

© Е. К. Васильева, И. В. Чубинский-Надеждин, 2025

МОДЕЛИРОВАНИЕ ПЕРВИЧНЫХ ДАННЫХ ОДНОМОЛЕКУЛЯРНОГО СЕКВЕНАТОРА. Ч. 1. СИМУЛЯЦИЯ ДЕТЕКТИРОВАНИЯ ИЗЛУЧЕНИЯ ВОЛНОВОДА НУЛЕВОЙ МОДЫ

В работе предложена стохастическая имитационная модель процесса детектирования флуоресценции меченных нуклеотидов при одномолекулярном секвенировании в четырех спектральных каналах. Обеспечена многофакторная имитация сигналов детектора. В первой части статьи описан алгоритмический шаблон для моделирования отклика светочувствительных элементов видеокамеры на свет, поступающий из одного волновода нулевой моды в одном спектральном диапазоне в процессе секвенирования одиночной молекулы ДНК. Введено понятие "базового изображения", имитирующего число фотоэлектронов в сенсорах за время экспозиции. Область симуляции такого изображения представляет собой квадратное поле размером (2×2) , ..., (5×5) пикселей, суммарная яркость которого дает единичный отсчет трейса в данном канале. Рассматриваются три режима формирования базового изображения: 1) только темновой ток, 2) фоновое излучение при отсутствии встраивания нуклеотида и 3) световое пятно флуоресцентной вспышки с фоновым уровнем при встраивании. Для моделирования базового изображения приведен пошаговый алгоритм: задание размеров и смещения пятна, амплитуд флуоресценции и фона; формирование двумерных гауссовых распределений для фона и сигнала и их суммирование; учет сигнал-зависимого шума с использованием аппроксимации результатов экспериментов с детектором и сигнал-независимого шума с использованием нормального распределения и параметров камеры; задание смещения нулевого уровня и финальное суммирование шумовых компонентов с изображением. Также описан алгоритм моделирования фонового сигнала с учетом спектральной помехи от соседних каналов. Разработана программа-симулятор первичных цифровых данных одномолекулярного секвенатора, обеспечивающая имитацию результатов секвенирования в форматах FASTA, TRACE и MOVIE. На вход симулятора может быть подана реальная последовательность ДНК, полученная из геномной базы.

Ключевые слова: симулятор последовательности, длинные риды, имитационное моделирование, одномолекулярное секвенирование, волновод нулевой моды

ВВЕДЕНИЕ

Одномолекулярное секвенирование — перспективный и бурно развивающийся метод получения генетической информации в тех областях, где длина считываемых ридов имеет решающее значение для интерпретации данных. Наиболее сложные задачи метагеномики и эпигенетики — секвенирование *de novo*, анализ большого числа повторяющихся и гомологичных участков генома, другие исследования, связанные с интерпретацией длинных геномных последовательностей, — не могут быть эффективно решены только методами NGS, обеспечивающими чтение коротких ридов.

Разработки компаний Pacific Biosciences (PacBio) и Oxford Nanopore Technologies (ONT)

представляют собой две основные технологии секвенирования третьего поколения, по которым сегодня развивается одномолекулярное секвенирование. Первая технология, названная SMRT (Single Molecule Real Time sequencing), вначале обеспечивала получение непрерывных длинных чтений (continuous longreads, или CLR) с уровнем ошибок 15%, а затем получила развитие в методе Hi-Fi ридов, снижающем уровень ошибок до величины $< 1\%$. Вторая технология дает возможность получать ультрадлинные риды до 4 Mbp, но с высоким уровнем ошибок, которые можно снизить почти до уровня метода Hi-Fi, используя технологии "дуплексного" секвенирования [1]. Обе технологии послужили для завершающего этапа расшифровки генома человека — секвенирования

Y-хромосомы, имеющей сложную повторяющуюся структуру [2].

В открытом доступе существует несколько симуляторов длинных последовательностей, имитирующих данные секвенаторов третьего поколения. Это, например, программы SimLoRD, NPBSS, PBSIM, PaSS [3–6]. Они позволяют получить последовательности длинных ридов (longreads) в формате FASTA-файлов, моделируя различные ошибки чтения, что дает возможность использовать эти данные в различных задачах биоинформатики для отладки алгоритмов вторичной обработки. Такие симуляторы дают высокую степень приближения к результатам, полученным при реальном секвенировании. Однако они не позволяют получить первичные, или сырые (raw), данные, т.е. изображения (movies) и трейсы (traces) для SMRT-секвенаторов. От них отличается симулятор Nanopore Simulation, который специально предложен как симулятор первичных данных для технологии ONT-сигналов ионного тока [7].

Симуляторы первичных данных для секвенаторов, регистрирующих сигналы флуоресценции, отсутствуют в публичном доступе. PacBio в своих внутренних тестах, очевидно, использует собственные генераторы симуляций, которые не публикуются в открытом виде. Это представляет проблему в случае необходимости имитации первичных данных секвенаторов PacBio, например, для разработки и оптимизации собственных алгоритмов первичной обработки данных типа movie и trace для оптимизации процедуры поиска баз (base-calling).

Таким образом, для решения задач распознавания флуоресцирующих объектов на исходных изображениях и дальнейшей интерпретации сигналов в методе одномолекулярного секвенирования ввиду недостатка и дороговизны получения экспериментальных данных и отсутствия симуляторов становятся актуальными задачи имитационного моделирования и программирования.

Разработка собственного симулятора — задача сложная и нетривиальная, поскольку реалистичная симуляция требует моделирования кинетики полимеразы, характеристик красителей, используемых для мечения нуклеотидов, особенностей оптики, шумов системы детектирования и прочих параметров, большая часть которых имеет вероятностный характер. Очевидно, что для качественной имитации данных требуется разработка нового метода моделирования. Наиболее целесообразным представляется подход "виртуального секвенирования", при котором обеспечивается многофакторная имитация сигналов, возникающих на светочувствительных элементах видеокамеры, сформированных с учетом перечисленных выше параметров.

СИСТЕМА ДЕТЕКТИРОВАНИЯ ОДНОМОЛЕКУЛЯРНОГО СЕКВЕНАТОРА

Систему детектирования при одномолекулярном секвенировании по технологии, близкой к PacBio [8–12], можно представить как специализированный четырехканальный, по числу спектральных диапазонов, флуоресцентный микроскоп с цифровой регистрацией изображений флуоресцирующих объектов ВНМ¹ (волноводов нулевой моды). Каждый спектральный диапазон соответствует одному из четырех флуоресцентно меченных нуклеотидов, традиционно обозначаемых А, Т, G и С. ВНМ представляет собой наноструктуру, в которой протекает реакция секвенирования. Это цилиндрический наноклодез с глубиной и шириной около 100 нм, сформированный в тонкой пленке алюминия, объем которого не превышает 10^{-19} л. В каждом ВНМ в процессе встраивания полимеразой очередного меченого нуклеотида из реакционного буферного раствора происходит эмиссия флуоресценции соответствующей метки под действием возбуждающего лазерного излучения, падающего на прозрачную поверхность измерительной ячейки снизу. Ячейка содержит большое число ВНМ и находится в поле зрения микрообъектива высокого разрешения с целью одновременного наблюдения флуоресценции во всех наноструктурах.

Выходящие из объектива пучки света несут информацию о пространственном распределении ВНМ и спектральном составе излучения. Для разделения спектральных диапазонов регистрируемой флуоресценции световые пучки проходят через дисперсионную призму [13]. При построении изображения излучение от каждого ВНМ будет разделяться в направлении дисперсии призмы в соответствии со спектральным составом флуоресценции четырех излучающих меток (рис. 1).

Свет попадает на фотосенсоры видеокамеры, где в течение заданного времени экспозиции генерируется изображение. Для отображения всех ВНМ используется одна крупноформатная высокочувствительная CMOS-видеокамера [14]. На фотосенсорах видеокамеры для каждого ВНМ можно выделить четыре смежных группы пикселей, в которых сигналы генерируются как световые пятна в соответствии с интенсивностью и спектральным составом попавшего на них света.

Размер и форма световых пятен будут определяться в основном характеристиками и разрешающей способностью оптической системы при построении изображения субмикронного объекта.

¹ В англоязычной литературе ZMW (Zero-Mode Waveguide).

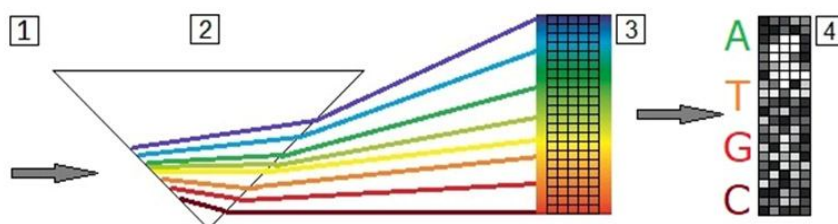


Рис. 1. Соответствие каналов детектирования дисперсии света на призме. 1 — индуцированное лазером излучение от ВМ; 2 — дисперсионная призма; 3 — дисперсия света на сенсорах видеокамеры; 4 — сформированное изображение ВМ в четырех спектральных каналах. Используется участок спектра 550–750 нм

Размеры изображения в пикселях зависят от дисперсии призмы, выбранного увеличения и физического размера сенсоров видеокамеры. Изображения световых пятен для разных нуклеотидов частично перекрываются в соответствии с перекрытием спектров флуоресцентных меток.

Полученное для каждого кадра изображение передается в компьютер. Серия таких изображений формирует набор сырых данных movies, которые подвергаются дальнейшей обработке. В процессе первичной обработки в компьютере проводится локализация изображений всех флуоресци-

рующих ВМ на общем изображении кадра и осуществляется трассировка сигналов флуоресценции для каждого спектрального канала, в результате чего формируются трейсы — наборы данных, где хранится информация об изменении интенсивности флуоресценции во времени. После предварительной фильтрации трейсов по выбранным критериям качества проводится basecalling-обнаружение и интерпретация импульсов в каждом спектральном канале, что в итоге обеспечивает формирование нуклеотидной последовательности (рис. 2).

