УДК 57.087, 543.07, 543.08 **DOI** 10.15622/ia.23.4.3

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

АЛГОРИТМЫ ПЕРВИЧНОГО АНАЛИЗА ЛОКАЛЬНЫХ ОБЪЕКТОВ ФЛУОРЕСЦЕНЦИИ В СЕКВЕНАТОРЕ ДНК «НАНОФОР СПС»

Манойлов В.В., Бородинов А.Г., Заруцкий И.В., Петров А.И., Сараев А.С., Курочкин В.Е. **Алгоритмы первичного анализа локальных объектов флуоресценции в секвенаторе ДНК «Нанофор СПС».**

Аннотация. В секвенаторе ДНК «Нанофор СПС», разработанном в Институте аналитического приборостроения РАН, реализован метод массового параллельного секвенирования для расшифровки последовательности нуклеиновых кислот. Этот метод позволяет определять последовательность нуклеотидов в ДНК или РНК, содержащих от нескольких сотен до сотен миллионов звеньев мономеров. Таким образом, имеется возможность получения подробной информации о геноме различных биологических объектов, в том числе человека, животных и растений. Важнейшей частью этого прибора является программное обеспечение, без которого невозможно решение задач по расшифровке генома. Выходными данными оптической детекции в секвенаторе являются набор изображений по четырем каналам, соответствующим типам нуклеотидов: A, C, G, T. С помощью специального программного обеспечения определяется положение молекулярных кластеров и их интенсивностные характеристики вместе с параметрами окружающего фона. В ходе создания программного обеспечения прибора были разработаны алгоритмы и программы обработки сигналов флуоресценции, рассмотренные в работе. Также, для отладки и тестирования рабочих программ созданы модели построения изображений, аналогичных реальным данным, получаемым в ходе работы секвенатора. Данные модели позволили получить значительный массив информации без запуска дорогостоящих экспериментов. За последние годы достигнуты значительные успехи в области машинного обучения, в том числе и в области биоинформатики, что привело к реализации наиболее распространенных моделей и возможности их применения для практических задач. Однако, если на этапе вторичного анализа биоинформационных данных эти методы широко зарекомендовали себя, то их потенциал для первичного анализа остается недостаточно раскрытым. В данной работе особое внимание уделяется разработке и внедрению методов машинного обучения для первичного анализа оптических изображений сигналов флуоресценции в реакционных ячейках. Описаны методы кластеризации и их апробация на моделях и на изображениях, полученных на приборе. Цель этой статьи – продемонстрировать возможности алгоритмов первичного анализа сигналов флуоресценции, получающихся в процессе секвенирования на приборе «Нанофор СПС». В работе описаны основные задачи анализа сигналов флуоресценции и сравниваются традиционные методы их решения с использованием технологий машинного обучения.

Ключевые слова: секвенирование, нуклеиновая кислота, методы обработки сигналов флуоресценции ДНК и РНК, анализ изображений, машинное обучение.

1. Введение. В секвенаторе «Нанофор СПС» реализован метод массового параллельного секвенирования, который еще называют методом секвенирования нового поколения (NGS) [1]. Отличительной

особенностью технологии является возможность анализировать одновременно множество участков генома [2, 3]. В процессе секвенирования используется техника удлинения цепей отдельных частей ДНК и РНК кислот.

В секвенаторе «Нанофор СПС» флуоресцентное химическое соединение (флуорофор или краситель) «присоединяется» к нуклеотидам и может повторно излучать свет при его возбуждении, например, лазерным излучением. Каждый нуклеотид, помеченный красителем, излучает свет на длине волны, соответствующей его типу. После присоединения красителя к фрагментам нуклеиновой кислоты производится возбуждение красителя лазерным излучением. Полученный после возбуждения сигнал флуоресценции проходит через светофильтры разных длин волн. Длины волн светофильтров соответствуют длинам волн, которые излучают нуклеотиды, помеченные красителями. После прохождения через светофильтры сигнал флуоресценции регистрируется видеокамерами. В секвенаторе имеются четыре видеокамеры, каждая из которых фиксирует сигналы определенного типа нуклеотида (канала): аденин – «A», цитозин – «C», гуанин – «G» и тимин – «T».

После регистрации видеокамерами изображений сигналов флуоресценции по всей длине реакционной ячейки происходит переход к следующему этапу. На этом этапе через камеру пропускают реактивы, которые отделяют краситель (флуорофор) и прекращают процесс синтеза. Затем добавляются другие реактивы, чтобы начать новый процесс синтеза – новый цикл.

Программное обеспечение (ПО) секвенатора «Нанофор СПС» решает следующие задачи обработки данных генетического анализа, полученных по результатам экспериментов:

1) Чтение изображений с видеокамер;
2) Фокусировка полученных изображений аппаратными и математическими методами;
3) Исключение фона в исходном изображении;
4) Распознавание и определение характеристик кластеров сигналов флуоресценции на реакционной ячейке;
5) Определение характеристик изображений «слипшихся» кластеров;
6) Исключение взаимовлияния флуоресценции в различных каналах;
7) Оценка качества результатов проведенного эксперимента после коррекции влияния химических процессов изменяющих

значения обрабатываемых сигналов: фазирование, перефазирование, затухание сигнала и др.

Решения задач обработки данных генетического анализа реализованы в программном обеспечении секвенатора «Нанофор СПС» и в нескольких зарубежных приборах, описанных в литературе [4, 5]. При разработке нижеперечисленных методов использовались алгоритмы из работ [6, 7, 8]. В последнее время технологии машинного обучения стали широко применяться для обработки биоизображений. Упомянутые методы были протестированы на изображениях, полученных с реальных приборов, а также на модельных изображениях. Методы построения моделей изображений описываются в настоящей статье.

2. Алгоритм обнаружения сигналов флуоресценции на основе свертки. В секвенаторе «Нанофор СПС» установлены четыре черно-белые видеокамеры, по одной на каждый канал. Эти камеры способны регистрировать изображения с 4096 оттенками серого. Изображения с камер передаются в компьютер в форме растровых массивов двоичных слов, где каждое слово содержит код яркости пикселя. На рисунке 1 представлен фрагмент изображения сигналов флуоресценции для канала «А» – аденин.

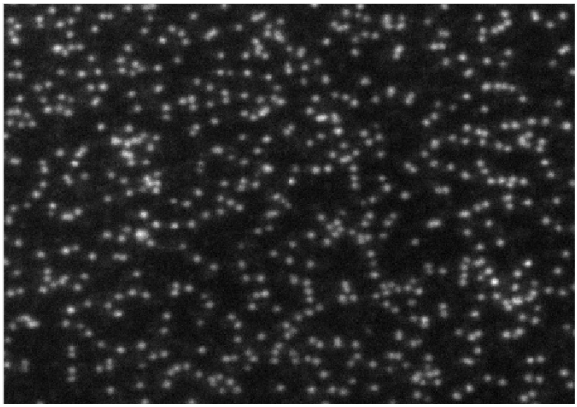


Рис. 1. Фрагмент изображения сигналов флуоресценции для канала «А» (аденин)

Для обеспечения высокого качества решения задач по обработке изображений сигналов флуоресценции, полученные из видеокамер цифровые данные подвергаются медианной фильтрации. Этот шаг помогает избежать негативного влияния «выбросов» регистрируемых сигналов и дефектных пикселей на видеокамерах, искажающих

изображения. Затем решается задача обнаружения кластеров нуклеиновых кислот. Задача обнаружения описывается как поиск точек (пикселей) на изображении, соответствующих центрам кластеров, что позволяет интерпретировать изображение, как результат искажения исходного сигнала различными причинами, например, плохой фокусировкой, шумами и другими причинами, которые могут испортить исходный сигнал [9 – 13]. Исходный сигнал представляет собой набор функций с координатами в центрах кластеров, а форма изображения кластера определяется функцией искажения. Задача обнаружения кластера сводится к задаче восстановления сигнала, решение которой является решением обратной задачи. Одним из методов решения обратных задач является деконволюция, для которой требуется знание формы функции искажения. Математически задача деконволюции решается с использованием прямого и обратного преобразования Фурье в двумерном пространстве частот с независимыми переменными [9 – 13].

Для обнаружения объектов, изображенных на рисунке 1, используется свертка с трехмерным образом второй производной гауссовой функции (1) с уменьшенной шириной, приблизительно равной половине средней ширины самих кластеров.

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

где t – независимая переменная, μ – параметр ширины.

Формула (1) является результатом вычисления второй производной функции $exp\left[-\left(\frac{t}{\mu^2}\right)^2\right]$.

Вторую производную гауссовой функции называют «мексиканской шляпой» (Mexican hat) и используют в программном обеспечении ряда приборов, в получаемых изображениях на которых необходимо обнаруживать сигналы флуоресценции, в том числе и в секвенаторе фирмы «Illumina» [4].

Получение трехмерного образа этой функции производится путем ее вращения по вертикальной оси, проходящей через максимум. Такое вращение формирует трехмерную функцию. После этого с помощью преобразования Фурье получается двумерный Фурье-образ в этой функции. Полученный Фурье-образ используется в алгоритме решения обратной задачи с помощью деконволюции. С помощью деконволюции происходит выделение полезного сигнала из шума,

«сужение» («обострение») обнаруженных сигналов и исключение влияния фоновой составляющей. «Обострение» обнаруженных сигналов необходимо для решения задачи разделения «наложившихся» кластеров.

Условия использования формулы (1) в рассматриваемой задаче следующие: соотношение сигнал шум находится в диапазоне: от 7 до 40, среднее квадратичное значение шума примерно 15 условных единиц, фоновая составляющая в приборе «Нанофор СПС», как правило, представляет собой нелинейную функцию со значениями интенсивностей от 100 до 200 единиц. Исходный размер изображения в приборе «Нанофор СПС» составляет 2000 х 2400 пикселей. В связи с тем, что ширина кластера на его полувысоте составляет от 6 до 15 пикселей, то фоновую составляющую под кластером можно считать линейной.

