

УДК 519.856, 004.942, 519.876.5, 57.081.23

EDN: MRCJZL

© Е. К. Васильева, И. В. Чубинский-Надеждин, 2026

МОДЕЛИРОВАНИЕ ПЕРВИЧНЫХ ДАННЫХ ОДНОМОЛЕКУЛЯРНОГО СЕКВЕНАТОРА. Ч. 2. СИМУЛЯЦИЯ ПОСЛЕДОВАТЕЛЬНОСТИ СОБЫТИЙ ВСТРАИВАНИЯ МЕЧЕНЫХ НУКЛЕОТИДОВ В ЦЕПЬ ДНК И ПОЛНОФОРМАТНОЕ МОДЕЛИРОВАНИЕ ИЗОБРАЖЕНИЙ МНОЖЕСТВА ВОЛНОВОДОВ НУЛЕВОЙ МОДЫ

Рассмотрены особенности детектирования сигналов флуоресценции меченых нуклеотида в волноводе нулевой моды. Введено понятие "сигнал-событие" — для единичного акта излучения в одном из спектральных каналов при комплементарном встраивании меченого нуклеотида. Учтены длительность свечения метки, особенности формирования последовательности сигналов-событий и фоновых интервалов между ними. Моменты наступления сигналов моделируются как стационарный одномерный пуассоновский поток, описываемый свойствами бесконечного счетного множества событий, стационарности, ординарности и отсутствия последствия. Отмечена необходимость применения нестационарных моделей или комбинаций процессов при учете эпигенетических эффектов (например, метилирования). Результатом работы стала программная реализация симулятора первичных данных для одномолекулярного секвенатора третьего поколения с генерацией выходных данных в форматах FASTA, MOVIE и TRACE с возможностью формирования серий изображений, имитирующих изображения массива волноводов нулевой моды в измерительной ячейке.

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

ВВЕДЕНИЕ

Вторая часть статьи завершает описание имитационной модели процесса детектирования флуоресценции в волноводах нулевой моды одномолекулярного секвенатора. Секвенирование одиночных, не объединенных в кластеры молекул ДНК, в реальном времени — метод третьего поколения, нацеленный на получение рядов, длительность которых составляет тысячи пар нуклеотидов. Получение таких рядов с известными характеристиками может быть осуществлено посредством симуляции длинных последовательностей событий встраивания нуклеотидов в ДНК. Такая симуляция для метода SMRT (Single Molecule Real Time sequencing) [1–3] может иметь важное утилитарное значение в разных задачах — от разработки и тестирования алгоритмов вторичной обработки данных до оптимизации химических процессов и повышения точности секвенирования.

Для разработки надежных алгоритмов base-calling, часто основанных на машинном обучении, требуются огромные массивы данных с известной истинной последовательностью. Моделирование

позволяет генерировать синтетические данные в любом объеме, точно контролируя последовательность и параметры сигнала, что невозможно при реальных экспериментах. Тестирование устойчивости алгоритмов вторичной обработки — важная задача, для которой реалистичное имитационное моделирование позволяет проверить, как алгоритмы base-calling реагируют на различные шумы, артефакты, изменения скорости полимеразы, пропуски или удлинение интервалов между событиями встраивания, что имитирует реальные биологические или инструментальные сбои.

Метод SMRT позволяет не просто считывать последовательность нуклеотидов в цепи ДНК, он показывает кинетику работы полимеразы. Моделирование позволяет точно воспроизвести распределение длительностей интервалов между событиями встраивания и определить оптимальное время экспозиции видеокамеры, чтобы максимально точно захватить короткие флуоресцентные импульсы и при этом минимизировать размытие сигнала или потерю данных.

Моделирование последовательности событий является лишь одним из этапов в полном симуляторе SMRT. Важнейшей частью полной симуляции данных является интеграция двух этапов — симуляции буквенной последовательности нуклеотидов и симуляции базовых изображений событий излучения флуоресценции, где последовательность событий является входным сигналом для этапа моделирования оптического отклика. Моделирование десятков тысяч волноводов нулевой моды (ВНМ) требует эффективного и точного генератора событий для каждого волновода, чтобы имитировать работу всей ячейки секвенирования.

СИМУЛЯЦИЯ ПОСЛЕДОВАТЕЛЬНОСТИ СОБЫТИЙ ВСТРАИВАНИЯ МЕЧЕНЫХ НУКЛЕОТИДОВ В ЦЕПЬ ДНК

Параметры для моделирования первичных данных

Назовем термином *сигнал-событие* событие, которое состоит в излучении света в одном из спектральных каналов в процессе комплемен-

тарного встраивания флуоресцентно меченного нуклеотида в цепь ДНК. Формирование из отсчетов базовых сигналов последовательности сигналов-событий, согласованной с заданной последовательностью буквенных обозначений нуклеотидов, составляет суть второго этапа моделирования. Длительность каждого такого события должна соответствовать длительности свечения флуоресцентной метки в ходе встраивания меченого нуклеотида.