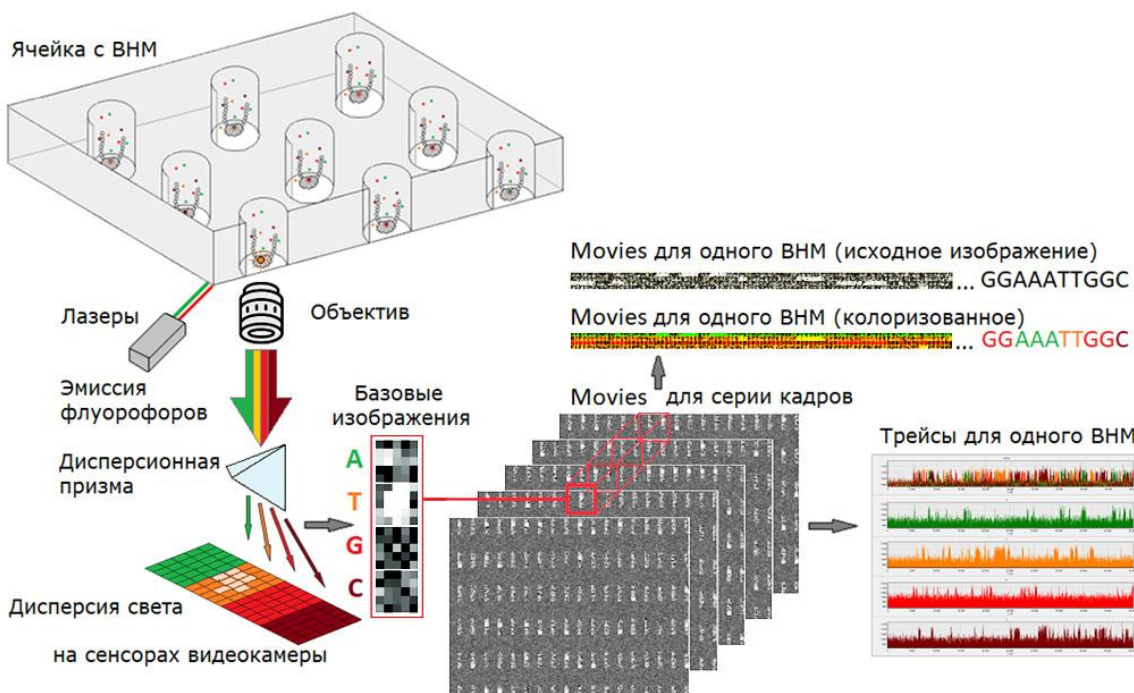


Рис. 2. Обобщенная схема детектирования флуоресценции при одномолекулярном секвенировании с использованием дисперсионной призмы.

МОДЕЛИРОВАНИЕ БАЗОВЫХ ИЗОБРАЖЕНИЙ И СИГНАЛОВ ДЛЯ ОДНОГО ВНМ

Принцип формирования базового модельного изображения

На первом этапе формируется алгоритмический и программный шаблон, предназначенный для моделирования процессов, протекающих на светочувствительных элементах видеокамеры при попадании на них света в каком-либо одном спектральном диапазоне от одного ВНМ.

В результате может быть получено *базовое изображение*, имитирующее количество фотоэлектронов, накопленных в соответствующих сенсорах видеокамеры за некоторое время экспозиции. Базовое изображение одиночного ВНМ формируется внутри заданной области симуляции.

Область симуляции светового пятна для одного спектрального диапазона представляет собой квадратное изображение, размер которого может быть установлен выборочно от 2×2 до 5×5 пикселей. Один пиксель такого базового изображения соответствует одному сенсору видеокамеры. Уровень яркости в одном пикселе соответствует числу фотоэлектронов, накопленных в соответствующем сенсоре за время однократной экспозиции. Сумма яркостей всех пикселей в области симуляции формирует интегральный *базовый сигнал*, который

становится единичным отсчетом трейса в одном спектральном канале (рис. 3).

На рис. 3:

– Лазер 1 индуцирует флуоресценцию в ВНМ 2 с иммобилизованной на дне полимеразой, фрагментом цепи ДНК и мечеными нуклеотидами в растворе.

– В процессе встраивания флуоресцентно-меченного нуклеотида из ВНМ под действием лазерного излучения испускается свет, который, попадая на сенсоры видеокамеры 3, в каждом формирует эквивалентный электрический сигнал в фотоэлектронах (ф. эл.).

– Накопленные в сенсорах камеры сигналы могут быть смоделированы в виде пикселей разной яркости в области симуляции заданного размера 4, где уровень яркости в пикселе пропорционален числу фотоэлектронов в соответствующем сенсоре видеокамеры.

– Суммарный сигнал всех пикселей базового изображения формирует интегральный базовый сигнал, соответствующий одному отсчету на трейсе 5 в моделируемом спектральном канале.

В качестве примера реального трейса на рис. 4 показан фрагмент длительностью 30 с. Данные получены при регистрации флуоресценции в канале G для одиночной молекулы стрептавидина, иммобилизованной на дне ВНМ.

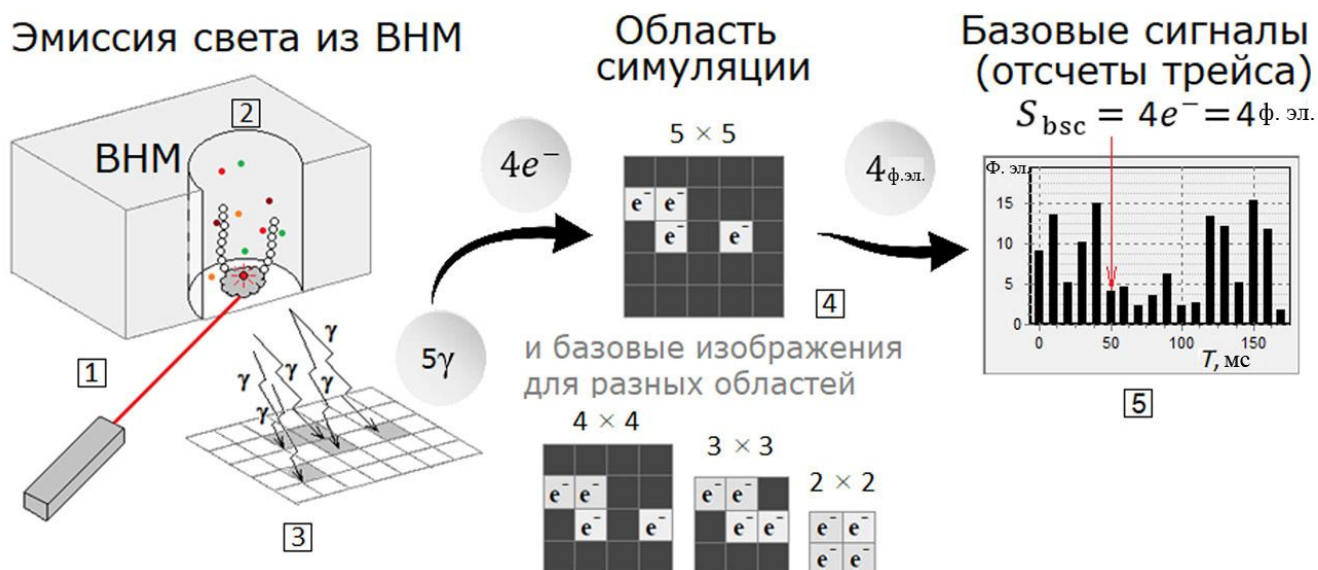


Рис. 3. Формирование базового изображения и базового сигнала ВНМ в одном спектральном канале. 1 — лазер, индуцирующий флуоресценцию; 2 — волновод нулевой моды (ВНМ); 3 — сенсоры видеокамеры; 4 — пиксели возможных моделей базовых изображений; 5 — отсчеты трейса моделируемого спектрального канала

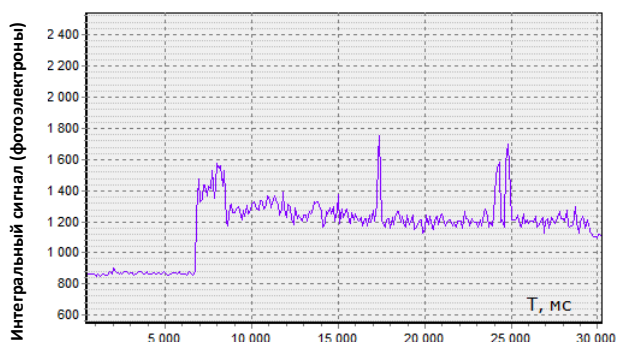


Рис. 4. Фрагмент трейса

Молекула стрептавидина помечена красителем Су5. В ВНМ добавлены меченные этим же красителем нуклеотиды, свободно плавающие в буферном растворе. На отметке 6.5 с включили красный лазер и регистрировали выгорание красителя на стрептавидине, после 15 с добавили меченные нуклеотиды. Интегральный сигнал интенсивности свечения ВНМ определялся на участке изображения размером 3×3 пикселя. Интенсивность фоно-

вого свечения ячейки после включения лазера составляет примерно 1200 фотоэлектронов, сигналы флуоресценции в ВНМ (базовые сигналы, или отсчеты трейса) превышают по амплитуде 200–400 фотоэлектронов.