Величина амплитуды результирующего сигнала после свертки (деконволюции) зависит от значения параметра A в формуле (1). Параметр A подбирался опытным путем для обеспечения амплитуды результирующего сигнала равной или большей амплитуды исходного сигнала. В будущем планируется сделать адаптивный выбор параметра A, как функции экспозиции.

Для обеспечения высокого качества решения обратной задачи с помощью деконволюции к обрабатываемому изображению необходимо предварительно применить медианную фильтрацию. Этот шаг помогает избежать негативного влияния «выбросов» регистрируемых сигналов и дефектных пикселей на видеокамерах, искажающих изображения.

Качественное распознавание и определение характеристик кластеров на изображении затруднено из-за фоновой составляющей. Для решения этой проблемы используется алгоритм, описанный в работе [6]. В процессе вычислений могут возникать отрицательные значения яркости изображения, которые не играют ключевой роли. Фон обычно изменяется плавно и его можно вычислить, усредняя сигнал изображения в окне, скользящем по кадру. Ширина окна должна быть в два раза больше, нежели размер самого крупного кластера на изображении.

Вычисления по алгоритму деконволюции производятся следующим образом. Исходные данные видеоинформации $S(x, y)$ в геометрическом пространстве (x, y) с помощью алгоритма быстрого преобразования Фурье преобразуются в данные видеоинформации в пространстве Фурье $S(u, v)$. Затем данные $S(u, v)$ умножаются на Фурье-образ функции искажения $Fl(u, v)$. Полученное

произведение подвергается обратному преобразованию Фурье и получаются данные $S_p(x, y)$, которые затем сравниваются с порогом. В данном описании x, y – координаты исходного геометрического пространства, u, v – координаты пространства Фурье (частоты).

После выполнения операций по вычитанию фона и применению алгоритма деконволюции происходит решение задач по поиску и оценке положений координат кластеров флуоресценции (КФ). Задача поиска включает в себя обнаружение объектов и определение координат их центров. Обнаружение кластеров представляет собой процесс выделения областей на изображении, соответствующих искомым кластерам [13]. Для реализации этой операции важным моментом является определение порога, который позволит эффективно отделить «сигнал» (объект) от шума. Для определения порога используется метод гистограмм распределения интенсивности сигналов [10].

Величины интенсивностей сигналов различных нуклеотидов (A, C, G, T) отличаются друг от друга, и поэтому значения порога для изображений сигналов флуоресценции каждого из нуклеотидов будут разные. Назовем изображения, полученные для сигналов флуоресценции нуклеотидов A, C, G, T каналами A, C, G, T, соответственно. Для определения значений порога для каждого из каналов строятся гистограммы распределения нормированных на максимальное значение интенсивностей сигналов в каждом пикселе.

Распределение интенсивностей сигналов флуоресценции представляет собой одномодальную ассиметричную функцию. Величины интенсивностей отличные от шума «вносят вклад» в ассиметричную (правую) часть функции. Определение порога с помощью гистограмм, имеющих ассиметричную функцию распределения, немного сложнее, чем для гистограмм, имеющих двухмодальную функцию распределения, которая описывается в работе Отсу [14]. Для двухмодальных функций распределения порог определяется как среднее значение между двумя максимумами в функции распределения. В нашем случае с помощью гистограмм, соответствующих сигналам флуоресценции секвенатора, порог определялся следующим образом: производилась оценка среднеквадратичного значения (СКО) шума. Оценка СКО шума представляет собой значение полуширины на полувысоте пика гистограммы. Величина порога равна произведению коэффициента k на оценку СКО. Коэффициент k подбирался экспериментально путем обработки большого количества изображений сигналов

флуоресценции. Для большинства задач определения порога пригодным оказался коэффициент k=9.

Работа программы обнаружения заканчивается отметкой специальными маркерами (например, кружками) границ обнаруженных кластеров на исходном изображении, например, так, как показано на рисунке 2, а также построением матрицы, содержащей вертикальные (v) и горизонтальные координаты (h) обнаруженных кластеров.

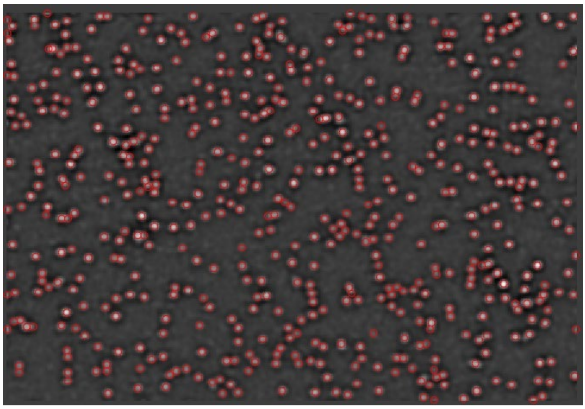


Рис. 2. Границы обнаруженных кластеров на исходном изображении, помеченные кружками

3. Построение математических моделей изображений совокупностей кластеров сигналов флуоресценции для проверки алгоритмов обнаружения. Для отладки и тестирования программного обеспечения обработки сигналов флуоресценции по рассмотренным выше алгоритмам были разработаны программы построения моделей изображений, практически соответствующих реальным изображениям, получаемым с прибора [15].

При построении программы моделирования сигналов флуоресценции приборов массового параллельного секвенирования выполняются следующие операции:

1) Генерация идеального одиночного объекта флуоресценции;

2) Генерация множества изображений одиночных объектов флуоресценции со случайными значениями амплитуды и ширины в соответствии с заданным количеством кластеров;

3) Генерация случайных значений координат x и y объектов флуоресценции на изображении, аналогично получаемому из видеокамеры;

4) Построение изображения на основе множества одиночных объектов флуоресценции со случайными значениями амплитуды и ширины для полученных в п. 3 значений координат x и y;

5) Добавление к изображению, полученного в п. 4, случайного шума с заданными параметрами;

6) Добавление к изображению, полученного в п. 5, нелинейного фона с заданными параметрами.

Идеальный одиночный объект флуоресценции строится на основе трехмерной гауссовой функции в соответствии с формулой (2).

$$g(u,v) = A * exp\left[-\left(\frac{(u-x)^2}{2\ {\sigma_u}^2} + \frac{(v-y)^2}{2\ {\sigma_v}^2}\right)\right], \quad (2)$$

где A – амплитуда сигнала, x и y – координаты максимального значения, σ_u и σ_v – параметры ширины по координатам u и v.

На основе экспериментальных данных параметры σ_u и σ_v можно считать одинаковыми и равными σ. Параметры σ_u и σ_v отвечают за площадь кластера на изображении.

Амплитуда объектов флуоресценции в реальных экспериментах может меняться от 100 до 600 условных единиц. Площадь объектов флуоресценции на реальном изображении может меняться от 6 до 10 пикселей на уровне половины максимального значения. Для создания модели объектов флуоресценции изображений, используемых при отладке алгоритмов обработки информации, рассмотренных выше, необходимо сгенерировать N объектов флуоресценции с указанным диапазоном амплитуд A и шириной σ посредством использования метода Монте-Карло. Создание таких объектов осуществлялось при помощи генератора случайных чисел, имеющих плотность распределения вероятностей подчиняющейся нормальному закону.

Генерация координат x и y объектов флуоресценции осуществлялась с помощью генератора равномерно распределенных случайных чисел для заданного количества объектов N и заданного размера моделируемого изображения, например 320 х 456 пикселей. В результате получался набор из N пар чисел. Каждая пара чисел (x и y) из этого набора соответствовала определенному кластеру с амплитудой A и шириной σ из указанных выше диапазонов. Значения координат x, y, амплитуды A и ширины σ подставлялись в формулу (2). Таким образом, оказались сгенерированными N кластеров и получено изображение, содержащее модели сигналов, построенных на основе моделей идеального одиночного объекта флуоресценции.

Моделирование «слипшихся» объектов производится в той же самой программе, что и моделирование одиночных объектов. При этом для получения эффекта слипания увеличивается общее количество моделируемых объектов. Координаты объектов являются случайными числами и поэтому, чем большее количество объектов моделируется на изображении заданного размера, тем большая вероятность того, что часть объектов будет иметь близкие координаты – такие объекты окажутся наложенными друг на друга и станут «слипшимися». На реальных изображениях, получаемых в ходе эксперимента, часть объектов флуоресценции сливается друг с другом, т.к. технология их «выращивания» также является случайным процессом и локальные центры этих объектов могут быть близки. Количество синтезируемых в секвенаторе фрагментов нуклеиновых кислот, локальные центры кластеров которых близки друг к другу, зависит от концентрации анализируемой пробы. Чем выше концентрация, тем большее количество букв нуклеотидов может быть получено в результате эксперимента. Однако, слишком высокая концентрация анализируемой пробы может привести к ошибочным результатам из-за неправильной оценки параметров кластеров. Оптимальная концентрация анализируемой пробы подбирается опытным путем таким образом, чтобы не допустить, с одной стороны, большого количества «слипшихся» кластеров и, с другой стороны, получить желаемое количество букв нуклеотидов. Для оценки параметров «слипшихся» объектов применяются специальные алгоритмы. Один из таких алгоритмов описан в работе [7].