Все параметры, которые задействованы при полной симуляции первичных данных одномолекулярного секвенатора, можно разделить на 4 группы в соответствии с задачей, которая решается на каждом этапе формирования данных (рис. 1, табл.).

На 1-м этапе происходит ввод исходных данных в алгоритм симуляции — считывание буквенной последовательности нуклеотидов. Буквенная последовательность в формате FASTA [4, 5] может быть получена в виде файла из геномной базы данных или сгенерирована случайным образом непосредственно в программе-симуляторе.

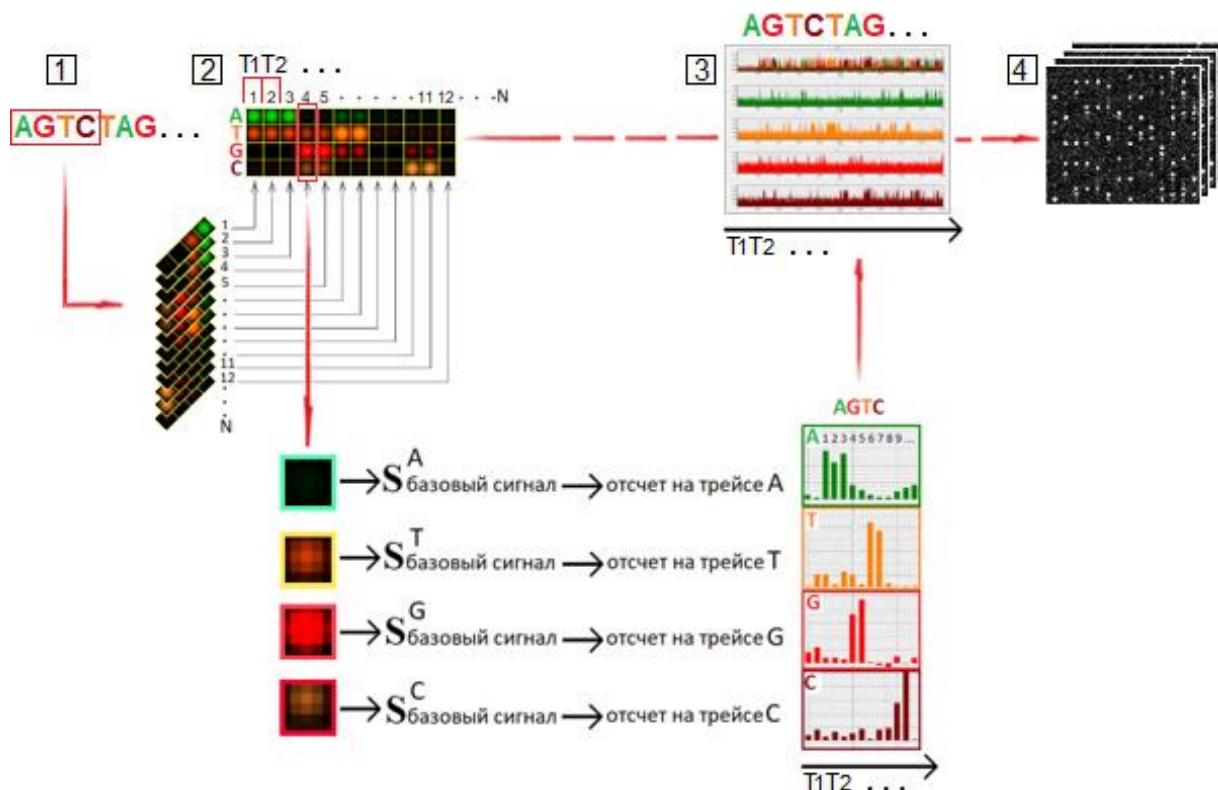


Рис. 1. Этапы симуляции первичных данных.