Имитация фона, полезного сигнала и темнового шума

Если рассматривать базовый сигнал с точки зрения того, произошло или не произошло событие встраивания нуклеотида, базовое изображение может быть сформировано как:

1. Симуляция наличия темнового тока в сенсорах видеокамеры в случае, если свет от ВНМ не излучается вообще (ситуация "а").
2. Симуляция сигнала фоновое излучения для случая, когда встраивания нуклеотида в данном спектральном диапазоне в течение времени текущей экспозиции не произошло (ситуация "b").
3. Симуляция светового пятна от флуоресцентной вспышки, в том числе и сигнала фона, для имитации процесса встраивания нуклеотида (ситуация "с").

Эти ситуации представлены на рис. 5.

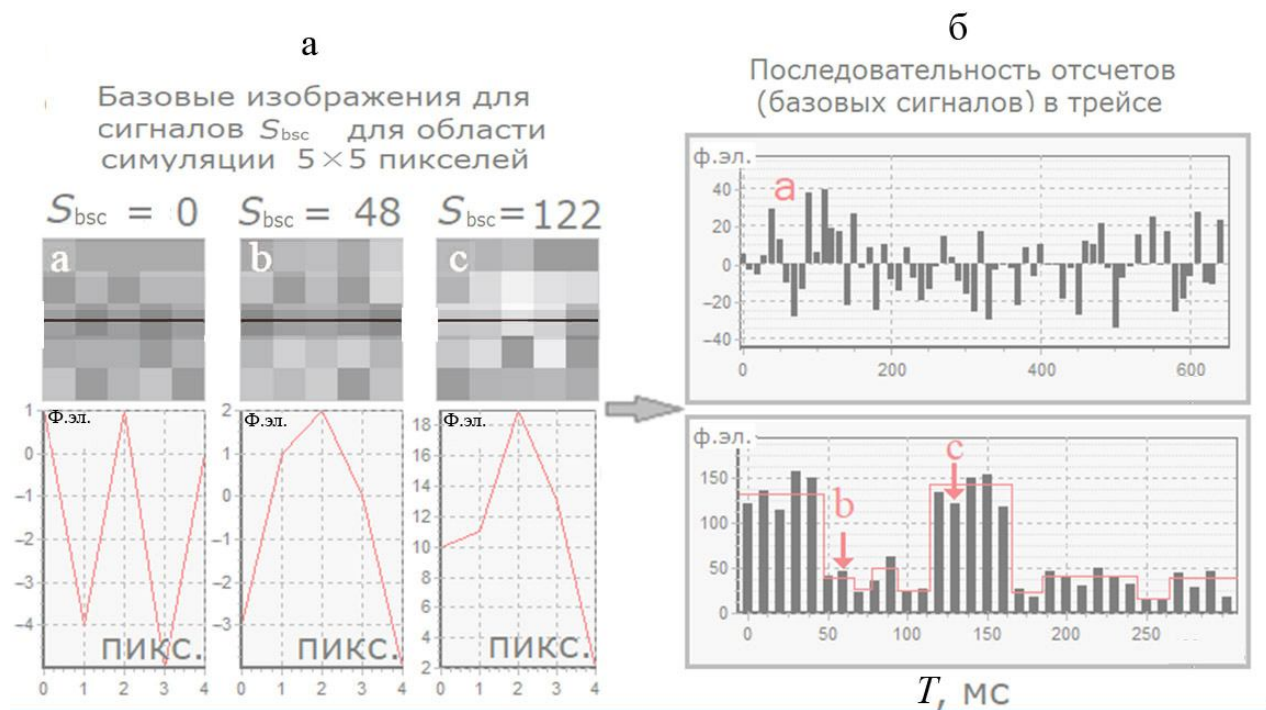


Рис. 5. Базовые изображения области симуляции размером 5×5 пикселей (а) и профили яркости для них (б). На б сверху — отсчеты базовых сигналов при имитации темнового шума камеры (ситуация "а"). На б внизу — отсчеты базовых сигналов при имитации фона ячейки (ситуация "b") и событий встраивания нуклеотидов (ситуация "с").

За время симуляции, равное 310 мс, произошло два события встраивания. Каждое событие образовано пятью отсчетами с интенсивностью, превышающей 100 фотоэлектронов (ф.эл.). Отсчеты фоновых сигналов сформированы базовыми сигналами меньшей интенсивности

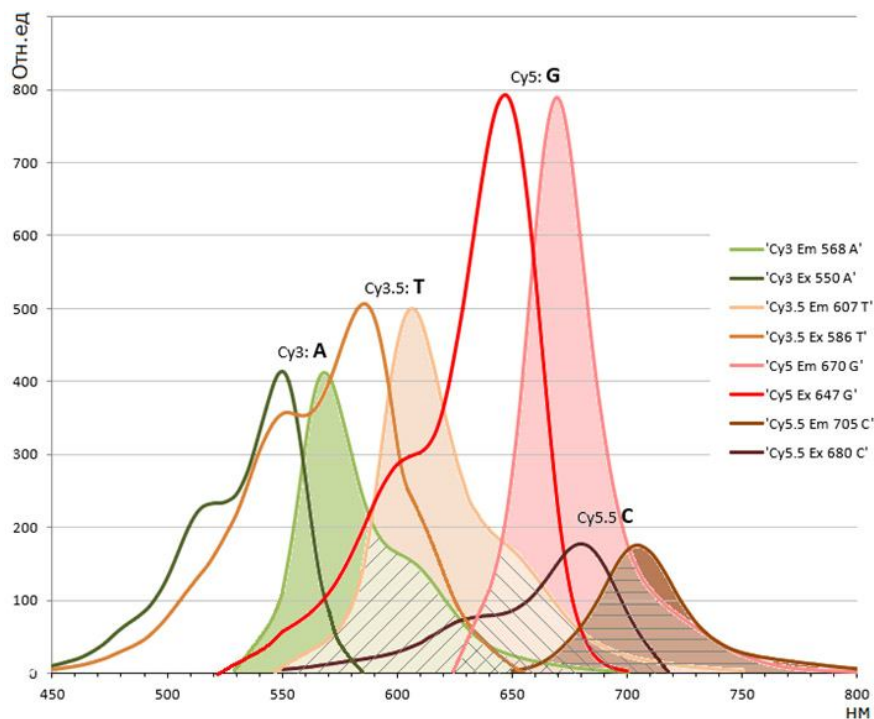


Рис. 6. Спектры возбуждения и эмиссии флуоресцентных красителей, используемых для мечения нуклеотидов. Заштрихованы области взаимного спектрального перекрытия между каналами А и Т, Т и G, G и С

Перекрестные спектральные помехи

При синтезе комплементарной цепи ДНК два разных нуклеотида не могут одновременно встроиться в первичную цепь, поэтому в одном ВНМ не могут одновременно светиться два флуорофора. Однако в момент свечения какого-либо одного нуклеотида в соседние с ним каналы детекции могут попадать перекрестные спектральные помехи (crosstalk). В модельном сигнале величина такой помехи может быть задана в процентах от интенсивности сигнала флуоресценции в канале, который является источником помехи. Величина помехи добавляется к интегральному сигналу яркости в моделируемом канале и зависит от формы спектров возбуждения и испускания используемых красителей (рис. 6). Помеху из канала А в канал Т обозначим как $CrTk^{AT}$ и, наоборот, помеху из канала Т в канал А обозначим как $CrTk^{TA}$. Логика обозначений для других спектральных каналов аналогична.

Для моделирования разделения лучей от четырех флуоресцентных красителей при помощи дисперсионной призмы следует расположить четыре базовых изображения друг под другом по вертикали в порядке увеличения длины волны спектров эмиссии А (зеленый), Т (желтый), G (красный) и С

(темно-красный). Каждое из четырех базовых изображений генерируется отдельно в соответствии с заданными для данного спектрального канала параметрами.

Алгоритм моделирования базового изображения Sbasic

Алгоритм моделирования (вариант 1) базового изображения в области симуляции для канала А (SbasicA) в случае встраивания меченного нуклеотида А (рис. 7):

1. Задать размер области симуляции в пикселях и смещение моделируемого светового пятна внутри области.
2. Задать амплитуду сигнала флуоресценции А в фотозлектронах — Fl^A .
3. Задать амплитуду фонового сигнала $BkGrd^A$, т.е. величину сигнала собственного свечения измерительной ячейки в отсутствие сигнала флуоресценции от встраиваемого нуклеотида в моделируемом канале.
4. Сформировать двумерные изображения гауссова распределения яркости в области симуляции для фона и флуоресценции согласно заданным параметрам. Суммировать оба изображения.