Для приближения полученного изображения, основанного на моделях идеального одиночного объекта флуоресценции, к изображению, получаемого с прибора массового параллельного секвенирования, необходимо добавить случайный шум и фон. Случайный шум создается с помощью генератора случайных чисел, подчиняющихся нормальному распределению с математическим ожиданием нуль и средним квадратичным отклонением, задаваемым оператором. Величина СКО составляет примерно 5…10 % от величины средней амплитуды сигнала флуоресценции кластера. Для генерации нелинейного фона использовалась модель двумерной гауссовой функции, аналогичная формуле (2), но с шириной, отличающейся от ширины сигналов флуоресценции кластеров примерно в 4000 раз. Центр такой двумерной гауссоиды моделировался в левом нижнем или правом верхнем углу изображения в зависимости от того, в каком месте находился источник света,

создающий флуоресценцию в реальном приборе. Значения ширины, центра и амплитуды функции фона задавались оператором.

На рисунке 3 представлены профили одиночного сигнала объекта в реальном приборе (пунктирная линия) и в модельном изображении (сплошная линия). Как видно из рисунка 3 профили сигналов объекта реального и моделируемого изображений близки. Численная оценка близости моделируемого и реального изображений выполнена в работе [16].

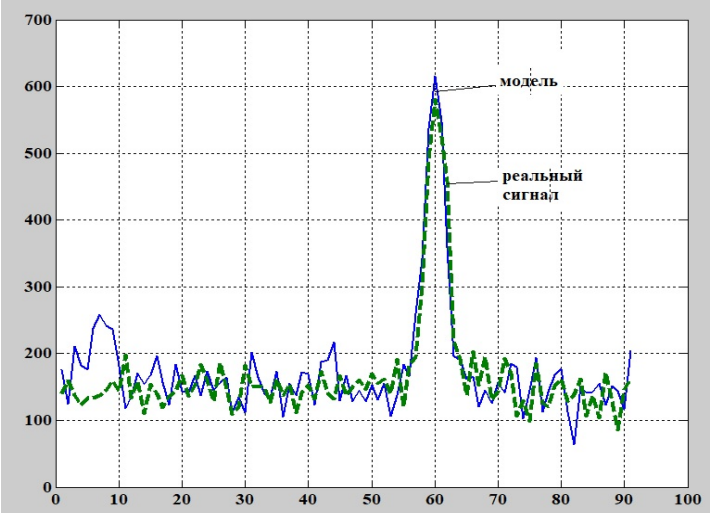


Рис. 3. Профили одиночного сигнала в реальном приборе (пунктирная линия) и модельном изображении (сплошная линия). По горизонтальной оси номер пикселя. По вертикальной оси величина интенсивности в условных единицах

Сравнение изображений, получаемых в ходе генерации модельных кластеров с изображениями кластеров, полученных в реальных экспериментах, показало их высокую идентичность. Подробнее описано в работе А.С. Сараева [16].

Разработанные программы моделирования кластеров молекул ДНК на приборах массового параллельного секвенирования позволили произвести отладку рабочих программ прибора «Нанофор СПС» без использования изображений сигналов флуоресценции, получаемых в ходе дорогостоящих экспериментов. Полученные результаты обеспечили возможность решения задач по обнаружению, оценке параметров кластеров в условиях шума, нелинейности фона и «наложения» кластеров при высокой концентрации объектов.

4. Оценка амплитуды и ширины объекта флуоресценции по известным координатам методом наименьших квадратов на основе трехмерной гауссовой функции. Одиночный объект флуоресценции (кластер) можно описать с помощью формулы (2). Предположим, что мы выделили из исходного изображения фрагмент размером 9x9 пикселей с координатами x и y, например, как показано на рисунке 4(а). 3D изображение яркостей пикселей этого фрагмента, из которых вычтено значение «базовой» линии, показано на рисунке 4(б).

Предположим, что объект флуоресценции симметричный $\sigma_u{=}\sigma_v{=}\sigma$. Тогда значение яркости *i*-го пикселя можно записать следующим образом:

$$Z_i = A * exp\left[-\left(\frac{(u_i-x)^2}{2\,\sigma^2}+\frac{(v_i-y)^2}{2\,\sigma^2}\right)\right], \quad (3)$$

где u_i и v_i горизонтальная и вертикальная координата *i*-го пикселя внутри фрагмента. Произведем логарифмирование и сделаем следующие обозначения:

$$b_0 = ln\,A \text{ и } b_1 = -\frac{1}{2\sigma}. \quad (4)$$

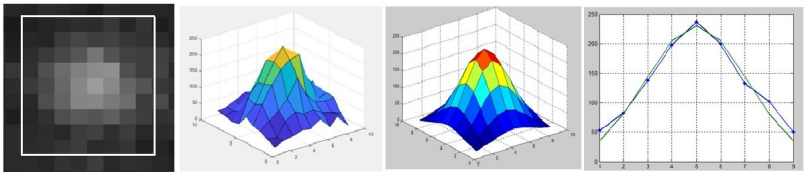
Теперь сумму Q квадратов отклонений логарифмов яркостей пикселей от аппроксимирующей функции, задаваемой по формуле (2) можно записать:

$$Q = \sum_{i=1}^{81}[b_0 + b_1(u_i - x)^2 + b_1(v_i - y)^2 - ln\,Z_i]^2. \quad (5)$$

Для нахождения минимума суммы квадратов отклонений приравняем частные производные $\frac{\partial Q}{b_0}$ и $\frac{\partial Q}{b_1}$ к нулю и после алгебраических преобразований получаем систему из двух линейных уравнений с двумя неизвестными. Решая эту систему, получаем значения b_0 и b_1. Учитывая обозначения (5), получаем формулы для определения амплитуды *A* и ширины σ пика кластера флуоресценции.

Найдя амплитуду и ширину, можно построить по формуле (2) фигуру, представленную на рисунке 4(в) и сравнить ее с фигурой, представленной на рисунке 4(б). Для оценки качества аппроксимации подсчет среднего квадратичного отклонения между исходными данными и данными, полученными по результатам аппроксимации по формуле (2) показал, что это среднее значение не превышает уровня шума в исходных данных. На рисунке 4(г) представлены профили

строк изображений, проходящих через максимум для исходных данных и для данных, полученных в результате аппроксимации. Из информации, представленной на рисунке 4(г) видно, что данные, полученные в результате аппроксимации и исходные данные, близки.



а б в г

Рис. 4. а – фрагмент изображения размером 9x9, показанный белыми линиями; б – 3D изображение яркостей пикселей фрагмента изображения 9x9 пикселей, из которых вычтено значение «базовой» линии; в – 3D изображение яркостей пикселей фрагмента изображения 9x9 пикселей, полученного в результате аппроксимации с использованием формулы (3); г – профиль изображения яркостей пикселей фрагмента исходного изображения (сплошная линия) и изображения, полученного в результате аппроксимации (линия с «*»)

5. Методы машинного обучения для первичного анализа сигналов флуоресценции в технологии массового параллельного секвенирования. В последние десятилетия искусственный интеллект (ИИ), (Artificial Intelligence, AI), машинное обучение (Machine Learning, ML) и глубокое обучение (Deep Learning, DL) стали высокоэффективными подходами, имеющими множество применений в сфере информатики.

В данной работе особое внимание уделяется разработке и внедрению методов машинного обучения для обработки и интерпретации оптических изображений сигналов флуоресценции в реакционных ячейках в приборе «Нанофор СПС». При этом если методам машинного обучения на этапе вторичного анализа биоинформационных данных секвенирования уделено уже достаточно много внимания в современных исследованиях [17, 18], то применение машинного обучения для первичного анализа только начало привлекать внимание исследователей [8]. Оптические изображения реакционных ячеек секвенаторов генерируют огромные объемы данных, но при этом демонстрируют высокую вариабельность из-за различий в подготовке образцов, методах визуализации, оборудовании и используемых параметрах сбора данных микроскопии. Присущая таким изображениям изменчивость создает проблемы для анализа классическими методами обработки изображений. Следовательно,

методы машинного обучения (ML), а в перспективе и методы глубокого обучения (DL) предлагают привлекательное решение для повышения скорости, точности, адаптируемости, воспроизводимости и эффективности анализа изображений.

5.1. Методы ML. Подходы ML/DL, используемые для анализа изображений в секвенаторе «Нанофор СПС», делятся на два типа (рисунок 5): обучение с учителем (Supervised learning) и обучение без учителя (Unsupervised learning). В обучении с учителем используется размеченный набор данных. Размеченный набор данных – это совокупность данных, разделенных на подмножества: обучающий набор (training set), тестовый набор (test set) и иногда проверочный набор данных (validation set). Методы обучения с учителем используются для прогнозного моделирования и подразделяются на задачи классификации и регрессии. В задаче обработки данных секвенатора типичной задачей классификации является задача base-calling. Base-calling – это процесс определения нуклеотидного основания, который генерирует соответствующее значение интенсивности в каналах флуоресценции для различных длин волн на кадрах изображения реакционной ячейки для различных циклов секвенирования методом синтеза [19]. Разнообразные алгоритмы такой классификации ML включают модель персептрона, логистическую регрессию, деревья решений, метод опорных векторов (SVM), случайный лес и k-ближайшие соседи (KNN).

Проблема base-calling, сформулированная в виде обобщенной модели [20], сводится к последовательности задачи регрессии для оценки параметров этих моделей. Соответствующие алгоритмы стремятся подогнать точки данных посредством регрессионного анализа, например, логистической регрессии, полиномиальной регрессии и линейной регрессии. Тип используемой регрессии будет зависеть от количества независимых и зависимых переменных и формы распределения точек данных. Примеры такой реализации решения проблемы base-calling содержатся в работах [21].