[1] — формирование или чтение буквенной последовательности нуклеотидов; [2] — формирование базовых изображений и базовых сигналов (отсчетов) для одного ВНМ; [3] — формирование последовательности сигналов-событий в четырех спектральных каналах для одного ВНМ; [4] — формирование серии изображений кадров с симуляцией данных для множества ВНМ

Табл. Параметры алгоритма симуляции на разных этапах формирования данных

1-й этап формирование последовательности нуклеотидов	2-й этап формирование базовых изображений и базовых сигналов	3-й этап формирование последовательности сигналов-событий	4-й этап формирование серии полно- форматных изо- бражений
Группа 1 параметров	Группа 2 параметров	Группа 3 параметров	Группа 4 параметров
Количество и поряд- ок букв в после- довательности нук- леотидов	Размер области симуля- ции для одного спек- трального канала в пик- селях	Время экспозиции в мс	Размер изобра- жения кадра в пикселях
	Размер пятна яркости и смещение пятна внут- ри области симуляции в пикселях	Средняя длительность сиг- нала-события и СКО дли- тельности (параметр PW — pulsewidth) в мс	Расстояние меж- ду ВНМ по X и Y в пикселях
	Интенсивность фоново- го сигнала для каждого канала в фотоэлектро- нах (4 параметра, по числу каналов регистра- ции)	Закон распределения дли- тельности PW (нормаль- ный, логнормальный, кон- станта)	Адрес папки с файлами movie для множества ВНМ
	Интенсивность флуо- ресценции сигнала- события для каждого канала в фотоэлектро- нах (4 параметра, по числу каналов регистра- ции)	Средняя частота появления сигнала-события в секунду — параметр A в распреде- лении Пуассона: $P(m) = \frac{A^m}{m!} e^{-A}$	
	Перекрестные флуорес- центные помехи – cross- talk в % от величины интенсивности флуо- ресценции (6 парамет- ров)	Интервал времени без ге- нерации полезных сигна- лов в секундах (участок с фоновым сигналом)	
	Шумы, зависящие от уровня сигнала (имита- ция PRNU и дробового шумов)		
	Шумы, не зависящие от уровня сигнала (имита- ция DSNU и ReadOut шумов)		
Смещение уровня сиг- нала в пикселе (offset в камере)			

Параметры из группы 2 необходимы для форми-
рования базовых изображений и интегральных базо-

вых сигналов в области симуляции для четырех
спектральных каналов в соответствии с описанным

в первой части статьи алгоритмом [6]. Одновременно с формированием этих данных и на их основе выстраивается общая последовательность сигналов-событий, генерируются и сохраняются трейсы и файлы изображений в формате movie для каждого ВНМ. При этом используется 3-я группа параметров, регламентирующая характер появления сигналов-событий и их длительность.

На последнем этапе с использованием параметров из группы 4 осуществляется сборка в кадр изображений отдельных ВНМ, данные для которых были предварительно сгенерированы и сохранены на диск, и затем сборка в серию множества кадров полноформатных изображений.

Последовательность сигналов-событий как пуассоновский случайный процесс

Высвечивание флуорофора метки начинается в момент фиксации нуклеотида полимеразой в зоне действия лазерного излучения и длится до тех пор, пока метка удерживается в этой области и не теряет способность флуоресцировать. По разным оценкам это занимает время от единиц до сотен мс, что, как правило, превосходит время экспозиции одного кадра. В течение этого времени система детектирования будет регистрировать сигнал флуоресценции. После выхода метки из зоны удержания полимеразой видеокамера будет фиксировать фоновый сигнал для ВНМ до тех пор, пока к полимеразе не подойдет новый меченый нуклеотид. Таким образом, в процессе секвенирования происходит последовательная регистрация относительно ярких сигналов-событий (pulse) и менее ярких фоновых сигналов, излучаемых ВНМ между событиями (interpulse). Скорость процесса встраивания, т.е. частота появления сигналов-событий может меняться в разных экспериментах.

Пусть моментом наступления случайного сигнала-события будет являться начало высвечивания флуоресценции в реакции встраивания нуклеотида, тогда последовательность таких моментов образует поток стартов реакций. Он, в свою очередь, порождается пуассоновским потоком событий захвата нуклеотидов полимеразой. В рассматриваемой модели он стационарный.

Согласно [7, 8], поток случайных однородных событий представляет собой одномерный стационарный пуассоновский поток и может быть описан процессами, подчиняющимися условиям стационарности, ординарности и отсутствия последствия. Для описания и имитации эпигенетических процессов, таких как метилирование ДНК [9, 10], потребуется использовать нестационарный пуассоновский поток либо комбинацию двух случайных процессов с разной частотой реализации событий, поскольку длительности интервалов IPD

(interpulse duration) между событиями встраивания обычных и метилированных нуклеотидов существенно отличаются [11, 12].

В работах [8, 13] показано, что распределение, определяющее периоды времени между следующими друг за другом однородными событиями в пуассоновском потоке, т.е. предсказывающее длины временных интервалов между ними, может быть описано экспоненциальной функцией. В [14] отмечено, что такой подход применим к моделированию интервалов времени между сигналами флуоресценции в методе SMRT.