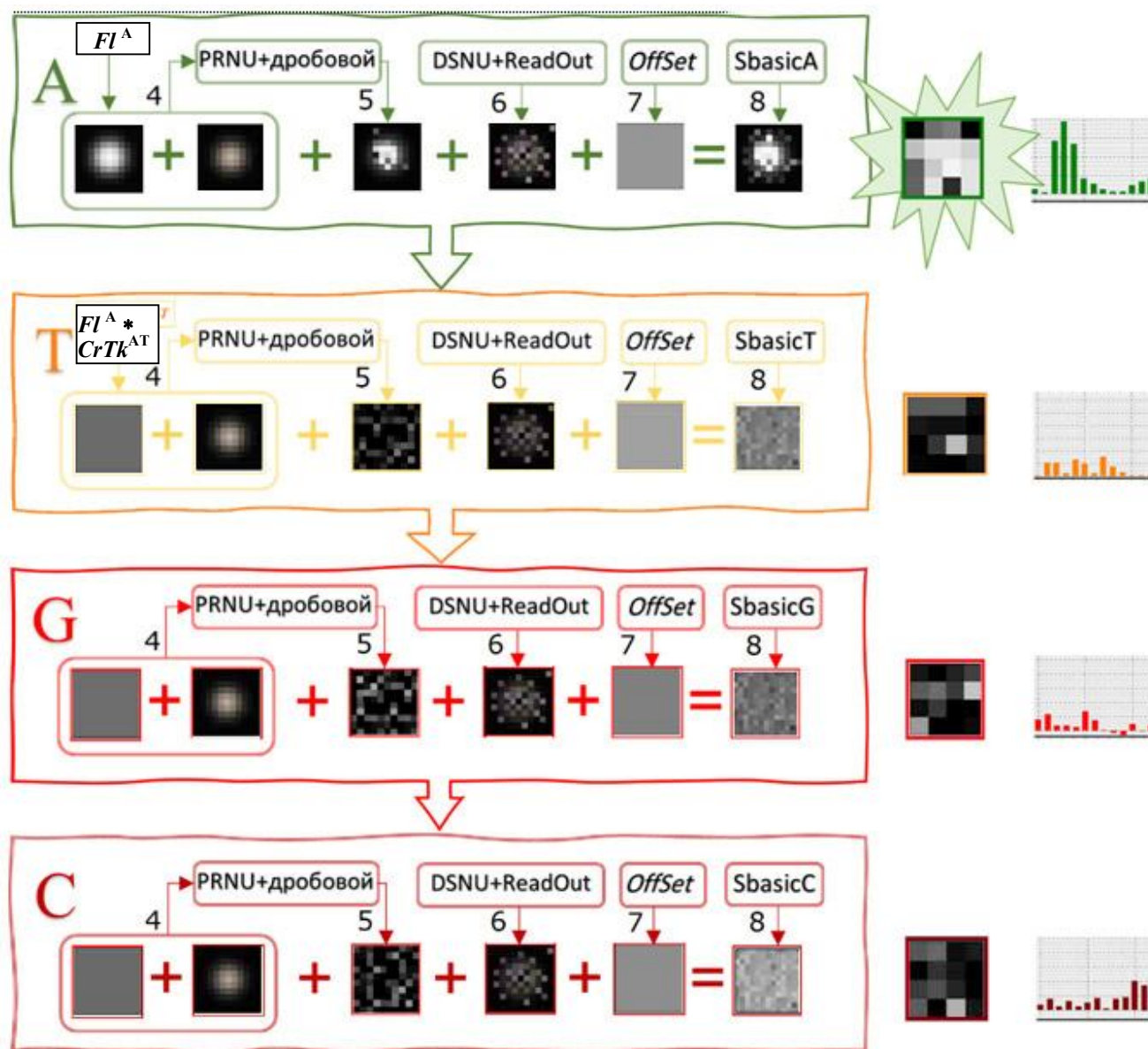


Рис. 7. Формирование базового изображения $Sbasic$ для каждого из четырех спектральных каналов. В канале A моделируется полезный сигнал (наличие флуоресценции). Цифрами показаны соответствующие этапы алгоритма моделирования

5. Моделировать шум, зависящий от величины сигнала, накопленного в сенсорах за время экспозиции, как сумму PRNU (Photo Response Non-Uniformity) и дробового шума (Shotgunnoise) — N_{dp} . Для этого рассчитать величину шумов в каж-

дом пикселе для полученного в п. 4 суммарного изображения.²

² С помощью аппроксимации экспериментально полученной зависимости СКО интегрального базового сигнала, измеренного в заданной квадратной области $M \times M$ пикселей, от величины этого сигнала.

6. Моделировать шум, не зависящий от величины сигнала, накопленного в сенсорах, как сумму темнового шума DSNU (Dark Signal Non-Uniformity) и шума считывания камеры ReadOut — N_{undp} . Для этого рассчитать величину шумов в каждом пикселе с помощью нормального распределения со средним, равным нулю, и СКО, указанным в паспорте видеокамеры для данных видов шумов.

7. Задать смещение интенсивности сигнала относительно нуля для каждого пикселя — $Offset$.³

8. Добавить рассчитанные шумы и смещение к изображению фона и флуоресценции, полученному в п. 4.

Алгоритм моделирования (вариант 2) базового изображения в области симуляции для канала Т (SbasicT) (фон + спектральная помеха из А) в случае встраивания меченного нуклеотида А (рис. 7):

1. Задать размер области симуляции в пикселях и смещение моделируемого светового пятна внутри области.

2. Задать спектральную помеху (crosstalk) в канале Т в процентах от амплитуды сигнала из канала А, создающего эту помеху — $Fl^A \times CrTk^{AT}$.

3. Задать амплитуду фонового сигнала $BkGrd^T$ — величину сигнала собственного свечения измерительной ячейки в отсутствие сигнала флуоресценции от встраиваемого нуклеотида в моделируемом канале.

4. Сформировать двумерные изображения гауссова распределения яркости в области симуляции для фона и флуоресценции согласно заданным параметрам. В данном случае собственная флуоресценция в канале Т равна нулю, но формируется паразитный флуоресцентный сигнал от засветки из канала А. Суммировать оба изображения.

5. Моделировать шум, зависящий от величины сигнала, накопленного в сенсорах за время экспозиции, как сумму PRNU (Photo Response Non-Uniformity) и дробового шума (Shotgunnoise) — N_{dp} . Для этого рассчитать величину шумов в каждом пикселе для полученного в п. 4 суммарного изображения.

6. Моделировать шум, не зависящий от величины сигнала, накопленного в сенсорах, как сумму темнового шума DSNU (Dark Signal Non-Uniformity) и шума считывания камеры ReadOut — N_{undp} . Для этого рассчитать величину шумов в каждом пикселе с помощью нормального распределения со средним, равным нулю, и СКО, указанным в паспорте видеокамеры для данных видов шумов.

7. Задать смещение интенсивности сигнала относительно нуля для каждого пикселя — $Offset$.

8. Добавить рассчитанные шумы и смещение к изображению фона и флуоресценции, полученному в п. 4.

При необходимости моделировать темновой ток в сенсорах используются только пункты 6 и 7 алгоритма.

Пояснение к п. 4 алгоритма. Из ВНМ на видеокамеру попадает суммарный поток света, порождаемый как сигналом флуоресценции нуклеотида, если произошло встраивание, так и общим фоновым излучением ВНМ, обусловленным свечением поверхности ячейки и слабым сигналом от меченых нуклеотидов, плавающих в растворе. В обоих случаях изображение светового пятна в области симуляции может быть смоделировано двумерным гауссовым распределением яркости (1) с заданным σ ($\sigma_x = \sigma_y = \sigma$), которое соответствует радиусу светящегося объекта, и заданным случайным смещением центра пятна m_x, m_y . При этом сам оптический сигнал соответствует величине двумерного гауссового интеграла (рис. 8)

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(x-m_x)^2}{2\sigma_x^2} - \frac{(y-m_y)^2}{2\sigma_y^2}}. \quad (1)$$

Смещение центра симметрии пятна задается опционально случайными величинами с равномерным распределением в пределах области симуляции и позволяет учесть проявление пространственной нестабильности изображений.

Общая яркость всех пикселей изображения в области симуляции определяется заданной величиной базового сигнала Fl (например, только для флуоресцентного сигнала). Программно базовый сигнал формируется как сумма яркостей всех пикселей в базовом изображении. Для этого каждый пиксель в области симуляции получает яркость $A_{i,j}$ с соответствующим весовым коэффициентом, как показано в (2):

$$A_{i,j} = \frac{Fl}{2\pi\sigma^2} e^{-\frac{(i-m_x)^2}{2\sigma^2} - \frac{(j-m_y)^2}{2\sigma^2}}. \quad (2)$$

В итоге алгоритм формирования модельного двумерного базового изображения S_{bsc} в области симуляции может быть так описан формулами.

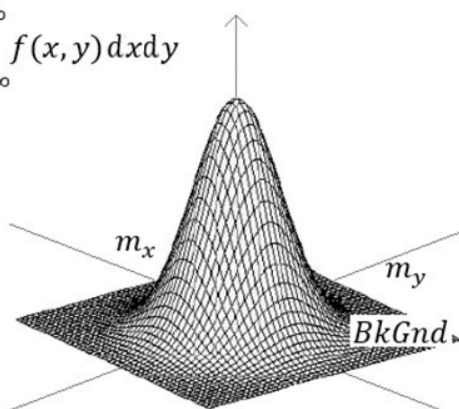
• Моделируем наличие флуоресценции (пример для канала А):

$$S_{bsc}^{A-Fl} = (Fl^A * f(x, y) + BkGrd^A * f(x, y)) + N_{dp} + N_{undp} + Offset. \quad (3)$$

³ Параметр $Offset$ видеокамеры, задающий смещение нулевого уровня, в рассматриваемом случае равен 100 единиц яркости на пиксель.

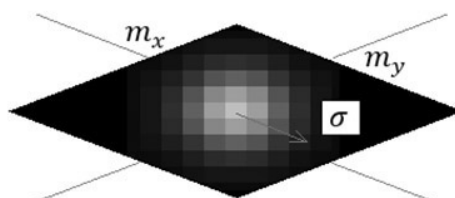
Объем под поверхностью определяет интегральный оптический сигнал Fl

$$Fl = \iint_{-\infty}^{\infty} f(x, y) dx dy$$



Базовое изображение (двумерный гауссиан), где m_x, m_y – положение центра

$$f(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-m_x)^2}{2\sigma^2} - \frac{(y-m_y)^2}{2\sigma^2}}$$



Амплитуда яркости в точке $\{i, j\}$: $A_{i,j} = \frac{Fl}{2\pi\sigma^2} e^{-\frac{(i-m_x)^2}{2\sigma^2} - \frac{(j-m_y)^2}{2\sigma^2}}$

Амплитуда яркости в центре, в точке $\{0, 0\}$: $A_{0,0} = \frac{Fl}{2\pi\sigma^2}$

Рис. 8. Гауссова модель изображения сигнала флуоресценции ВНМ

- Моделируем только фон (пример для канала Т):

$$\begin{aligned} S_{bsc}^{T_Bg} = & \\ = & (Fl^A * CrTk^{AT} * f(x, y) + BkGrd^T * f(x, y)) + \\ + & N_{dp} + N_{undp} + Offset. \end{aligned} \quad (4)$$

- Моделируем темновой кадр:

$$S_{bsc}^{Drk} = N_{undp} + Offset. \quad (5)$$

Обоснование способа моделирования шумов

Пояснение к п. 5 алгоритма. Экспериментальное изучение зависимости шумов видеокамеры от величины регистрируемого сигнала позволило определить СКО сигнала яркости изображения измерительной ячейки, интегрированного в области 3×3 пикселя. На рис. 9 представлена зависимость относительного СКО (определенного в процентах по отношению к соответствующей величине сигнала) такого сигнала яркости для пустой ячейки, засвеченной зеленым и красным лазерами. Под "хорошими" пикселями здесь подразумевается основное множество пикселей камеры, для ко-

торых измеренный разброс сигнала яркости близок к указанному в паспорте видеокамеры значению. Под "плохими" пикселями мы имеем в виду некоторые пиксели с уровнем разброса сигнала, превышающим паспортные значения в 5 и более раз. Попадание таких пикселей в область интегрирования сигнала яркости при построении трейса может увеличить СКО шума фона и затруднить последующий поиск сигналов-событий⁴.

