [19,20]. Контигом является набор перекрывающихся последовательностей ДНК-фрагментов, полученных из одного биологического источника (организма, ткани, клетки).

10. Методики работы с *k*-мерами

В работе [18] предложен метод оценки качества геномной сборки, заключающийся установлении соответствия между уникальными *k*-мерами в собранном геноме и *k*-мерами в ридах. Чтением или ридом называется последовательность нуклеотидов одного кластера.

Процедура выглядит следующим образом:

1. Построение гистограммы встречаемости *k*-меров для ридов, полученных при секвенировании исследуемого генома.

2. Выбор некоторой окрестности пика уникальных *k*-меров на гистограмме встречаемости *k*-меров в ридах.

3. Построение гистограммы встречаемости *k*-меров для каждой из полученных сборок.

4. Расчет меры *Q*, как доли различных *k*-меров, взятых из окрестности пика на гистограмме встречаемости *k*-меров в чтениях, среди уникальных *k*-меров для собранного генома). Каждый *k*-мер из множества уникальных *k*-меров ридов проверяется на вхождение во множество уникальных *k*-меров собранного генома.

В работе [21] предложен метод исправления ошибок, оптимизированный для работы с ридами, содержащими как ошибки замены, так и ошибки вставки и удаления. Поскольку ошибки происходят с небольшой вероятностью, вероятность того, что один и тот же *k*-мер будет прочитан несколько раз с одинаковым набором ошибок, очень мала. Из этого вытекает, что те *k*-меры, которые встречаются в наборе ридов мало раз, являются ошибочными („плохими“), остальные же являются реальными подстроками генома („хорошими“).

K-mer спектр — это графическое представление набора данных, показывающее, сколько коротких слов фиксированной длины (*k*-мер) появляется определенное количество раз. Частота встречаемости нанесена на ось *x*, а число *k*-меров на оси *y*.

Мы представляем, сколько элементов каждой частоты в спектре последовательности нуклеотидов (ридов) оказались: 0 — не включены в референтный геном (в нашем случае Phix174), включены 1 раз, 2 дважды и т. д.

На рис. 7 представлен близкий к идеальному пример, полученный на приборе Нанофор СПС. Ошибки в сборке отсутствуют (черный пик вне красного). Основной уникальный контент находится в референтном наборе ровно один раз.

Заключение

Рассмотренные алгоритмы и программы, разработанные на их основе, производят регистрацию и обработку изображений сигналов объектов флуоресценции системы параллельного секвенирования „Нанофор СПС“. Среди этих алгоритмов и программ важное значение имеют алгоритмы автоматической фокусировки изображений, автоматической оценки сдвига изображений, получающихся в различных каналах и циклах секвенирования, программы коррекции фона, обнаружения и оценки параметров координат кластеров в поле зрения и другие.

Для достижения наилучшего качества автоматической фокусировки изображений, получаемых видеокамерами, был выбран критерий вычисления дисперсии амплитуд пикселей изображения как наиболее простой в программной реализации.

Использование алгоритма вычисления координат максимального значения двумерной кросскорреляционной функции двух изображений позволяет выполнить автоматическую оценку сдвига изображений, получаемых видеокамерами в различных каналах секвенатора.

Алгоритм свертки со второй производной гауссовой функции позволяет фильтровать шум, „обострять“ пики и корректировать влияние фона.

При построении последовательности нуклеотидов важное значение имеют алгоритмы для оценки показателей качества для всех индивидуальных ридов. Одним из путей оценки качества является использование алгоритмов, основанных на *k*-мерах.

Основанные на *k*-мерах методы целесообразно использовать для эффективного создания геномных сборок.

Программа KAT (*k*-mer Analysis Toolkit) может анализировать данные секвенирования для определения уровней случайных ошибок, систематических ошибок и контаминации. Информация, полученная в ходе этого анализа, может помочь исследователям решить, следует ли продолжать выполнение последующих задач, таких как сборка генома. Затем KAT может перепроверить проведенную сборку генома, определив полноту и точность сборки без каких-либо внешних справочных данных.

Объединение рассмотренных алгоритмов в единый комплекс позволяет решить ряд важных практических и научных задач по построению последовательностей нуклеотидов анализируемого генома различных объектов. Полученные результаты являются полезными в следующих областях: молекулярная биология, генетика, сельское хозяйство, медицина, охрана окружающей среды и других.