При программном моделировании длин временных интервалов между событиями одномерного пуассоновского потока для получения последовательности случайных чисел с заданным законом распределения может быть использовано соотношение, связывающее случайные числа с заданным законом и числа, равномерно распределенные в интервале от 0 до 1 [7–9, 15]:

$$t_i = -\frac{1}{A} \ln(1 - R_i), \quad (1)$$

где t_i — длина i -го временного интервала, A — средняя интенсивность потока реализации событий, R_i — случайное значение, полученное в интервале от 0 до 1 при помощи генератора случайных чисел.

Таким образом, программно формируется массив начальных моментов времени для интервалов $t_1, t_2, \dots, t_i$, соответствующих моментам старта реакций встраивания нуклеотидов. Одновременно с этим формируется или считывается из файла массив буквенных обозначений нуклеотидов. После чего, моделируя случайные длительности сигнала-события $PW_1, PW_2, \dots, PW_i$ (pulsewidth) для каждого нуклеотида из указанного массива букв, можно ассоциировать их с полученными моментами старта реакций (рис. 2). Сопоставляя обозначение нуклеотида, момент старта реакции и длительность реакции, моделируем базовые изображения и сигналы при помощи описанных в первой части статьи алгоритмических шаблонов с индивидуальным набором параметров для каждого нуклеотида. При этом формируем пять последовательностей — одну общую и четыре для букв А, Т, Г и С, соответствующие каждой своему нуклеотиду.

Длительности сигналов-событий могут быть заданы постоянными или вариативными при помощи нормального или логнормального распределений. Нуклеотиды не могут встраиваться одновременно, но могут встраиваться один за другим так, что в этом случае в пределах длительности одной экспозиции находится окончание одной реакции и начало следующей. В случае, если длительность события PW_i , начавшегося в момент времени t_i , перекрывает момент старта следующего