Данные для построения зависимости получены в результате обработки серии кадров, зарегистрированных в режиме темного поля (при отсутствии внешнего источника света) и при включенных лазерах с последовательно увеличивающимся временем экспозиции от 10 мс до 250 мс.

⁴ Сигнал-событие — событие, которое состоит в излучении света в одном из спектральных каналов в процессе комплементарного встраивания флуоресцентно меченного нуклеотида в цепь ДНК. Состоит из совокупности базовых сигналов, которые представляют собой интегральные сигналы яркости для изображения ВНМ, полученные в заданной области на кадре за время однократной экспозиции. На трейсе выглядит, как импульс.

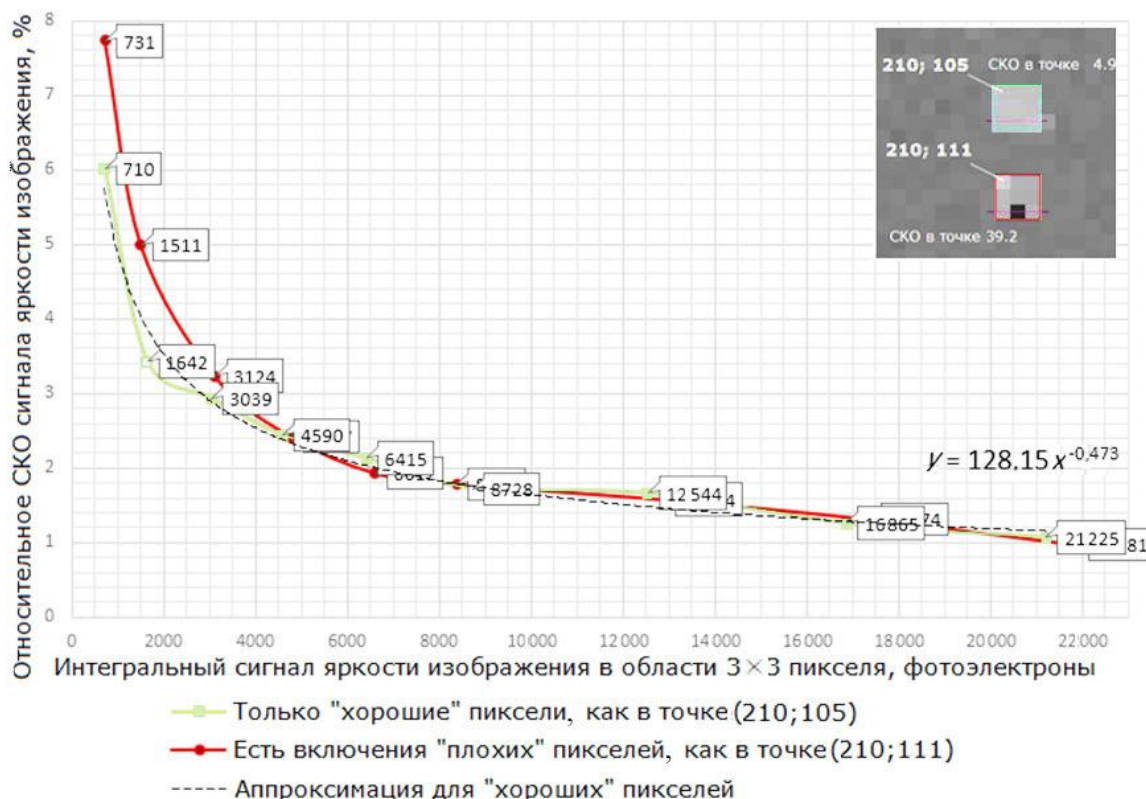


Рис. 9. Зависимость относительного СКО сигнала яркости (интенсивности) изображения фона ячейки в области интегрирования 3×3 пикселя от величины этого сигнала с вычетом смещения уровня *Offset*

Это позволило измерить уровни яркости изображения собственного свечения ячейки в широком диапазоне значений. Для корректной обработки данных из величины интегрального сигнала был вычтен уровень смещения, равный 100 фотоэлектронов на пиксель, искусственно добавляемый в видеокамере для коррекции нулевого уровня. С помощью аппроксимации результатов измерений получена функциональная зависимость величины СКО интегрального сигнала N от уровня интенсивности этого интегрального сигнала S в области 3×3 пикселя:

$$N = 128.15 \times S^{-0.473}. \quad (6)$$

Полученное выражение используется при моделировании базового изображения по формулам (3), (4) для определения величины шумов, зависящих от уровня сигнала $N_{\text{дн}}$. Для этого вместо величины S в (6) следует подставить сумму сигналов фона и флуоресценции для моделируемого пикселя в области симуляции.

РЕЗУЛЬТАТЫ

Таким образом, на симулированном базовом изображении мы будем иметь дело с асимметричным пятном, заданным двумерным гауссовым распределением сигналов фона и флуоресценции, с добавлением спектральных помех и флуктуациями яркости в каждом пикселе. Интегральная яркость базового изображения — сумма яркости в пикселях в пределах области симуляции — также будет флуктуировать в разных реализациях даже при одинаковых параметрах симуляции, создавая таким образом стохастическую имитационную модель процесса детектирования излучения от ВНМ.

Используя такой алгоритм формирования базовых сигналов и изображений многократно, можно моделировать сигналы флуоресценции, превышающие по длительности время экспозиции и соответствующие событиям встраивания нуклеотидов в цепь ДНК во всех четырех спектральных каналах. Представленный способ обеспечивает реалистичную генерацию базовых изображений

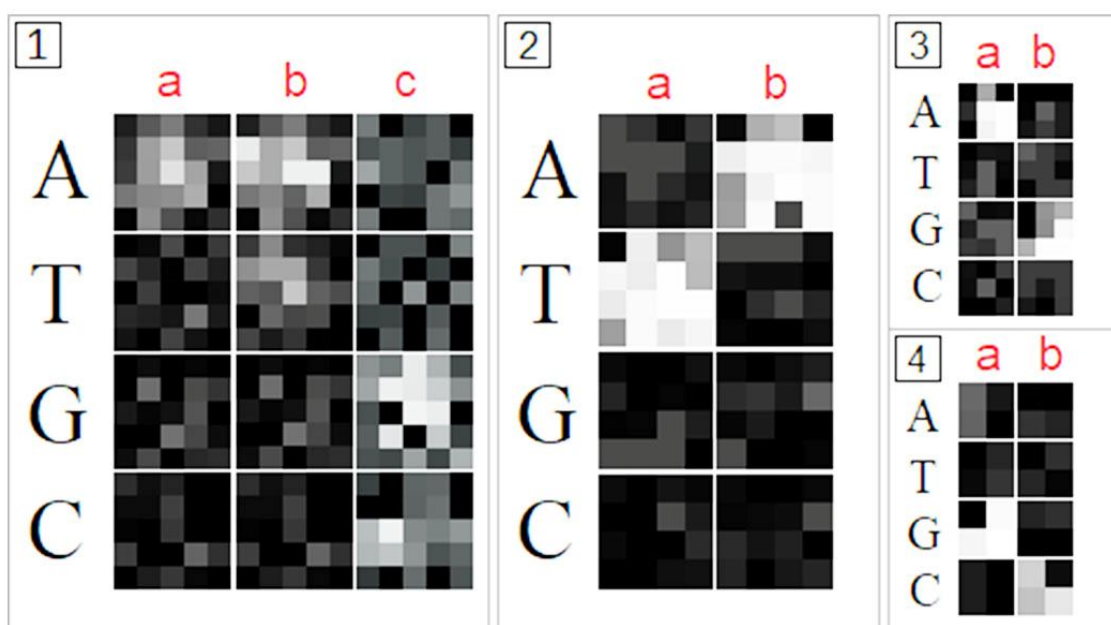


Рис. 10. Примеры базовых изображений нуклеотидов.
Пояснения см. в тексте

для использования их в дальнейшем полноформатном моделировании серии кадров первичных изображений секвенатора, которые могут быть сформированы для тестирования алгоритмов обработки первичных данных и повышения точности распознавания флуоресцентных событий.

На рис. 10 показаны программные симуляции базовых изображений ВНМ в четырех спектральных каналах для разных размеров области симуляции с разными входными параметрами, включая спектральные помехи для некоторых изображений.

На рис. 10:

[1] — область 5×5 пикс.; а — флуоресценция в канале А; б — флуоресценция в канале А со спектральным перекрытием канала Т; с — флуоресценция в канале G со спектральным перекрытием канала С.

[2] — область 4×4 пикс.; а — флуоресценция в канале Т; б — флуоресценция в канале А.

[3] — область 3×3 пикс.; а — флуоресценция в канале А; б — флуоресценция в канале G.