Прогнозируемые выходные данные контролируемых моделей ML могут быть «переобучены» (overfitted) или недообучены в результате некорректного обучения, которое приводит к искаженным или смещенным результатам с высокой погрешностью. Переобучение является результатом слишком точного обучения модели к точкам данных тестового набора, при этом набор обучающих данных слишком мал для обучения или слишком зашумлен. Задача переобучения (overfitting) очень часто возникает при применении моделей машинного обучения в задачах регрессии. Здесь она решается с помощью

регуляризации соответствующих регрессионных моделей [22]. Также эффективным средством борьбы с такого рода переобучением являются методы снижения размерности модели. Различные методы снижения размерности для задачи base-calling рассмотрены в работе [23].

Рис. 5. Разные типы методов в машинном обучении, используемые при анализе изображений в биоинформатике. CNN – сверточная нейронная сеть; DBSCAN – основанная на плотности кластеризация для приложений с шумами; HDBSCAN – иерархическая версией DBSCAN; BIRCH – сбалансированное итеративное сокращение и кластеризация с помощью иерархий; GMM – модель гауссовой смеси; KNN – метод k-ближайших соседей; OPTICS – упорядочение точек для обнаружения кластерной структуры; PCA – анализ главных компонент; SVD – разложение по сингулярным значениям; SVM – метод опорных векторов; SVR – регрессия опорных векторов; t-SNE – стохастическое вложение соседей с t-распределением; UMAP – Uniform Manifold Approximation and Projection

Наконец, типичным классом задач обучения без учителя являются задачи кластеризации. При этом, не требуется наличие помеченного набора данных (labelled dataset). Набор данных не сгруппирован в обучающий и тестовый набор – весь набор данных вводится в модель машинного обучения для анализа данных. Кластерный анализ – это метод, который группирует схожие точки данных в кластеры в зависимости от их относительного сходства. В задачах обработки изображений сигналов флуоресценции задача кластеризации связана с обнаружением клональных кластеров, получаемых из фрагментов геномной библиотеки посредством мостиковой амплификации. Такая задача в области массового параллельного секвенирования сильно осложнена огромным количеством таких кластеров (сотни тысяч), небольшим количеством

пикселей изображений, приходящихся в среднем на каждый кластер, неопределенностью паттерна каждого отдельного кластера, плотной упаковкой кластеров в рамках одного изображения. Естественным образом, при анализе таких сложных физико-химических процессов необходимой задачей является определение зон чрезмерной плотности кластеров или аномальных зон оптической яркости для их последующей фильтрации в области обнаружения кластеров.

Таким образом, практически все типовые задачи обработки изображений (классификация, регрессия, кластеризация, снижение размерности, выявление аномалий) востребованы на стадии первичного анализа данных секвенирования.

5.2. Процесс анализа изображений. Типичный процесс анализа изображений состоит из нескольких этапов: предварительная обработка, детектирование объектов, сегментация объектов, извлечение признаков для интеллектуального анализа данных и других [24]. На рисунке 6 представлена структурная схема процесса анализа изображений в секвенаторах.

Рис.6. Общий рабочий процесс анализа изображений в секвенаторах

5.2.1. Предварительная обработка (Image preprocessing). Процесс обычно начинается с предварительной обработки необработанных изображений после регистрации. Предварительная обработка включает в себя такие этапы, как контроль качества изображения и различные типы манипуляций с изображением (например, изменение размера и улучшение отношения сигнал/шум (SNR). Этот шаг особенно полезен, например, при работе в условиях

низкой освещенности, что приводит к низкой интенсивности сигнала во время визуализации образца в реальном времени. Предварительная обработка может также включать удаление выбросов и шумоподавление в изображениях.

5.2.2. Детектирование объектов, сегментация и выделение признаков (object detection, object segmentation, feature extraction). Обнаружение объектов включает в себя локализацию и классификацию объектов на изображении. Его цель – идентифицировать конкретные объекты, представляющие интерес, и определить их ограничивающие рамки, что имеет решающее значение для таких задач, как отслеживание поведения объектов. Задача обработки изображений, полученных секвенатором, усложняется отсутствием образцов кластеров как таковых и небольшим количеством пикселей на каждый объект. Сегментация – это процесс разделения изображения или видео на значимые области для идентификации и дифференциации объектов. Она служит таким целям, как понимание границ объектов, извлечение детальной информации и обеспечение дальнейшего анализа. Одним из этапов сегментации является бинаризация изображений. Этот метод используется для распознавания объектов, т.к. позволяет отличить интересующий объект от фона, на котором он находится. Следующим этапом служит сегментация экземпляров (Instance Segmentation) [25]. Сегментация экземпляров связывает наборы пикселей с каждым отдельным экземпляром объекта и используется, когда необходимо различать уникальные экземпляры одной и той же категории объектов на изображении. Наконец, в процессе выделения признаков сегментированные объекты наделяются рядом признаков, которые служат в дальнейшем для последующего анализа. В частности, выделенные кластеры на изображениях флуоресцентных сигналов наделяются рядом характеристик, позволяющих, во-первых, отфильтровать низко информативные кластеры по признакам chastity и purity, во-вторых, эффективно провести классификацию base-calling.

5.2.3. Отслеживание объектов и их классификация. Алгоритмы машинного обучения могут сыграть заметную роль в задаче выравнивания изображений сигналов флуоресценции, полученных на различных циклах работы секвенатора. Наконец, финальной частью обработки изображения для секвенаторов является задача классификации по типу нуклеотида на каждом цикле секвенатора, что и является сутью base-calling.

5.3. Алгоритмы кластеризации в обработки изображений секвенатора. Чтобы не путать определения кластерного анализа, как одного из методов классификации в алгоритмах машинного обучения

и кластера в приборах для секвенирования нуклеиновых кислот дадим поясняющие определения.

Кластерный анализ, как алгоритм классификации – категория методов обучения без учителя, которая позволяет нам обнаруживать скрытые структуры в данных, где мы заранее не знаем правильного ответа. Цель кластеризации – найти естественную группировку данных, чтобы элементы в одном кластере были более похожи друг на друга, нежели на элементы из разных кластеров.

Кластер в приборах для секвенирования – клональная группа матричной ДНК, связанная с поверхностью реакционной ячейки. Каждый кластер засевается одной цепью матричной ДНК и клонально амплифицируется посредством мостиковой амплификации до тех пор, пока кластер не наберет примерно 1000 копий исследуемой ДНК. Каждому кластеру реакционной ячейки в идеале соответствует одно прочтение (рид) исследуемой геномной последовательности.

Рассмотрим классификацию методов кластерного анализа, приведенного в новейшем обзоре [26], которые представлены на рисунке 7.

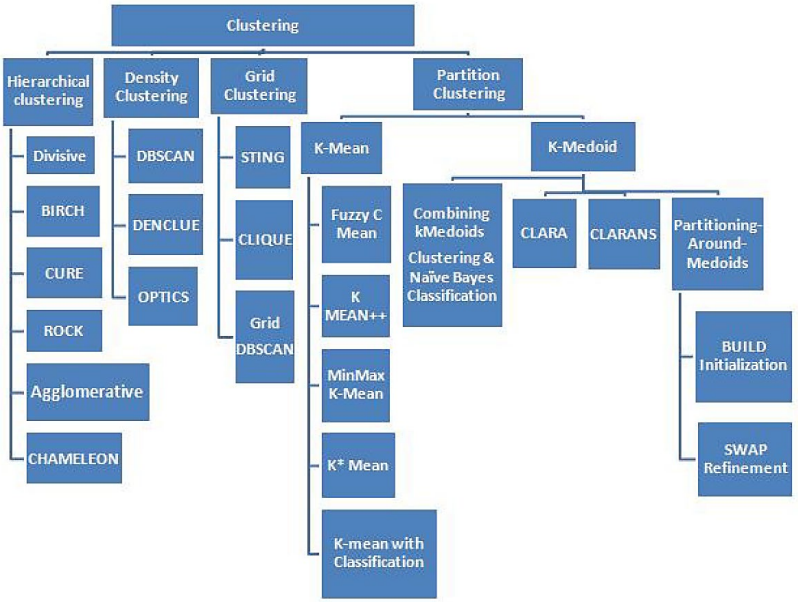


Рис.7. Классификация методов кластерного анализа [44]

5.3.1. Иерархическая кластеризация (Hierarchical Clustering) – популярная стратегия обучения без учителя для группировки схожих элементов данных. Она создает кластер иерархической структуры путем многократного слияния и разделения кластеров на основе сходства и несходства. Основная концепция, лежащая в основе иерархической кластеризации, заключается в построении дендрограммы, которая представляет собой древовидную структуру, изображающую связи между точками данных и кластерами. Дендрограмма начинается с каждой точки данных как отдельного кластера и в конечном итоге объединяет связанные группы в соответствии со значением сходства и расстояния. Алгоритм оценивает близость кластеров и выбирает кластеры для объединения на каждом этапе. В зависимости от данных и проблемной области метрики сходства и расстояния, используемые в иерархической кластеризации, могут различаться. Обычно используемые меры расстояния включают евклидово расстояние, манхэттенское расстояние и корреляционное расстояние. Из методов, доступных в библиотеке Scikit-learn для анализа изображений в секвенаторе, были опробованы агломеративная иерархическая кластеризация [27], а также сбалансированное итеративное сокращение и кластеризация с иерархиями (BIRCH) [28].

5.3.2. Кластеризация на основе плотности (Density-Based Clustering) – это тип алгоритма кластеризации, который организует точки данных в пространстве объектов в зависимости от их плотности. Кластеризация на основе плотности – это метод кластеризации, который организует точки данных в пространстве данных на основе их плотности. Методы этого класса стремятся обнаруживать кластеры любой формы и размера путем распознавания мест со значительной плотностью данных, при этом не делается предположений относительно количества кластеров в наборе данных или предустановленных форм кластеров. Из методов, предоставляемых библиотекой Scikit-learn были испытаны методы DBSCAN и HDBSCAN.