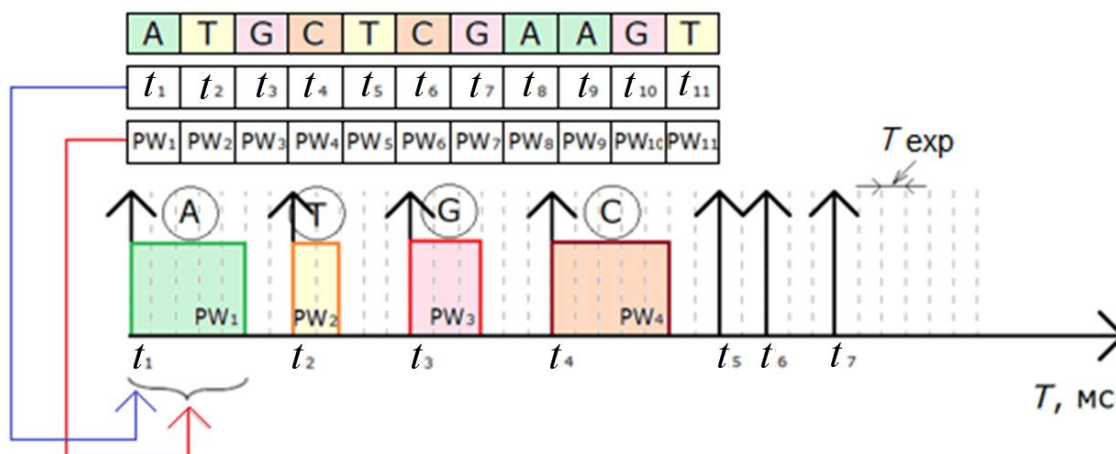


Рис. 2. Формирование последовательности сигналов-событий

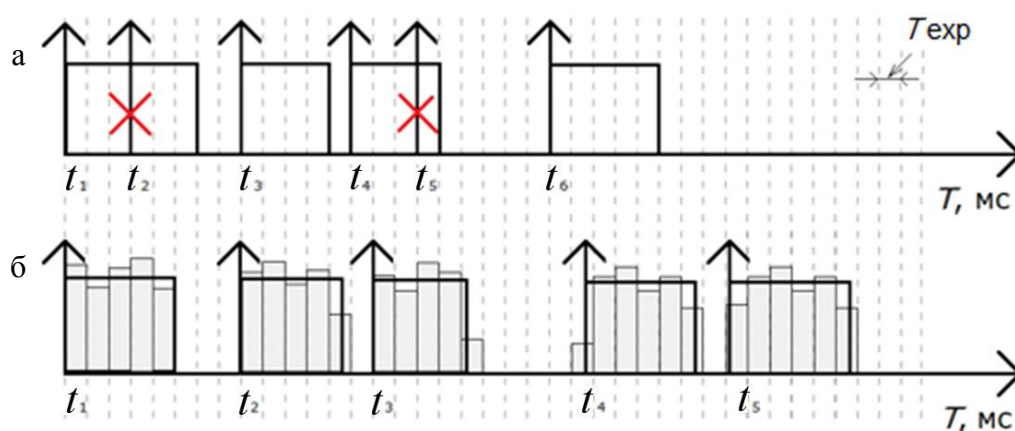


Рис. 3. Особенности формирования последовательности сигналов-событий

события $t_{(i+1)}$, момент $t_{(i+1)}$ должен быть пропущен и использован следующий за ним момент $t_{(i+2)}$ (рис. 3, а). Если и этот момент перекрывается длительностью события PW_i , то должен быть выбран следующий момент старта $t_{(i+3)}$ и т.д. Относительно большая длительность сигналов-событий либо высокая частота появления событий в секунду, например, 5 событий в секунду со средней длительностью 200 мс, автоматически приведет к удлинению общего времени реализации потока за счет уменьшения частоты появления событий, при этом пуассоновский характер их появления сохранится.

Момент старта либо завершения события встраивания попадает случайным образом внутрь очередного временного интервала дискретизации, поэтому почти всегда длительность сигнала-

события не может быть дискретизирована целым числом интервалов, равных времени экспозиции (рис. 3, б). Для корректного моделирования принимаем, что в случае, если событие встраивания началось в первой половине очередного интервала экспозиции или закончилось во второй половине, как события на рис. 3 t_2 и t_4 (задний фронт) и t_5 (передний и задний фронты), то заданная интенсивность флуоресценции базового изображения и сигнала умножаются на коэффициент 0.6. Другими словами, присваиваем базовому сигналу $2/3$ интенсивности от заданной при моделировании, поскольку большая часть рассматриваемого интервала экспозиции "перекрывается" моделируемым импульсом. Если же событие встраивания началось во второй половине очередного интервала экспозиции или закончилось в первой, как

события на рис. 3 t_3 (задний фронт), t_4 (передний фронт), то заданная интенсивность флуоресценции базового изображения и сигнала умножаются на коэффициент 0.3. То есть, если интервал длительности сигнала-события не кратен времени экспозиции, то первый и/или последний отсчет сигнала-события, длящийся, очевидно, меньше времени экспозиции, уменьшается по амплитуде. Это эквивалентно случаю, когда при секвенировании свечение флуоресцентной метки заняло не полное время экспозиции, а больше или меньше половины этого времени, что привело к ослаблению общего накопленного сигнала флуоресценции за время экспозиции.

Из базовых сигналов складывается зависимость интенсивности свечения меток в одном спектральном канале от времени. На графике такой зависимости каждому интервалу дискретизации, равному времени экспозиции, можно поставить в соответствие интегральный сигнал, имитирующий сигнал, накопленный на видеокамере в течение этого времени экспозиции, формируя таким образом трейс.

Генерируя базовые изображения и вычисляя базовые сигналы для четырех спектральных каналов, постепенно формируем всю последовательность сигналов-событий.

По трейсам, заданным в цифровом виде для четырех спектральных каналов детекции, впоследствии может осуществляться base-calling при обработке задач распознавания последовательности нуклеотидов [16, 17].

Таким образом, последовательность сигналов-событий может быть сформирована, как череда единичных отсчетов — базовых сигналов, непрерывно следующих друг за другом с интервалом дискретизации, равным времени экспозиции. Сама последовательность таких событий представляет собой реализации пуассоновского потока событий, а вариации длительностей временных интервалов между этими сигналами могут быть описаны, как реализации экспоненциального закона распределения.

Логика формирования последовательности показана на рис. 4.