[4] — область 2×2 пикс.; а — флуоресценция в канале G; б — флуоресценция в канале С.

СПИСОК ИСТОЧНИКОВ

1. Mahmoud M., Huang Y., Garimella K. et al. Utility of long-read sequencing for All of Us // Nat Commun. 2024. Vol. 15. Id. 837. DOI: 10.1038/s41467-024-44804-3
2. Rhie A., Nurk S., Cechova M. et al. The complete sequence of a human Y chromosome // Nature. 2023. Vol. 621. P. 344–354. DOI: 10.1038/s41586-023-06457-y
3. Stöcker B.K., Köster J., Rahmann S. SimLoRD: Simulation of Long Read Data // Bioinformatics. 2016. Vol. 32, iss. 17. P. 2704–2706. DOI: 10.1093/bioinformatics/btw286 // Epub 2016 May 10.
4. Wei Z.-G., Zhang Sh.-W. NPBS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model // BMC Bioinformatics. 2018. Vol. 19. Id. 177. DOI: 10.1186/s12859-018-2208-0
5. Ono Y., Asai K., Hamada M. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores // Bioinformatics. 2021. Vol. 37, iss. 5. P. 589–595. DOI: 10.1093/bioinformatics/btaa835
6. Zhang W., Jia B., Wei Ch. PaSS: a sequencing simulator for PacBio sequencing // BMC Bioinformatics. 2019. Vol. 20. Id. 352. DOI: 10.1186/s12859-019-2901-7
7. Rohrandt C., Kraft N., Gießelmann P., et al. . Nanopore SimulatION – a raw data simulator for Nanopore Sequencing // 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid, Spain. P. 1–8. URL: <https://ieeexplore.ieee.org/document/8621253>
8. Eid J., Fehr A., Gray J., Luong K., et al. Real-Time DNA Sequencing from Single Polymerase Molecules // Science. 2009. Vol. 323. Id. 5910. P. 133–138. DOI: 10.1126/science.1162986
9. Korlach J., Bjornson K.P., Chaudhuri B.P., et al. Ch. 20 – Real-Time DNA Sequencing from Single Polymerase Molecules // Methods in Enzymology. Pacific Biosciences, Menlo Park, California, USA, 2010. Vol. 472. P. 431–455. DOI: 10.1016/S0076-6879(10)72001-2

10. Levene M.J., Korlach J., Turner S.W., et al. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations // *Science, New Series*. 2003. Vol. 299, no. 5607. P. 682–686.
URL: <http://www.jstor.org/stable/3833436>
11. Бессмельцев В.П., Терентьев В.С., Вилейко В.В. и др. О создании платформы для исследования нуклеиновых кислот (ДНК-секвенатора) // *Прикладная фотоника*. 2016. Т. 3, № 4. С. 388–412. DOI: 10.15593/2411-4367/2016.04.03
12. Chen J., Dalal R.V., Petrov A.N., et al. High-throughput platform for real-time monitoring of biological processes by multicolor single-molecule fluorescence // *PNAS*. 2014. Vol. 111, no. 2. P. 664–669.
DOI: 10.1073/pnas.1315735111
13. Мелентьев П.Н., Букатин А.С. и др. Оптические методы детектирования единичных биомолекул: визуализация, сенсорики, секвенирование молекул ДНК // *Успехи физических наук*. 2024. Т. 194, № 11. С. 1130–1145. DOI: 10.3367/UfNr.2024.07.039720
14. Camera Dhyana 400BSI V3, Tucsen Photonics Co., Ltd. <https://www.tucsen.com/scmos-dhyana-400bsi-v3-product>

Финансирование

Работа выполнена в Институте аналитического приборостроения Российской академии наук (Санкт-Петербург) в рамках Федеральной научно-технической программы развития генетических технологий на 2019–2027 годы (Рег. N НИОКР 122012500127-1), соглашение № 075-15-2021-1057 от 28.09.2021.

Благодарности

Авторы выражают свою искреннюю благодарность ст. науч. сотр. ИАП РАН, канд. физ.-мат. наук Андрею Геннадьевичу Бородину, ст. науч. сотр. ИАП РАН, д-ру техн. наук Владимиру Владимировичу Манойлову и зам. дир. по научной работе ИАП РАН, канд. физ.-мат. наук Букатину Антону Сергеевичу за ценные замечания и участие в обсуждении темы данной работы.

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

Контакты: Васильева Екатерина Кимовна,
virtkat@mail.ru

Материал поступил в редакцию 29.11.2025

MODELING OF RAW DATA FROM A SINGLE-MOLECULE SEQUENCER. PART I: SIMULATING DETECTION OF EMISSION FROM A ZERO-MODE WAVEGUIDE

E. K. Vasil'eva, I. V. Chubinskiy-Nadezhdin

Institute for Analytical Instrumentation of RAS, Saint Petersburg, Russia

This work proposes a stochastic simulation model for the fluorescence detection process of labeled nucleotides during single-molecule sequencing across four spectral channels. A multi-factor detector signal simulation is implemented. The first part of the article describes an algorithmic template for modeling the response of a camera's photosensitive elements to incident light from a zero-mode waveguide within a single spectral range during DNA single-molecule sequencing. The concept of "base image" is introduced, representing the number of photoelectrons in sensors during an exposure time. The area simulation of such an image is a square field measuring 2×2 to 5×5 pixels, with total brightness corresponding to a single trace count in the given channel. Three modes of base image formation are considered: 1) dark current only, 2) background radiation in the absence of incorporation of a nucleotide, and 3) a fluorescent spot with a background level. A step-by-step algorithm for modeling the base image is provided. It includes: defining the size and offset of the spot, fluorescence and background amplitudes; generating two-dimensional Gaussian distributions for background and signal of fluorescence; and summing them; accounting for signal-dependent noise using experimental data; modeling signal-independent noise with a normal distribution; setting the zero-level offset; and finally summing all noise components to create the image. The algorithm for modeling background signal considering spectral crosstalk from the nearest signal registration channel is also described. A program simulator for primary digital data from a single-molecule sequencer has been developed, allowing for the simulation of sequencing results in FASTA, trace, and movie formats. The simulator accepts a real DNA sequence from a genomic database as input.

Keywords: sequencing simulator, long reads, simulation modeling, SMRT, zero-mode waveguide

REFERENCES

1. Mahmoud M., Huang Y., Garimella K. et al. Utility of long-read sequencing for All of Us. *Nat Commun.*, 2024, vol. 15, id. 837. DOI: 10.1038/s41467-024-44804-3
2. Rhie A., Nurk S., Cechova M. et al. The complete sequence of a human Y chromosome. *Nature*, 2023, vol. 621, pp. 344–354. DOI: 10.1038/s41586-023-06457-y
3. Stöcker B.K., Köster J., Rahmann S. SimLoRD: Simulation of Long Read Data. *Bioinformatics*, 2016, vol. 32, iss. 17, pp. 2704–2706. DOI: 10.1093/bioinformatics/btw286 //Epub 2016 May 10.
4. Wei Z.-G., Zhang Sh.-W. NPBS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. *BMC Bioinformatics*, 2018, vol. 19, id. 177. DOI: 10.1186/s12859-018-2208-0
5. Ono Y., Asai K., Hamada M. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics*, 2021, vol. 37, iss. 5, pp. 589–595. DOI: 10.1093/bioinformatics/btaa835
6. Zhang W., Jia B., Wei Ch. PaSS: a sequencing simulator for PacBio sequencing. *BMC Bioinformatics*, 2019, vol. 20, id. 352. DOI: 10.1186/s12859-019-2901-7
7. Rohrandt C., Kraft N., Gießelmann P., et al. Nanopore SimulatION – a raw data simulator for Nanopore Sequencing. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, pp. 1–8. URL: <https://ieeexplore.ieee.org/document/8621253>
8. Eid J., Fehr A., Gray J., Luong K., et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 2009, vol. 323, id. 5910. pp. 133–138 DOI: 10.1126/science.1162986
9. Korlach J., Bjornson K.P., Chaudhuri B.P., et al. Ch. 20 – Real-Time DNA Sequencing from Single Polymerase Molecules. *Methods in Enzymology*, Pacific Biosciences, Menlo Park, California, USA, 2010, vol. 472, pp. 431–455. DOI: 10.1016/S0076-6879(10)72001-2
10. Levene M.J., Korlach J., Turner S.W., et al. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science, New Series*, 2003, vol. 299, no. 5607, pp. 682–686. URL: <http://www.jstor.org/stable/3833436>
11. Bessmeltsev V.P., Terent'ev V.S., Vileiko V.V. et al. [Development of the platform for nucleic acid analysis (DNA sequencer)]. *Prikladnaya fotonika* [Applied photonics], 2016, vol. 3, iss. 4, pp. 388–412. DOI: 10.15593/2411-4367/2016.04.03 (In Russ.).
12. Chen J., Dalal R.V., Petrov A.N., et al. High-throughput platform for real-time monitoring of biological processes by multicolor single-molecule fluorescence. *PNAS*, 2014, vol. 111, no. 2, pp. 664–669. DOI: 10.1073/pnas.1315735111

13. Melentiev P.N., Bukatin A.S. et al. [Optical methods for detection of single biomolecules: visualization, sensorics, sequencing of DNA molecules]. *Uspekhi fizicheskikh nauk* [Advances in Physical Sciences], 2024, vol. 194, iss. 11, pp. 1130–1145. (In Russ.).
DOI: 10.3367/UFNr.2024.07.039720
14. *Camera Dhyana 400BSI V3, Tucsen Photonics Co., Ltd.*
<https://www.tucsen.com/scmos-dhyana-400bsi-v3-product>

Contacts: *Vasil'eva Ekaterina Kimovna*,
virtkat@mail.ru

Article received by the editorial office on 29.11.2025

INTRODUCTION

Single-molecule sequencing is a promising and rapidly developing method for obtaining genetic information in areas where the length of reads is crucial for data interpretation. The most complex tasks in metagenomics and epigenetics — *de novo* sequencing, analysis of large numbers of repetitive and homologous regions of the genome, and other studies related to the interpretation of long genomic sequences—cannot be effectively solved solely by NGS methods, which provide short reads. Developments by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) represent the two main third-generation sequencing technologies that are driving the development of single-molecule sequencing today. The first technology, called SMRT (Single Molecule Real Time sequencing), initially provided continuous longreads (CLR) with an error rate of 15% and was later developed into the Hi-Fi read method, which reduces the error rate to $< 1\%$. The second technology makes it possible to obtain ultra-long reads up to 4 Mbp, but with a high error rate, which can be reduced almost to the level of the Hi-Fi method using duplex sequencing technologies [1]. Both technologies served for the final stage of decoding the human genome—sequencing the Y chromosome, which has a complex repetitive structure [2].