Основными особенностями DBSCAN являются следующие [29]. Метод основан на поиске областей с высокой плотностью точек, разделенных областями с низкой плотностью. Такие плотные скопления и формируют кластеры. Метод может выявлять кластеры произвольной формы, в том числе неконвексные, не требует указания заранее известного числа кластеров, устойчив к выбросам и шумовым точкам, выделяет их в отдельные «кластеры». DBSCAN имеет два основных параметра: радиус окрестности эпсилон и минимальное

число точек в окрестности MinPts. Именно они определяют порог плотности для кластеризации. Таким образом, DBSCAN позволяет находить скопления объектов произвольной формы в данных без необходимости заранее задавать число кластеров. Это делает его гибким и удобным методом кластеризации во многих практических задачах.

Еще более применимым к обработке изображений в секвенаторе оказался метод HDBSCAN. Как и DBSCAN, метод HDBSCAN разработан для обнаружения кластеров различной формы в пространстве признаков, в отличие от методов K-средних или иерархической кластеризации. Вот основные различия между DBSCAN и HDBSCAN:

1. Параметры: DBSCAN требует два параметра: эпсилон (eps, радиус окружности для соседей) и минимальное количество точек (общее количество точек, которые должны находиться в пределах радиуса eps, чтобы образовать кластер). HDBSCAN требует только параметра minPts (минимальное количество точек), так как он динамически оценивает эпсилон на основе плотности данных.
2. Точки шума: DBSCAN отмечает все точки, лежащие вне minPts эпсилон-сферы окружности, как шум. HDBSCAN, тем не менее, имеет более сложный способ обработки шума и может назначить точки шума к кластерам на основе стойкости.
3. Гибкость: DBSCAN на самом деле сложно использовать на практике, так как он чувствителен к выбору параметра эпсилон. В свою очередь HDBSCAN более гибок, успевает работать со сложными структурами данных, способен обрабатывать и неравномерные плотности данных.
4. Количество кластеров: в отличие от алгоритма K-средних, число кластеров в DBSCAN и HDBSCAN не указывается заранее. HDBSCAN, как модифицированная версия DBSCAN, позволяет получить даже вложенную иерархию кластеров.
5. Иерархическая кластеризация: HDBSCAN выполняет иерархическую кластеризацию, создавая иерархию кластеров с разными уровнями гранулярности или уровня детализации, в то время как DBSCAN не создает явной структуры иерархии кластера.

5.3.3. Методы кластеризации, основанные на сетке (Grid-Based Clustering). Кластеризация на основе сетки – это подход к кластеризации, который использует поле данных, разделенное на ячейки сетки. Он может предоставить эффективные и масштабируемые методы кластеризации, которые чрезвычайно полезны для больших наборов данных [30]. В отличие

от традиционных алгоритмов кластеризации, таких как K-Means, которые работают непосредственно с точками данных в их исходном пространстве признаков, Grid-Based Clustering использует другой подход, заключающийся в нескольких этапах реализации:

1. Дискретизация: разделяет пространство данных на сетку ячеек.
2. Расчет плотности: рассчитывается плотность точек данных внутри каждой ячейки.
3. Формирование кластеров. Непрерывные ячейки с плотностью выше определенного порога группируются с образованием кластеров.

Данная работа ограничивается методами кластеризации, реализованными в рамках библиотеки Scikit-learn, которая на настоящее время не имеет специального модуля Grid-Based Clustering. В дальнейших планах авторов реализация такого метода на основе комбинации классов KBinsDiscretizer для создания разбиения пространства на ячейки сети, KernelDensity для расчета плотности точек и scipy.ndimage.label для идентификации кластеров. Огромное количество потенциально присутствующих кластеров на анализируемых изображениях при секвенировании служит веским основанием уделить этим методам особое внимание в дальнейших исследованиях.

5.3.4. Разделительные методы кластеризации (Partitioning-Based Clustering). Это тип алгоритмов кластеризации, который делит наборы данных на отдельные группы путем оптимизации целевой функции. Например, разбивает данные на заданное число кластеров, минимизируя расстояние внутри кластера и максимизируя расстояние между кластерами. При этом в ходе кластеризации алгоритм непрерывно распределяет точки данных по кластерам и сохраняет точки центроида кластера до тех пор, пока не произойдет сходимость по определенному критерию.

В нашей работе для обнаружения объектов флуоресценции был использован метод К-средних (K-means) из библиотеки Scikit-learn. Метод К-средних – один из основных алгоритмов кластеризации, целью которого является разделение набора наблюдений на K кластеров, в которых каждое наблюдение принадлежит кластеру с ближайшим средним значением, служащим прототипом кластера.

6. Итоговый алгоритм обнаружения сигналов флуоресценции в секвенаторах. Изображения, детектируемые

на протяжении циклов в процессе массового параллельного секвенирования, характеризуются рядом особенностей [31]:

– Огромное количество объектов (кластеров), требующих обнаружения (до нескольких сот тысяч объектов);

– Плотность объектов высока и неравномерна по кадрам, что требует процедур разделения кластеров и/или фильтрации плохо разрешимых участков изображения;

– Изображения получаются в четырех разных каналах флуоресценции и требуют аккуратного совмещения кадров по циклам и каналам (alignment);

– Уровень шумов на изображениях высок и требует эффективного отделения сигналов реальных кластеров от шумовой составляющей;

– Количество пикселей на каждый кластер достаточно мало и форма отдельных кластеров нерегулярна;

– Для оптимального функционирования дальнейших этапов требуется корректное определение центров кластеров;

– Производительность алгоритмов кластеризации крайне важна, поскольку лимитирована временем обработки каждого цикла секвенирования и последовательностью обработки различных участков изображения реакционной ячейки.

Исходя из вышеприведенных требований был предложен многостадийный этап кластеризации:

1) На первом этапе применен метод K-means++ с целью отделения полезного сигнала от фона. При этом число задаваемых кластеров варьировалось в диапазоне 3-5. По сути, на этом этапе мы проводим гибкую бинаризацию изображений для последующей обработки. Кластеризация происходит по значениям интенсивностей кластеров. Метод K-means++ обеспечивает высокую производительность при проведении этого этапа обработки. Главная цель при этом – снизить объем данных для последующей пространственной кластеризации.

2) На втором этапе из всего арсенала апробированных методов выбран BIRCH. BIRCH – это алгоритм иерархической кластеризации, разработанный непосредственно для сканирования больших наборов данных. Основные преимущества BIRCH заключаются в эффективности процесса кластеризации исходя из параметров близости пикселей на изображении и возможности выбирать пороговое значение и количество фрагментов в дереве BIRCH для оптимизации получаемой модели. Однако стоит учитывать, что BIRCH лучше всего подходит для равномерно распределенных

данных и может работать неудовлетворительно при кластеризации сложных структур, таких как кластеры с переменной плотностью или иррегулярной формы. В дальнейшем предлагается апробировать на этом этапе методы Density-Based Clustering, такие как DBSCAN и HDBSCAN с целью оптимального обнаружения кластеров нерегулярной формы.

3) На третьем этапе опционально может быть использована операция сегментации изображения, основанная на морфологических операциях библиотеки skimage [32]. Главная задача этого дополнительного этапа – получить большую информацию о кластерах в виде параметров их плотности и площади с целью отфильтровать плохо разрешимые кластеры и пренебречь случайными точками выбросов большой интенсивности на изображении.

Для пояснения работы предложенного комбинированного метода кластеризации приведены иллюстрации к каждому этапу. Изображение представляет собой моделированное распределение кластеров отдельных ридов в одном из каналов флуоресценции в одном из циклов секвенирования с плотностью, характерной для протокола работы секвенатора «Нанофор СПС». На рисунке 8 представлена стадия бинаризации изображения, т.е. его разделение на фоновую и сигнальную составляющую. В данном случае мы кластеризуем изображения по интенсивностям на 4 группы. Два класса самых интенсивных пикселей (зеленый цвет – самые интенсивные пиксели, желтый цвет – следующий класс пикселей по интенсивности) идут на следующий этап для пространственной кластеризации. На рисунке 9 показан результат пространственной кластеризации. Красными точками показаны истинные положения кластеров, взятые из программы моделирования. Голубая область вокруг них – пиксели, отнесенные кластеризацией BIRCH к данному кластеру. На следующем этапе (рисунок 10) изображение сегментируется (красные прямоугольники). Характеристики сегментов (плотность, площадь) позволяют применять фильтры, отсеивающие зоны наложения кластеров друг на друга. По результатам проведенного моделирования такой алгоритм кластеризации позволяет обнаруживать 85-90% процентов кластеров, при этом ошибка определения центра объекта для 80% процентов кластеров составляет меньше одного пикселя.