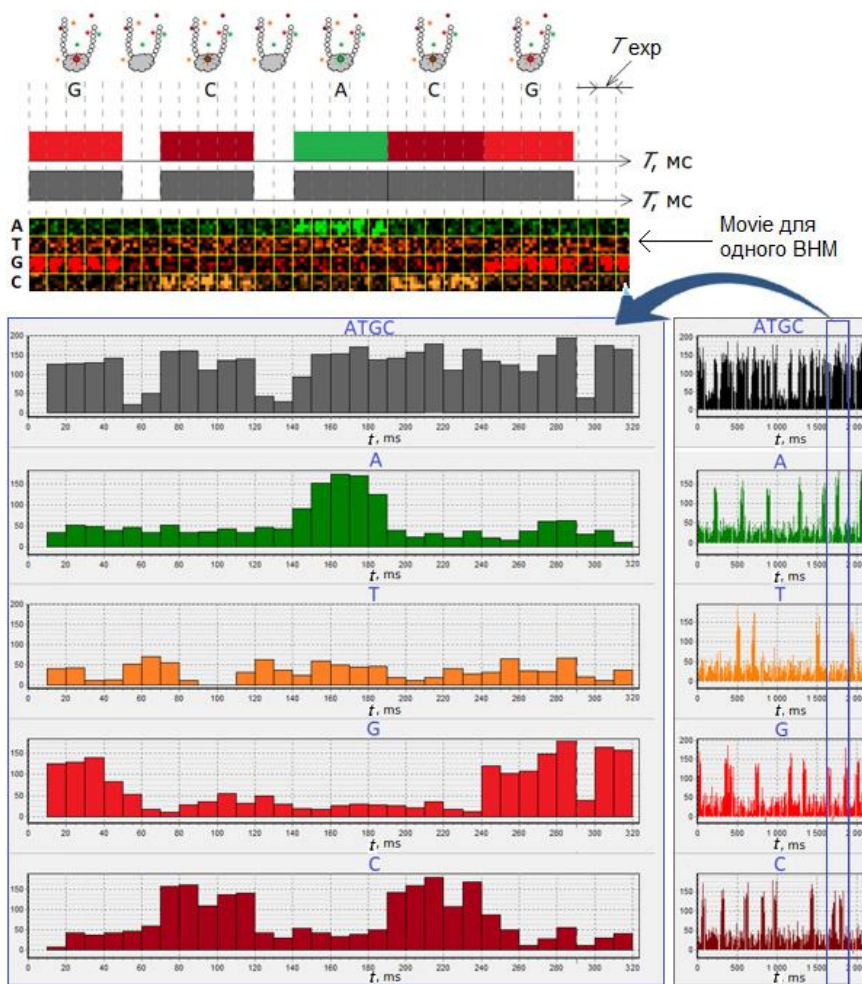


Рис. 4. Пример формирования трейсов и movie для короткой последовательности GCACG

Флуоресцентные метки вспыхивают в захваченных полимеразой молекулярных комплексах. Последовательным вспыхкам разных флуорохромов — в данном случае это красный, темно-красный, зеленый, темно-красный, красный — соответствуют по длительности сигналы-события, показанные четырьмя такими же цветами. Серым цветом показана общая последовательность импульсов на временной оси. На общей последовательности видно, что последние три сигнала-события слились воедино и не могут быть легко разделены без информации о спектральном канале, в котором они были зарегистрированы. При этом мы видим, что это три разных по спектральному составу оптических сигнала. (Цветовую аргументацию можно оценить только в цветном формате электронной версии статьи.)

С дискретизацией, равной времени экспозиции, формируются базовые изображения для всех спектральных каналов. Количество базовых изображений для каждого сигнала-события соответствует тому, сколько раз время экспозиции укладывается во временной интервал высвечивания соответствующего флуорохрома. Совокупность таких изображений для одного ВНМ представляет собой данные типа movie и сохраняется на диск в виде графического файла. Соответствующие базовые сигналы формируют трейсы для каждого канала и общую последовательность сигналов-событий. Трейсы для всех четырех каналов регистрации сохраняются на диск в виде одного общего для ВНМ файла.

Симуляция серии кадров

Задача симуляции набора первичных данных в виде полноформатных кадров movie размером 2048 на 2048 пикселей решается в два этапа. На первом этапе формируются базовые изображения последовательностей сигналов-событий для всех ВНМ из заданного для моделирования множества. Эти данные сохраняются на диск в виде файлов изображений в формате movie и генерируются для каждого ВНМ индивидуально по заданным параметрам при помощи описанных ранее алгоритмов.

На втором этапе из этих, сгенерированных заранее файлов, осуществляется "сборка" кадров. Для этого в каждом файле movie для одного ВНМ выбирается фрагмент изображения, соответствующий симуляции данных для очередного i -го момента времени регистрации (рис. 5), который должен быть перенесен на изображение i -го кадра в полной симуляции.