Several long-sequence simulators that imitate data from third-generation sequencers are publicly available. These include SimLoRD, NPBSS, PBSIM, and PaSS [3–6]. They produce longread sequences in FASTA file format, simulating various read errors. This makes it possible to use this data in various bioinformatics tasks to debug secondary processing algorithms. These simulators provide a high degree of approximation to the results obtained during real sequencing. However, they do not produce raw data, i.e., movies and traces, for SMRT sequencers. The NanoporeSimulatION simulator differs from these, being specifically designed as a raw data simulator for ONT ion current signaling technology [7].

Raw data simulators for sequencers that record fluorescence signals are not publicly available. PacBio apparently uses its own simulation generators in its internal tests, which are not published publicly. This presents a challenge when it comes to simulating raw data from PacBio sequencers, for example, to develop and optimize custom algorithms for primary data processing, such as movie and trace, to optimize base-calling procedures.

Therefore, due to the limited and expensive nature of experimental data and the lack of simulators, simulation modeling and programming become essential for recognizing fluorescent objects in raw images and interpreting signals in single-molecule sequencing.

Developing own simulator is a complex and non-trivial task, as realistic simulation requires modeling polymerase kinetics, the characteristics of dyes used for nucleotide labeling, optical properties, detection system noise, and other parameters, most of which are probabilistic. Clearly, high-quality data simulation requires the development of a new modeling method. The most appropriate approach appears to be "virtual sequencing", which provides a multifactorial simulation of signals generated by the light-sensitive elements of a video camera, formed taking into account the parameters listed above.

SINGLE-MOLECULAR SEQUENCING DETECTION SYSTEM

A detection system for single-molecule sequencing using technology similar to PacBio [8–12] can be represented as a specialized four-channel, in terms of the number of spectral ranges, fluorescence microscope with digital recording of images of fluorescent objects of the zero-mode waveguides (ZMWs). Each spectral range corresponds to one of four fluorescently labeled nucleotides, traditionally designated A, T, G, and C. The ZMW is a nanostructure in which the sequencing reaction occurs. It is a cylindrical nanowell with a depth and width of approximately 100 nm, formed in a thin aluminum film, the volume of which does not exceed 10^{-19} L. In each ZMW, as the polymerase incorporates the next labeled nucleotide from the reaction buffer solution, fluorescence of the corresponding label is emitted by the excitation laser light incident on the transparent surface of the measuring cell from below. The cell contains a large number of ZMW and is positioned within the field of view of a high-resolution microlens for the purpose of simultaneously observing fluorescence in all nanostructures.

The light beams emerging from the lens carry information about the spatial distribution of the ZMW and the spectral composition of the emission. To separate the spectral ranges of the recorded fluorescence, the light beams pass through a dispersive prism [13]. When constructing an image, the emission from each ZMW is divided in the direction of the prism's dispersion according to the spectral composition of the fluorescence of the four emitting labels (Fig. 1).

Fig. 1. Correspondence between the light dispersion detection channels on the prism.
1 — laser-induced emission from the ZMW;
2 — dispersive prism; 3 — light dispersion on the video camera sensors; 4 — the ZMW image formed in four spectral channels. The spectral range of 550–750 nm is used

The light hits the photosensors of a video camera, where an image is generated over a specified exposure time. A single large-format, high-sensitivity CMOS video camera is used to display all the ZMWs [14]. On the camera's photo sensors, four adjacent groups of pixels can be distinguished for each ZMW, in which signals are generated as light spots in accordance with the intensity and spectral composition of the light incident on them.

The size and shape of the light spots will be determined primarily by the characteristics and resolution of the optical system used to construct an image of a submicron object. The image size in pixels depends on the prism dispersion, the selected magnification, and the physical size of the video camera sensors. The light spot images for different nucleotides partially overlap, corresponding to the overlap of the spectra of the fluorescent labels.

The image obtained for each frame is transferred to a computer. A series of such images forms a set of raw data (movies), which undergo further processing. During the initial processing, the computer localizes the images of all fluorescent nucleotides within the overall frame image and traces the fluorescence signals for each spectral channel, resulting in the formation of traces — data sets storing information about changes in fluorescence intensity over time. After preliminary filtering of traces based on selected quality criteria, basecalling — detection and pulse interpretation are performed in each spectral channel, ultimately generating a nucleotide sequence (Fig. 2).

Fig. 2. Generalized scheme for fluorescence detection in single-molecule sequencing using a dispersive prism.

MODELING BASE IMAGES AND SIGNALS FOR A SINGLE ZMW

Principle of base model image generation

In the first stage, an algorithmic and software template is generated for simulating the processes occurring on the photosensitive elements of a video camera when exposed to light in a single spectral range from a single ZMW. This results in a base image simulating the number of photoelectrons accumulated in the corresponding sensors of the video camera over a certain exposure time. A base image of a single ZMW is generated within a specified simulation region.

The simulation region of a light spot for a single spectral range is a square image, the size of which can be selectively set from 2×2 to 5×5 pixels. One pixel of this base image corresponds to one video camera sensor. The brightness level of one pixel corresponds

to the number of photoelectrons accumulated in the corresponding sensor during a single exposure. The sum of the brightnesses of all pixels in the simulation region forms an integral base signal, which becomes a single trace sample in one spectral channel (Fig. 3).

Fig. 3. Formation of a base image and a base ZMW signal in a single spectral channel.

1 — fluorescence-inducing laser; 2 — zero-mode waveguide (ZMW); 3 — video camera sensors; 4 — pixels of possible base image models; 5 — trace samples of the simulated spectral channel

In Figure 3:

- Laser 1 induces fluorescence in ZMW 2 with polymerase immobilized at the bottom, a DNA strand fragment, and labeled nucleotides in solution.

- During the process of embedding a fluorescently labeled nucleotide from the ZMW, laser radiation emits light, which, upon striking the sensors of the video camera 3, generates an equivalent electrical signal in photoelectrons in each.

- The signals accumulated in the camera sensors can be modeled as pixels of varying brightness in a simulation region of a given size 4, where the brightness level in a pixel is proportional to the number of photoelectrons in the corresponding camera sensor.

- The total signal of all pixels in the base image forms an integrated base signal, corresponding to one sample in trace 5 in the simulated spectral channel.

As an example of a real trace, a 30-second fragment is shown in Fig. 4. The data were obtained by recording fluorescence in channel G for a single streptavidin molecule immobilized on the bottom of the ZMW.

Fig. 4. Trace fragment

The streptavidin molecule is labeled with Cy5 dye. Nucleotides labeled with the same dye, floating freely in a buffer solution, are added to the ZMW. At 6.5 s, the red laser was turned on, and dye fading on the streptavidin was recorded. After 15 seconds, the labeled nucleotides were added. The integrated ZMW fluorescence intensity signal was determined over a 3×3 pixel image area. The background fluorescence intensity of the cell after the laser was turned on is approximately 1200 photoelectrons; the fluorescence signals in the ZMW (baseline signals, or trace samples) exceed 200–400 photoelectrons in amplitude.

Simulation of background, useful signal and dark noise

If we consider the baseline signal in terms of whether nucleotide incorporation has occurred or not, the baseline image can be formed as follows:

1. Simulation of the presence of dark current in the video camera sensors in the case where no light is emitted from the ZMW (case "a").
2. Simulation of the background radiation signal for the case where nucleotide incorporation in a given spectral range has not occurred during the current exposure (case "b").
3. Simulation of the light spot from a fluorescent flash, including the background signal, to simulate the nucleotide incorporation process (case "c").

These cases are shown in Fig. 5.

Fig. 5. Baseline images of a 5×5 pixel simulation region (a) and their brightness profiles (b).
 6 (top) — shows baseline signal samples when simulating dark chamber noise (case "a").
 6 (bottom) — shows baseline signal samples when simulating cell background (case "b") and nucleotide incorporation events (case "c").
 During the simulation, two incorporation events occurred. Each event consists of five samples with an intensity exceeding 100 photoelectrons (ф.э.л.). The background signal counts are formed by baseline signals of lower intensity

Spectral crosstalk

When synthesizing a complementary DNA strand, two different nucleotides cannot be simultaneously incorporated into the primary strand, so two fluorophores cannot glow simultaneously in one ZMW. However, when a single nucleotide fluoresces, spectral crosstalk can enter adjacent detection channels. In the model signal, the magnitude of this crosstalk can be specified as a percentage of the fluorescence signal intensity in the channel that is the source of the crosstalk. This crosstalk is added to the integral brightness signal in the modeled channel and depends on the shape of the excitation and emission spectra of the dyes used (Fig. 6).

Fig. 6. Excitation and emission spectra of fluorescent dyes used for nucleotide labeling. The shaded regions show spectral overlap between channels A and T, T and G, and G and C

We denote interference from channel A to channel T as $CrTk^{AT}$, and, conversely, interference from chan-

nel T to channel A as $CrTk^{TA}$. The notation logic for other spectral channels is similar.