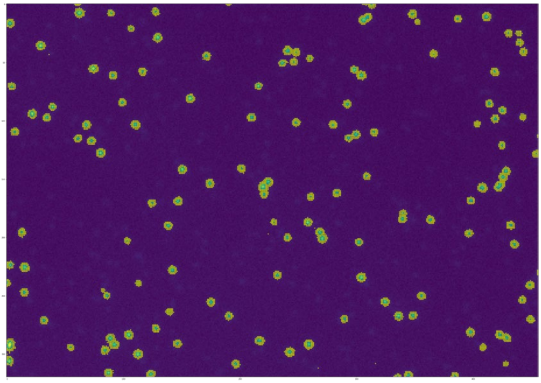


Рис. 8. Этап бинаризации

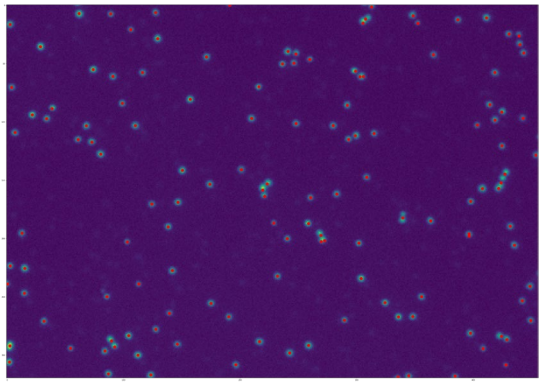


Рис. 9. Этап кластеризации методом Birch

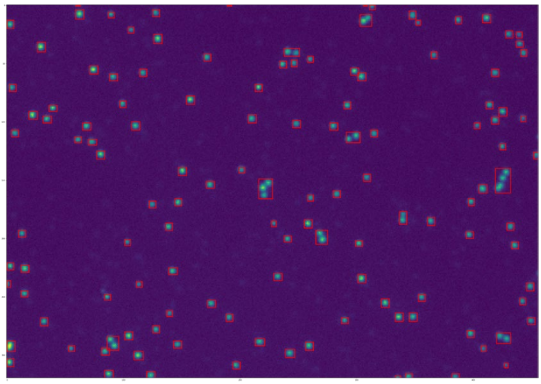


Рис. 10. Этап сегментации

Рассмотренный в данном разделе итоговый алгоритм обнаружения сигналов флуоресценции имеет два преимущества по сравнению с классическим алгоритмом обнаружения на основе свертки:

1. Для обнаружения сигналов не требуется предварительное знание функции формы обнаруживаемого кластера;
2. Этот алгоритм позволяет выделять «слипшиеся» кластеры, параметры которых, как правило, оцениваются не в режиме онлайн, а после окончания эксперимента по алгоритмам, описанным в работе [7].

7. Сравнение программного обеспечения первичного анализа локальных объектов флуоресценции в секвенаторе «НАНОФОР СПС» с зарубежным аналогом. Рассмотренные в настоящей работе алгоритмы первичного анализа локальных объектов флуоресценции в секвенаторе «НАНОФОР СПС» легли в основу разработанного в Институте аналитического приборостроения РАН программного обеспечения с предварительным названием PrImA. Данное программное обеспечение является первой отечественной разработкой для приборов, основанных на методе массового параллельного секвенирования.

Для сравнения этого программного обеспечения с его зарубежным аналогом RTA рассмотрим графики, полученные с помощью программного обеспечения UGENE. UGENE – это отечественное свободно распространяемое программное обеспечение для биоинформатики с открытым исходным кодом под системы Windows, macOS и Linux. UGENE собирает полную нуклеотидную последовательность из фрагментов, полученных по результатам первичного анализа локальных объектов флуоресценции в секвенаторе.

На рисунке 11 и 12 представлены распределения показателей качества отдельных букв нуклеотидов референсного генома бактериофага Phix174, для которых первичный анализ флуоресценции был выполнен с помощью программы RTA и программы PrImA. Референсный геном бактериофага Phix174 содержит 5386 нуклеотидов и в результате проведенного анализа на секвенаторе «Нанофор СПС» этот геном в среднем был покрыт примерно 100 раз.

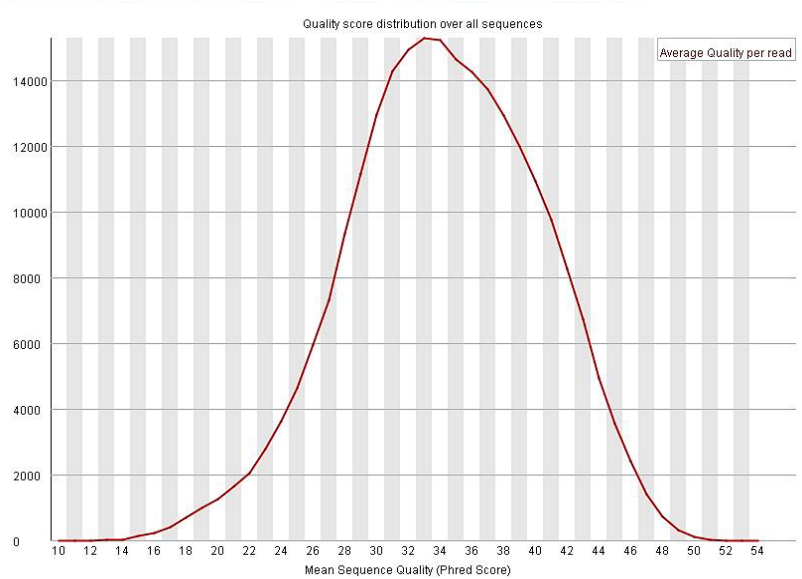


Рис. 11. Показатели качества последовательностей нуклеотидов, полученных с помощью программы RTA

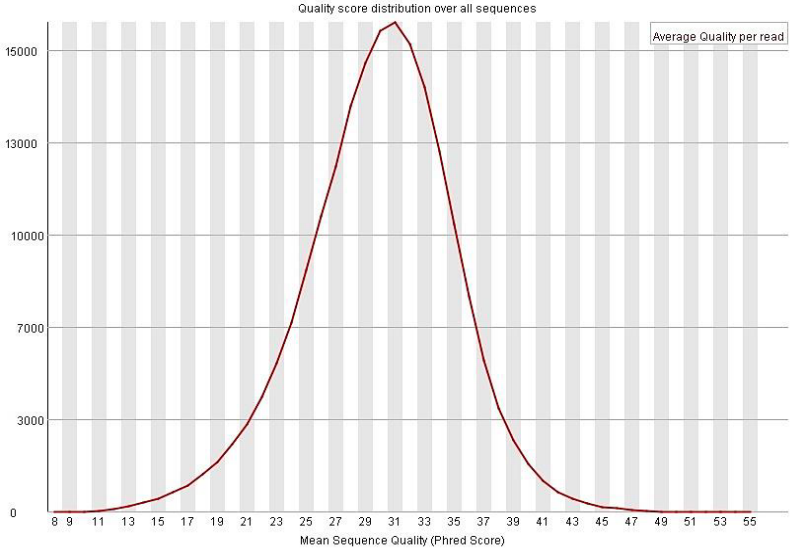


Рис. 12. Показатели качества последовательностей нуклеотидов, полученных с помощью программы PrImA

По горизонтальной оси на этих графиках отложены показатели качества по Phred Score, принятые в биоинформатике. Оценки качества Phred логарифмически связаны с вероятностью ошибок построения последовательности букв нуклеотидов и определяются как:

$$Q = -log_{10} P.$$

Это соотношение можно записать как:

$$P = 10^{\frac{-Q}{10}}.$$

Например, Phred присваивает букве оценку качества, равную 30. Вероятность того, что эта буква в последовательности была названа неправильно, равна 1 к 1000, или вероятность правильности буквы равна 99.9%.

По вертикальной оси откладывается количество букв, оценка которых произведена с указанной вероятностью.

Сравнение графиков на рисунках 11 и 12 показывает, что программы RTA и PrImA генерируют примерно одинаковые результаты.

8. Заключение. Описанные алгоритмы являются результатом исследований, проведенных в процессе разработки секвенатора «Нанофор СПС». Созданное программное обеспечение на основе данных методик в значительной степени предопределило успешный запуск серийной партии отечественных приборов. Для проверки работоспособности алгоритмов были использованы изображения, полученные непосредственно в ходе экспериментов на приборе, а также математические модели изображений кластеров ДНК.

Использование технологий машинного обучения в программном обеспечении секвенатора ДНК «Нанофор СПС» позволило расширить диапазон методов первичного анализа сигналов флуоресценции и показало возможность их применения в обработке без априорного знания функции искажения, а также для автоматического обнаружения слипшихся кластеров.

Программное обеспечение, разработанное на основе рассмотренных алгоритмов, позволяет улучшить качество проводимых экспериментов в сфере геномных исследований и расширить области их применения.

Литература

1. Курочкин В.Е., Алексеев Я.И., Петров Д.Г., Евстрапов А.А. Отечественные приборы для молекулярно-генетического анализа: разработки ИАП РАН и ООО «Синтол» // Известия Российской Военно-медицинской академии. 2021. Т. 40. № 3. С. 69–74. DOI: 10.17816/rmmar76918.
2. Ansorge W.J. Next-generation DNA sequencing techniques // Nature Biotechnology. 2009. vol. 25. no. 4. pp. 195–203.
3. Bentley R.D. Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., et al. Accurate whole human genome sequencing using reversible terminator chemistry // Nature. 2008. vol. 456. no. 7216. pp. 53–59.
4. Whiteford N. The Solexa pipeline. 2012. URL: http//41j.com/blog/wp-content/uploads/2012/04/pipeline.pdf (дата обращения: 20.02.2024).
5. Leshkowitz D. Introduction to Deep-Sequencing Data Analysis Illumina Primary Analysis Pipeline & Quality Control. 2017. URL: http://dors.weizmann.ac.il/course/course2017/Dena_IlluminaPrimaryAnalysisPipeline-course2017.pdf (дата обращения: 20.02.2024).
6. Манойлов В.В., Бородинов А.Г., Сараев А.С., Петров А.И., Заруцкий И.В., Курочкин В.Е. Алгоритмы обработки изображений в секвенаторе ДНК НАНОФОР СПС // Журнал технической физики. 2022. Т. 92. № 7. С. 985–992. DOI: 10.21883/JTF.2022.07.52655.318-21.
7. Манойлов В.В., Бородинов А.Г., Заруцкий И.В., Петров А.И., Курочкин В.Е. Алгоритмы обработки сигналов флуоресценции массового параллельного секвенирования нуклеиновых кислот // Труды СПИИРАН. 2019. Т. 18. № 4. С. 1010–1036. DOI: 10.15622/sp.2019.18.4.1010-1036.
8. Бородинов А.Г., Манойлов В.В., Заруцкий И.В., Петров А.И., Курочкин В.Е., Сараев А.С. Машинное обучение в задачах base-calling для методов секвенирования нового поколения // Информатика и автоматизация. 2022. Т. 21. № 3. С. 572–603. DOI: 10.15622/ia.21.3.5.
9. Журавель И.М. Краткий курс теории обработки изображений. URL: http://matlab.exponenta.ru/imageprocess/book2/49.php (дата обращения: 26.10.2023).
10. Вудс Р., Гонсалес Р. Цифровая обработка изображений / 3-е изд. // М.: Техносфера. 2012. 1104 с.
11. 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.
12. Sizikov V.S., Stepanov A.V., Mezhenin A.V., Burlov D.I., Eksemplyarov R.A. 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. 203–110.
13. Бардин Б.В., Чубинский-Надеждин И.В. Обнаружение локальных объектов на цифровых микроскопических изображениях // Научное приборостроение. 2009. Т. 19. № 4. С. 96–102.
14. Otsu N. A Threshold Selection Method from Gray-Level Histograms // IEEE Transactions on Systems. IEEE Transactions on Systems, Man and Cybernetics. 1979. vol. 9. pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.
15. Сараев А.С., Петров А.И., Манойлов В.В. Моделирование генерации кластеров молекул ДНК в приборах массового параллельного секвенирования // Тезисы докладов Четвертой международной конференции со школой молодых ученых «Физика – наукам о жизни» / СПб: ФТИ им. А.Ф. Иоффе. 2021. С. 153.