Последовательно копируя фрагменты изображений для разных одиночных ВНМ в заранее рассчитанные координаты на полноформатном изо-

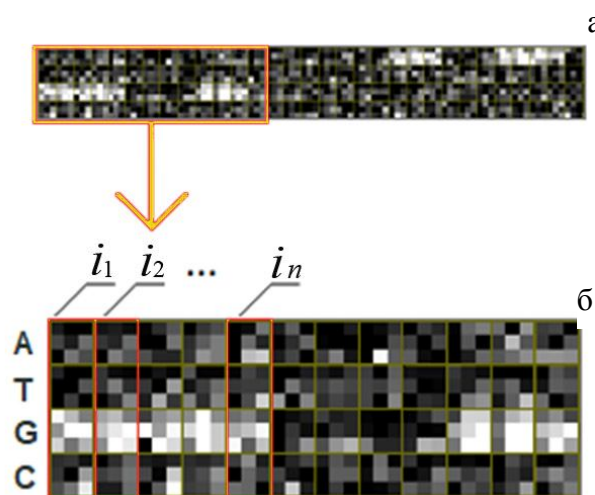


Рис. 5. Данные типа movie для одного ВНМ.

На (а) выделен фрагмент изображения с двумя событиями встраивания в канале G (пиксели с ярким свечением). На (б) показан этот же фрагмент в крупном масштабе с выделенными вертикальными участками, симулирующими флуоресцентные отклики, сформированные в разные моменты регистрации в четырех спектральных каналах. Первое событие встраивания дискретизируется на 5 временных интервалов ($n = 5$), равных времени экспозиции при регистрации; второе событие встраивания дискретизируется на 3 интервала ($n = 3$); между ними 4 фрагмента с фоновым сигналом

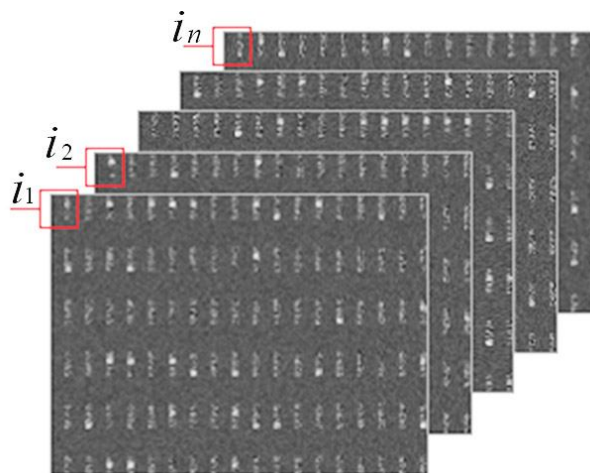


Рис. 6. Данные типа movie для множества кадров с разными ВНМ в моделируемой серии данных

бражении (рис. 6), можно сформировать изображение любого кадра со множеством ВНМ, имеющих в рассматриваемый момент времени флуоресцентные отклики в разных каналах.

Серия из большого количества таких кадров будет имитировать реальные сырые изображения одномолекулярного секвенатора, полученные в течение длительного времени, где фрагменты $i_1, i_2, \dots, i_n$ изображений movie для каждого ВНМ переносятся в соответствующие места на кадрах $i_1, i_2, \dots, i_n$. Таким образом, количество интервалов экспозиции будет точно соответствовать числу кадров в симулируемом наборе данных. Эти кадры являются имитацией самых первичных данных, которые при реальном секвенировании формируются на выходе видеокамеры детектора. К первичным данным относятся также и трейсы, но они в реальном процессе детектирования могут быть получены только после обработки данных типа movie.

ЗАКЛЮЧЕНИЕ

В работе предложена стохастическая имитационная модель процесса детектирования флуоресценции меченых нуклеотидов при одномолекулярном секвенировании в четырех спектральных каналах. Обеспечена многофакторная имитация сигналов детектора. Разработана программасимулятор первичных цифровых данных одномолекулярного секвенатора, обеспечивающая имитацию результатов секвенирования в форматах FASTA, trace и movie.

Точная программная симуляция позволяет оценить, как отношение сигнал/шум и частота событий встраивания (и следовательно, длительность межимпульсных интервалов) влияют на эффективность base-calling и общую точность чтения последовательности.

Моделирование оптических откликов и последовательности событий встраивания нуклеотидов в цепь ДНК является важными составляющими реалистичного симулятора первичных данных одномолекулярного секвенатора. Модель позволяет генерировать синтетические данные для обучения и валидации алгоритмов вторичной обработки, изучать кинетические эффекты полимеразы и оптимизировать условия проведения секвенирования. Это, в конечном итоге, может облегчить расшифровку фрагментов генома, включающих в себя длинные повторяющиеся и гомологичные участки.