Финансирование работы

Работа выполнена в рамках Государственного задания № 075-00280-21-00 по теме № 0074-2019-0013 Министерства науки и высшего образования РФ.

Конфликт интересов

Авторы заявляют, что у них нет конфликта интересов.

Список литературы

[1] F. Sanger, S. Niclein, A.R. Coulson. Proc. Natl. Acad. Sci. USA, **74**, 5463–5467 (1977). DOI: 10.1073/pnas.74.12.5463

[2] A.M. Maxam, W. Gilbert. Proc. Natl. Acad. Sci. USA, **74** (2), 560–564 (1977) DOI: 10.1073/pnas.74.2.560

[3] Д.В. Ребриков, Д.О. Коростин, Е.С. Шубина, В.В. Ильинский. *NGS: высокопроизводительное секвенирование*, под общей ред. Д.В. Ребрикова (БИНОМ, Лаборатория знаний, М., 2014), 232 с.

[4] В.В. Манойлов, А.Г. Бородинов, И.В. Заруцкий, А.И. Петров, В.Е. Курочкин. Журн. Труды СПИИ РАН, **18** (4), 1010–1036 (2019). DOI: 10.15622sp.2019.18.4.1010-1036

[5] В.В. Манойлов, И.В. Заруцкий. *Обработка сигналов флуоресценции массового параллельного секвенирования нуклеиновых кислот.* Св-во о гос. регистрации программы для ЭВМ. № 2019663248.

[6] S. Pertuz, D. Puig, M.Á. García. Pattern Recognition, **46** (5), 1415–1432 (2012). DOI: 10.1016/j.patcog.2012.11.011

[7] A. Santos, C.O. de Solorzano, J.J. Vaquero, J.M. Pena, N. Mapica, F.D. Pozo. J. Microscopy, **188**, 264–272 (1997).

[8] Yu Sun, S. Duthaler, B.J. Nelson. Microscopy Res. Tech., **65**, 139–149 (2004).

[9] Chun-Hung Shen, H.H. Chen. Robust Focus Measure for Low-Contrast Images. 2006 Digest of Technical Papers Intern. Conf. Consumer Electron., 69–70 (2006). DOI: 10.1109/ICCE.2006.1598314

[10] E. Krotkov, J.-P. Martin. Range From Focus. Proceed. IEEE Intern. Conf. Robotics and Automation, 1093–1098 (1986). DOI: 10.1109/ROBOT.1986.1087510

[11] Р. Вудс, Р. Гонсалес. *Цифровая обработка изображений* (Техносфера, М., 2012), 3-е изд., испр. и доп., 1104 с.

[12] В.С. Сизиков. *Прямые и обратные задачи в восстановления изображений, спектроскопии и томографии с Матлаб* (Лань, СПб., 2017), 412 с.

[13] V.S. Sizikov. J. Opt. Technol., **84** (2), 95–101 (2017).

[14] V.S. Sizikov, A.V. Stepanov, A.V. Mezhenin, R.A. Burlov, R.A. Éksemplyarov. J. Opt. Technol., **85** (4), 95–101 (2018).

[15] J.P. Lewis. *Fast Template Matching, Vision Interface*, 120–123 (1995).

[16] Ж. Макс. *Методы и техника обработки сигналов при физических измерениях: в 2-х томах*, пер. с франц. (Мир, М., 1983), т. 1, 312 с.

[17] N. Whiteford, T. Skelly, Ch. Curtis, M.E. Ritchie, A. Löhr, A.W. Zaranek, I. Abnizova, C. Brown. Bioinformatics, **25** (17), 2194–2199 (2009). DOI: 10.1093/bioinformatics/btp38

[18] К. В. Романенков. *Метод оценки качества сборки генома на основе частот k-меров*, Препринты ИПМ им. М.В. Келдыша, 2017, 11.

[19] T.J. Treangen, D.D. Sommer, F.E. Angly, S. Koren, M. Pop. Current Protocols in Bioinformatics, **11** (11.8), 1–18 (2011).

[20] А.В. Александров, А.А. Шалыто. Научно-технический вестник информационных технологий, механики и оптики, (1), 108–114 (2016).

[21] G. Marcais, C. Kingsford. Bioinformatics, **27** (6), 764–770 (2011).