16. Сараев А.С. Научно-квалификационная работа «Разработка алгоритма распознавания кластеров нуклеиновых кислот в микрофлюидной ячейке секвенатора «Нанофор СПС». СПб: ИАП РАН. 2023. С. 16–22.
17. Schmidt B., Hildebrandt A. Deep learning in next-generation sequencing // Drug discovery today. 2021. vol. 26. no. 1. pp. 173–180.
18. Ozgur S., Orman M. Application of deep learning technique in next generation sequence experiments // Journal of Big Data. 2023. vol. 10. no. 1. DOI: 10.1186/s40537-023-00838-w.
19. Tegfalk E. Application of machine learning techniques to perform base-calling in next-generation DNA sequencing. 2020. 45 p.
20. Cacho A., Smirnova E., Huzurbazar S., Cui X. A comparison of base-calling algorithms for illumina sequencing technology // Briefings in bioinformatics. 2016. vol. 17. no. 5. pp. 786–795.
21. Kircher M., Stenzel U., Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies // Genome biology. 2009. vol. 10(8). DOI: 10.1186/gb-2009-10-8-r83.
22. Ghojogh B., Crowley M. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial // arXiv preprint arXiv:1905.12787. 2019.
23. Бородинов А.Г., Ямановская А.Ю., Манойлов В.В., Петров А.И. Оптимальный выбор признаков для уменьшения размерности моделей машинного обучения в задаче base-calling // Тезисы докладов Второй ежегодной всероссийской молодежной конференции по методам и приборам для анализа биологических объектов «АналитБиоПрибор-2023» (Санкт-Петербург, 23–24 ноября 2023 г.). Санкт-Петербург: Издательско-полиграфическая ассоциация высших учебных заведений, 2023. С. 135–138.
24. Whiteford N., Skelly T., Curtis C., Ritchie M.E., Lohr, A., Zaranek A.W., Abnizova I., Brown C. Swift: primary data analysis for the Illumina Solexa sequencing platform // Bioinformatics. 2009. vol. 25. no. 17. pp. 2194–2199.
25. Hafiz A.M., Bhat G.M. A survey on instance segmentation: state of the art // International journal of multimedia information retrieval. 2020. vol. 9. no. 3. pp. 171–189.
26. Chaudhry M., Shafi I., Mahnoor M., Vargas D.L.R., Thompson E.B., Ashraf I.A. Systematic literature review on identifying patterns using unsupervised clustering algorithms: a Data mining perspective // Symmetry. 2023. vol. 15. no. 1679. DOI: 10.3390/sym15091679.
27. Khandare A., Pawar R. Data clustering algorithms: experimentation and comparison // Intelligent Computing and Networking: Proceedings of IC-ICN 2021. 2022. pp. 86–99.
28. Sarang P. BIRCH: Divide and Conquer // Thinking Data Science: A Data Science Practitioner's Guide. Cham: Springer International Publishing. 2023. pp. 229–236.
29. Ester M., Kriegel H.P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // kdd. 1996. vol. 96. no. 34. pp. 226–231.
30. Zhao Y., Cao J., Zhang C., Zhang S. Enhancing grid-density based clustering for high dimensional data // Journal of Systems and Software. 2011. vol. 84. no. 9. pp. 1524–1539.
31. Wolowski V.R. High-quality, high-throughput measurement of protein-DNA binding using HiTS-FLIP. Dissertation, LMU Munchen: Fakultat fur Chemie und Pharmazie 2016. DOI: 10.5282/edoc.19445.
32. Chityala R., Pudipeddi S. Image processing and acquisition using Python // Chapman and Hall/CRC. 2020. 452 p.

33. Kameshwaran K., Malarvizhi K. Survey on clustering techniques in data mining // International Journal of Computer Science and Information Technologies. 2014. vol. 5. no. 2. pp. 2272–2276.

Манойлов Владимир Владимирович — д-р техн. наук, доцент, заведующий лабораторией, лаборатория автоматизации измерений и цифровой обработки сигналов, ИАП РАН. Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 101. manoilov-vv@mail.ru; ул. Ивана Черных, 31-33А, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0720.

Бородинов Андрей Геннадьевич — д-р физ.-мат. наук, старший научный сотрудник, лаборатория методов и приборов иммунного и генетического анализа, ИАП РАН. Область научных интересов: математическая статистика, проблемы анализа, обработки и представления данных, искусственный интеллект. Число научных публикаций — 15. borodinov@gmail.com; ул. Ивана Черных, 31-33А, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0719.

Заруцкий Игорь Вячеславович — канд. техн. наук, старший научный сотрудник, лаборатория автоматизации измерений и цифровой обработки сигналов, ИАП РАН. Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 51. igorzv@yandex.ru; ул. Ивана Черных, 31-33А, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0719.

Петров Александр Иванович — заведующий сектором электроники и программного обеспечения, лаборатория методов и приборов иммунного и генетического анализа, ИАП РАН. Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 21. fataip@mail.ru; ул. Ивана Черных, 31-33А, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0719.

Сараев Алексей Сергеевич — инженер 1 категории, лаборатория методов и приборов иммунного и генетического анализа, ИАП РАН. Область научных интересов: представление и обработка сигналов и изображений в аналитических устройствах. Число научных публикаций — 4. alex.niispb@yandex.ru; ул. Ивана Черных, 31-33А, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0719.

Курочкин Владимир Ефимович — д-р техн. наук, профессор, руководитель научного направления, заведующий лабораторией, лаборатория методов и приборов иммунного и генетического анализа, 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; ул. Ивана Черных, 31-33А, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0719.

Поддержка исследований. Работа выполнена в соответствии с Государственным заданием Министерства науки и высшего образования РФ № 075-01157-23-00 от 29.12.2022 г.

DOI 10.15622/ia.23.4.3

V. MANOILOV, A. BORODINOV, I. ZARUTSKY, A. PETROV, A. SARAEV, V. KUROCHKIN

ALGORITHMS FOR THE PRIMARY ANALYSIS OF LOCAL FLUORESCENCE OBJECTS IN THE DNA SEQUENCER «NANOFOR SPS»

Manoilov V., Borodinov A., Zarutsky I., Petrov A., Saraev A., Kurochkin V. **Algorithms for the Primary Analysis of Local Fluorescence Objects in the DNA Sequencer «Nanofor SPS».**

Abstract. The DNA sequencer "Nanofor SPS", developed at the Institute of Analytical Instrumentation of the Russian Academy of Sciences, implements a method for massively parallel sequencing to decrypt the sequence of nucleic acids. This method allows for the determination of the nucleotide sequence in DNA or RNA, containing from several hundred to hundreds of millions of bases. Thus, there is the opportunity to obtain detailed information about the genome of various biological entities, including humans, animals, and plants. A crucial part of this device is the software, without which it is impossible to solve genome decoding tasks. The output data of optical detection in the sequencer are a set of images over four channels, corresponding to nucleotide types: A, C, G, T. Through specialized software, the position of molecular clusters and their intensity characteristics, along with parameters of the surrounding background, are determined. Algorithms and programs for processing fluorescence signals, discussed in the paper, were developed during the creation of the device software. Also, to debug and test the working programs, models of image construction similar to real data obtained in the course of sequencer operation were created. These models made it possible to obtain a significant amount of information without running expensive experiments. Significant progress has been made in the field of machine learning in recent years, including in the field of bioinformatics, leading to the implementation of the most common models and their potential for practical tasks. However, while these methods have amply proven their worth in secondary bioinformatics data analysis, their potential for the primary analysis remains untapped. This paper focuses on the development and implementation of machine learning methods for primary analysis of optical images of fluorescence signals in reaction cells. The methods of clustering and their testing on models and images obtained from the device are described. The aim of this paper is to demonstrate the capabilities of algorithms for primary analysis of fluorescence signals that arise during sequencing in the «Nanofor SPS» device. The paper describes the main tasks of fluorescence signal analysis and compares traditional methods of solving them and solutions using machine learning technologies.

Keywords: sequencing, nucleic acid, processing DNA and RNA fluorescence signals methods, image analysis, machine learning.