To simulate the beam splitting from four fluorescent dyes using a dispersive prism, four base images should be arranged vertically in order of increasing emission wavelength: A (green), T (yellow), G (red), and C (dark red). Each of the four base images is generated separately in accordance with the parameters specified for a given spectral channel.

Sbasic base image simulation algorithm

Simulation algorithm (option 1) for the base image in the simulation region for channel A (SbasicA) in the case of embedding labeled nucleotide A (Fig. 7):

Fig. 7. Generating the basic image Sbasic for each of the four spectral channels. In channel A, the useful signal (the presence of fluorescence) is simulated. The numbers indicate the corresponding stages of the simulation algorithm.

1. Specify the size of the simulation region in pixels and the offset of the simulated light spot within the region.
2. Specify the amplitude of the fluorescence signal A in photoelectrons — Fl^A .
3. Set the $BkGrd^A$ background signal amplitude, i.e., the magnitude of the intrinsic luminescence signal of the measuring cell in the absence of a fluorescence signal from the embedded nucleotide in the simulated channel.
4. Generate two-dimensional images of the Gaussian brightness distribution in the simulation region for the background and fluorescence according to the specified parameters. Sum both images.
5. Simulate noise, which depends on the signal magnitude accumulated in the sensors over the exposure time, as the sum of PRNU (Photo Response Non-Uniformity) and shotgun noise (N_{dp}). To do this, calculate the noise value in each pixel for the combined image obtained in step 4*.
6. Model the noise, independent of the signal accumulated in the sensors, as the sum of the dark signal non-uniformity (DSNU) noise and the camera readout noise (ReadOut) — N_{undp} . To do this, calculate the noise value in each pixel using a normal distribution with a mean equal to zero and the standard deviation specified in the camera specifications for these types of noise.

* By approximating the experimentally obtained dependence of the standard deviation of the integral base signal, measured in a given square area of $M \times M$ pixels, on the magnitude of this signal.

7. Set the signal intensity offset relative to zero for each pixel** — *Offset*.

8. Add the calculated noise and offset to the background and fluorescence image obtained in step 4.

Algorithm for modeling (option 2) the base image in the simulation region for channel T (SbasicT) (background + spectral crosstalk from A) in the case of embedding labeled nucleotide A (Fig. 7):

1. Specify the size of the simulation region in pixels and the offset of the simulated light spot within the region.

2. Specify the spectral crosstalk in channel T as a percentage of the signal amplitude from channel A that creates this crosstalk — $Fl^A \times CrTk^{AT}$.

3. Specify the background signal amplitude $BkGrd^T$ — the magnitude of the intrinsic fluorescence signal of the measuring cell in the absence of a fluorescence signal from the embedded nucleotide in the simulated channel.

4. Generate two-dimensional images of the Gaussian luminance distribution in the simulation region for the background and fluorescence according to the specified parameters. In this case, the intrinsic fluorescence in channel T is zero, but a parasitic fluorescence signal from the illumination from channel A is generated. Sum both images.

5. Model the noise, which depends on the signal accumulated in the sensors during the exposure time, as the sum of PRNU (Photo Response Non-Uniformity) and shotgun noise (N_{dp}). To do this, calculate the noise value in each pixel for the summary image obtained in step 4.

6. Model the noise, which does not depend on the signal accumulated in the sensors, as the sum of DSNU (dark signal non-uniformity) noise and ReadOut (camera readout noise — N_{undp}). To do this, calculate the noise level in each pixel using a normal distribution with a mean of zero and the standard deviation specified in the video camera's specifications for these types of noise.

7. Set the signal intensity offset relative to zero for each pixel — *Offset*.

8. Add the calculated noise and offset to the background and fluorescence image obtained in step 4.

If it is necessary to model the dark current in the sensors, only steps 6 and 7 of the algorithm are used.

Explanation of step 4 of the algorithm. The total light flux generated by both the nucleotide fluorescence signal (if incorporation has occurred) and the general background radiation of the ZMW, caused by

the glow of the cell surface and the weak signal from the labeled nucleotides floating in solution, enters the video camera from the ZMW. In both cases, the image of the light spot in the simulation region can be modeled by a two-dimensional Gaussian brightness distribution (1) with a given $\sigma(\sigma_x = \sigma_y = \sigma)$, which corresponds to the radius of the luminous object and a random offset of the spot center m_x, m_y . The optical signal itself corresponds to the magnitude of the two-dimensional Gaussian integral (Fig. 8)

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(x-m_x)^2}{2\sigma_x^2} - \frac{(y-m_y)^2}{2\sigma_y^2}}. \quad (1)$$

Fig. 8. Gaussian model of the image of the fluorescence signal of the ZMW

The offset of the spot's symmetry center is optionally specified by random variables uniformly distributed within the simulation region and allows for the manifestation of spatial instability of the image to be taken into account.

The overall brightness of all image pixels in the simulation region is determined by the specified value of the base signal Fl (e.g., only for the fluorescence signal). The base signal is generated by the software as the sum of the intensities of all pixels in the base image. To do this, each pixel in the simulation region receives a brightness $A_{i,j}$ with a corresponding weighting coefficient, as shown in (2):

$$A_{i,j} = \frac{Fl}{2\pi\sigma^2} e^{-\frac{(i-m_x)^2}{2\sigma^2} - \frac{(j-m_y)^2}{2\sigma^2}}. \quad (2)$$

As a result, the algorithm for generating a model two-dimensional base image S_{bsc} in the simulation region can be described by the following formulas.

- We simulate the presence of fluorescence (example for channel A):

$$S_{bsc}^{A-Fl} = (Fl^A * f(x, y) + BkGrd^A * f(x, y)) + N_{dp} + N_{undp} + Offset. \quad (3)$$

- Modeling only the background (example for channel T):

$$S_{bsc}^{T-Bg} = (Fl^A * CrTk^{AT} * f(x, y) + BkGrd^T * f(x, y)) + N_{dp} + N_{undp} + Offset. \quad (4)$$

** The *Offset* parameter of the video camera, which specifies the zero-level offset, in this case is equal to 100 brightness units per pixel.

- Modeling a dark framesubstraction:

$$S_{\text{bsc}}^{\text{Drk}} = N_{\text{undp}} + \text{Offset}. \quad (5)$$

Justification of the noise modeling method

Explanation of point 5 of the algorithm. An experimental study of the dependence of video camera noise on the magnitude of the recorded signal allowed us to determine the standard deviation of the luminance signal of the measurement cell image, integrated into a 3×3 pixel area. Figure 9 shows the dependence of the relative standard deviation (defined as a percentage relative to the corresponding signal magnitude) of this luminance signal for an empty cell illuminated with green and red lasers. By "good" pixels we mean the majority of camera pixels for which the measured luminance signal spread is close to the value specified in the video camera's specifications. By "bad" pixels we mean some pixels with a signal spread exceeding the specifications by 5 times or more. The inclusion of such pixels in the luminance signal integration region during trace construction can increase the standard deviation of background noise and complicate subsequent signal-eventsearches***.

Fig. 9.Dependence of the relative standard deviation of the brightness signal (intensity) of the background image of a cell in the 3×3 pixel integration area on the magnitude of this signal minus the *Offset* level shift.

The data for plotting the dependence were obtained by processing a series of frames recorded in darkfield mode (in the absence of an external light source) and with the lasers turned on, with successively increasing exposure times from 10 ms to 250 ms.

This allowed us to measure the brightness levels of the cell's intrinsic luminescence image over a wide range of values. To correctly process the data, an offset of 100 photoelectrons per pixel, artificially added in the video camera to correct the zero level, was subtracted from the integral signal value.

By approximating the measurement results, we obtained a functional dependence of the standard deviation of the integral signal N on the intensity level of this integral signal S in a 3×3 pixel region:

$$N = 128.15 \times S^{-0.473}. \quad (6)$$

*** An *signal-eventsearches* is an event that is the emission of light in one of the spectral channels during the complementary incorporation of a fluorescently labeled nucleotide into a DNA strand. It consists of a set of *base signals*, which are the integrated brightness signals for the ZMW-image, obtained in a given region of the frame during a single exposure. On a trace, itappearsas a pulse.

The resulting expression is used in modeling the base image using formulas (3) and (4) to determine the noise level dependent on the N_{dp} signal level. To do this, replace the value of S in (6) with the sum of the background and fluorescence signals for the simulated pixel in the simulation region.

RESULTS

Thus, in the simulated base image, we will be dealing with an asymmetric spot defined by a two-dimensional Gaussian distribution of background and fluorescence signals, with the addition of spectral noise and brightness fluctuations at each pixel. The integrated brightness of the base image — the sum of the brightnesses in pixels within the simulation region — will also fluctuate across different simulations, even with identical simulation parameters, thus creating a stochastic simulation model of the radiation detection process from the ZMW.

By using this algorithm for generating base signals and images repeatedly, it is possible to simulate fluorescence signals that exceed the exposure time and correspond to nucleotide incorporation events into the DNA strand in all four spectral channels. The presented method ensures realistic generation of base images for use in subsequent full-frame simulations of a series of sequencer raw image frames, which can be generated for testing raw data processing algorithms and improving the accuracy of fluorescence event recognition.

Figure 10 shows software simulations of the baseline ZMW images in four spectral channels for different simulation region sizes with different input parameters, including spectral noise for some images.

Fig. 10.Examples of basic nucleotide images.
See text for explanations

In Figure 10:

- [1] 5×5 pixel region; a — fluorescence in channel A; b — fluorescence in channel A with spectral overlap of channel T; c — fluorescence in channel G with spectral overlap of channel C.
- [2] 4×4 pixel region; a — fluorescence in channel T; b — fluorescence in channel A.
- [3] 3×3 pixel region; a — fluorescence in channel A; b — fluorescence in channel G.
- [4] 2×2 pixel region; a — fluorescence in channel G; b — fluorescence in channel C.