References

1. Kurochkin V.E., Alekseev Ya.I., Petrov D.G., Evstrapov A.A. [Domestic devices for molecular genetic analysis: developments of the IAP RAS and Syntol LLC]. Izvestiya Rossijskoj Voenno-medicinskoj akademii – Proceedings of the Russian Military Medical Academy. 2021. vol. 40. no. 3. pp. 69–74. DOI: 10.17816/rmmar76918. (In Russ.).
2. Ansorge W.J. Next-generation DNA sequencing techniques. Nature Biotechnology. 2009. vol. 25. no. 4. pp. 195–203.

3. Bentley R.D. Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008. vol. 456. no. 7216. pp. 53–59.
4. Whiteford N. The Solexa pipeline. 2012. Available at: http//41j.com/blog/wp-content/uploads/2012/04/pipeline.pdf (accessed 20.02.2024).
5. Leshkowitz D. Introduction to Deep-Sequencing Data Analysis Illumina Primary Analysis Pipeline & Quality Control. 2017. Available at: http://dors.weizmann.ac.il/course/course2017/Dena_IlluminaPrimaryAnalysisPipeline-course2017.pdf (accessed 20.02.2024).
6. Manojlov V.V., Borodinov A.G., Saraev A.S., Petrov A.I., Zaruckij I.V., Kurochkin V.E. [Image processing algorithms in the DNA sequencer NANOPHORE SPS]. Zhurnal tehnicheskoj fiziki – Journal of Technical Physics. 2022. vol. 92. no. 7. pp. 985–992. DOI: 10.21883/JTF.2022.07.52655.318-21. (In Russ.).
7. Manojlov V.V., Borodinov A.G., Zaruckij I.V., Petrov A.I., Kurochkin V.E. [Algorithms for processing fluorescence signals of mass parallel sequencing of nucleic acids]. Trudy SPIIRAN – SPIIRAS Proceedings. 2019. vol. 18. no. 4. pp. 1010–1036. DOI: 10.15622/sp.2019.18.4.1010-1036. (In Russ.).
8. Borodinov A.G., Manojlov V.V., Zaruckij I.V., Petrov A.I., Kurochkin V.E., Saraev A.S. Machine learning in base-calling tasks for next-generation sequencing methods. Informatika i avtomatizacija – Informatics and Automation. 2022. vol. 21. no. 3. pp. 572–603. DOI: 10.15622/ia.21.3.5. (In Russ.).
9. Zhuravel' I.M. Kratkij kurs teorii obrabotki izobrazhenij. Available at: http://matlab.exponenta.ru/imageprocess/book2/49.php (accessed 26.10.2023). (In Russ.).
10. Vuds R., Gonsales R. Cifrovaya obrabotka izobrazhenij [Digital image processing]. 3rd ed. Moscow: Tekhnosfera, 2012. 1104 p. (In Russ.).
11. 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.
12. Sizikov V.S., Stepanov A.V., Mezhenin A.V., Burlov D.I., Eksemplyarov R.A. 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. 203–110.
13. Bardin B.V., Chubinskij-Nadezhdin I.V. [Local object detection in digital microscopic images]. Nauchnoe priborostroenie – Nauchnoe Priborostroenie. 2009. vol. 19. no. 4. pp. 96–102. (In Russ.).
14. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems. IEEE Transactions on Systems, Man and Cybernetics. 1979. vol. 9. pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.
15. Saraev A.S., Petrov A.I., Manojlov V.V. [Modeling the generation of clusters of DNA molecules in massively parallel sequencing devices] Tezisy dokladov Chetvertoj mezhdunarodnoj konferencii so shkoloj molodyh uchenyh «Fizika – naukam o zhizni» [Abstracts of reports of the Fourth International Conference with the School of Young Scientists «Physics – Life Sciences»]. SPb: Ioffe Istitute. 2021. pp. 153 (In Russ.).
16. Saraev A.S. Nauchno-kvalifikacionnaya rabota «Razrabotka algoritma raspoznavaniya klasterov nukleinovyh kislot v mikroflyuidnoj yachejke sekvenatora «Nanofor SPS» [Scientific qualification work «Development of an algorithm for recognizing nucleic acid clusters in the microfluidic cell of the Nanofor SPS sequencer»]. SPb: IAIRAS, 2023. pp. 16–22. (In Russ.).
17. Schmidt B., Hildebrandt A. Deep learning in next-generation sequencing. Drug discovery today. 2021. vol. 26. no. 1. pp. 173–180.

18. Ozgur S., Orman M. Application of deep learning technique in next generation sequence experiments. Journal of Big Data. 2023. vol. 10. no. 1. DOI: 10.1186/s40537-023-00838-w.
19. Tegfalk E. Application of machine learning techniques to perform base-calling in next-generation DNA sequencing. 2020. 45 p.
20. Cacho A., Smirnova E., Huzurbazar S., Cui X. A comparison of base-calling algorithms for illumina sequencing technology. Briefings in bioinformatics. 2016. vol. 17. no. 5. pp. 786–795.
21. Kircher M., Stenzel U., Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome biology. 2009. vol. 10(8). DOI: 10.1186/gb-2009-10-8-r83.
22. Ghojogh B., Crowley M. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv preprint arXiv:1905.12787. 2019.
23. Borodinov A.G., Yamanovskaya A.Yu., Manojlov V.V., Petrov A.I. [Optimal feature selection for reducing the dimensionality of machine learning models in the base-calling problem] Tezisy dokladov Vtoroj ezhegodnoj vserossijskoj molodezhnoj konferencii po metodam i priboram dlja analiza biologicheskih obektov «AnalitBioPribor-2023» [Abstracts of reports of the Second Annual All-Russian Youth Conference on Methods and Instruments for the Analysis of Biological Objects «AnalitBioPribor-2023»]. St. Petersburg: Publishing and Printing Association of Higher Educational Institutions, 2023. pp. 135–138. (In Russ.).
24. Whiteford N., Skelly T., Curtis C., Ritchie M.E., Lohr, A., Zaranek A.W., Abnizova I., Brown C. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009. vol. 25. no. 17. pp. 2194–2199.
25. Hafiz A.M., Bhat G.M. A survey on instance segmentation: state of the art. International journal of multimedia information retrieval. 2020. vol. 9. no. 3. pp. 171–189.
26. Chaudhry M., Shafi I., Mahnoor M., Vargas D.L.R., Thompson E.B., Ashraf I.A. Systematic literature review on identifying patterns using unsupervised clustering algorithms: a Data mining perspective. Symmetry. 2023. vol. 15. no. 1679. DOI: 10.3390/sym15091679.
27. Khandare A., Pawar R. Data clustering algorithms: experimentation and comparison. Intelligent Computing and Networking: Proceedings of IC-ICN 2021. 2022. pp. 86–99.
28. Sarang P. BIRCH: Divide and Conquer. Thinking Data Science: A Data Science Practitioner's Guide. Cham: Springer International Publishing. 2023. pp. 229–236.
29. Ester M., Kriegel H.P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. kdd. 1996. vol. 96. no. 34. pp. 226–231.
30. Zhao Y., Cao J., Zhang C., Zhang S. Enhancing grid-density based clustering for high dimensional data. Journal of Systems and Software. 2011. vol. 84. no. 9. pp. 1524–1539.
31. Wolowski V.R. High-quality, high-throughput measurement of protein-DNA binding using HiTS-FLIP. Dissertation, LMU Munchen: Fakultat fur Chemie und Pharmazie 2016. DOI: 10.5282/edoc.19445.
32. Chityala R., Pudipeddi S. Image processing and acquisition using Python. Chapman and Hall/CRC. 2020. 452 p.
33. Kameshwaran K., Malarvizhi K. Survey on clustering techniques in data mining. International Journal of Computer Science and Information Technologies. 2014. vol. 5. no. 2. pp. 2272–2276.

Manoilov Vladimir — Ph.D., Dr.Sci., Associate Professor, Head of the laboratory, Laboratory of automation of measurements and digital signal processing, IAI RAS. Research interests: representation and processing of signals and images in analytical devices. The number of publications — 101. manoilov-vv@mail.ru; 31-33A, Ivan Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0720.

Borodinov Andrew — Ph.D., Senior researcher, Laboratory of methods and instruments for immune and genetic analysis, IAI RAS. Research interests: mathematical statistics, problems of analysis, processing and presentation of data, artificial intelligence. The number of publications — 15. borodinov@gmail.com; 31-33A, Ivan Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0719.

Zarutsky Igor — Ph.D., Senior researcher, Laboratory of automation of measurements and digital signal processing, IAI RAS. Research interests: representation and processing of signals and images in analytical devices. The number of publications — 51. igorzv@yandex.ru; 31-33A, Ivan Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0719.

Petrov Alexander — Head of the sector of electronics and software, Laboratory of methods and instruments for immune and genetic analysis, IAI RAS. Research interests: representation and processing of signals and images in analytical devices. The number of publications — 21. fataip@mail.ru; 31-33A, Ivan Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0719.

Saraev Alexey — Engineer of the 1st category, Laboratory of methods and instruments for immune and genetic analysis, IAI RAS. Research interests: representation and processing of signals and images in analytical devices. The number of publications — 4. alex.niispb@yandex.ru; 31-33A, Ivan Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0719.

Kurochkin Vladimir — Ph.D., Dr.Sci., Professor, Head of the scientific direction, head of the laboratory, Laboratory of methods and instruments for immune and genetic analysis, IAI RAS. Research interests: research and optimization of electromigration analysis methods, development of analytical methods for capillary electrophoresis, study of optical methods of detection, development of methods and instruments for DNA analysis, development of methods for preparing samples and specialized reagents. The number of publications — 200. lavrovas@yandex.ru; 31-33A, Ivan Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0719.

Acknowledgements. This research was performed in accordance with the State Assignment of the Ministry of Science and Higher Education of the Russian Federation No. 075-01157-23-00 dated 12/29/2022.