СПИСОК ИСТОЧНИКОВ

1. 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
2. Korlach J., Bjornson K.P. et al. 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
3. Lundquist P.M., Zhong C.F., Zhao P. et al. Parallel confocal detection of single molecules in real time // *Optics Letters*. 2008. Vol. 33, no. 9. P. 1026–1028. DOI: 10.1364/OL.33.001026
4. Lipman D., Pearson W. Rapid and sensitive protein similarity searches // *Science*. 1985. Vol. 227. Id. 4693. P. 1435–1441. DOI: 10.1126/science.2983426
5. Pearson W., Lipman D. Improved tools for biological sequence comparison // *Proc. Natl. Acad. Sci. USA*. 1988. Vol. 85. P. 2444–2448. DOI: 10.1073/pnas.85.8.2444
6. Васильева Е.К., Чубинский-Надеждин И.В. Моделирование первичных данных одномолекулярного секвенатора. Ч.1. Симуляция детектирования излучения волнового нулевой моды // *Научное приборостроение*. 2026. Т. 36, № 1. С. 3–16. URL: <http://iairas.ru/mag/2026/abst1.php#abst1>
7. Вентцель Е.С. Теория вероятностей. Учеб. для вузов. М.: Высшая школа, 1999. 576 с.
8. Калашникова В.И., Козодаев М.И. Детекторы элементарных частиц. М.: Наука, 1966. 408 с.
9. Mattei A., Bailly N., Meissner A. DNA methylation: a historical perspective // *Trends in Genetics*. 2022. Vol. 38, no. 7. P. 676–707. DOI: 10.1016/j.tig.2022.03.010
10. Sigurpalsdottir B.D., Stefansson O.A., Holley G. et al. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes // *Genome Biology*. 2024. Vol. 25, no. 1. Id. 69. DOI: 10.1186/s13059-024-03207-9
11. Flusberg B.A., Webster D.R., Lee J.H. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing // *Nature Methods*. 2010. Vol. 7, no. 6. P. 461–465. DOI: 10.1038/nmeth.1459
12. Clark T., Spittle K., Turner S., Korlach J. Direct Detection and Sequencing of Damaged DNA Bases // *Genome Integrity*. 2011. Vol. 2. Id. 10. DOI: 10.1186/2041-9414-2-10
13. Артемкин Д.Е., Баринов В.В., Овечкин Г.В., Степнов И.М. Основы компьютерного моделирования систем. М.: Лаборатория Базовых Знаний, 2004. 152 с.
14. Schad E.E., Korlach J. et al. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases // *Genome Research*. 2013. Vol. 23. P. 129–141. DOI: 10.1101/gr.136739.111
15. Бусленко Н.П., Шрейдер Ю.А. Метод статистических испытаний (Монте-Карло) и его реализация на ЦВМ. М.: ФИЗМАТЛИТ, 1961. 226 с.
16. 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. Вып. 21, № 3. С. 572–603. DOI: 10.15622/ia.21.3.5

17. Andrade L., Manolakos E. Signal Background Estimation and Baseline Correction Algorithms for Accurate DNA Sequencing // Journal of VLSI Signal Processing. Kluwer Academic Publishers. Manufactured in The Netherlands. 2003. Vol. 35, P. 229–243.
DOI: 10.1023/B:VLSI.0000003022.86639.1f

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

Контакты: Васильева Екатерина Кимовна,
virtkat@mail.ru

Материал поступил в редакцию 23.03.2026

MODELING OF RAW DATA FROM A SINGLE-MOLECULE SEQUENCER. PART II: SIMULATION OF THE SEQUENCE OF EVENTS OF FLUORESCENTLY LABELED NUCLEOTIDES' INCORPORATION INTO A DNA STRAND AND COMPREHENSIVE MODELING OF MULTIPLE ZERO-MODE WAVEGUIDES IMAGES

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

Institute for Analytical Instrumentation of RAS, Saint Petersburg, Russia

The article discusses the characteristics of detecting fluorescence signals from labeled nucleotides in a zero-mode waveguide (ZMW). It introduces the concept of a "signal event" to describe a single emission act in one of the spectral channels during the incorporation of a labeled nucleotide. The model accounts for label fluorescence duration, the formation of signal event sequences, and the background intervals between them. The moments of signal occurrence are modeled as a stationary one-dimensional Poisson process, characterized by properties such as discrete countability, stationarity, and orderliness. The article emphasizes the necessity of employing non-stationary models or combinations of processes when considering epigenetic effects, such as methylation. Finally, the article presents a software simulator for primary data acquisition in a third-generation single-molecule sequencer. This software generates output in FASTA, "movie", and "trace" formats, enabling the creation of image series that simulate the fluorescent array of zero-mode waveguides in a measurement cell.

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

REFERENCES

1. Eid J., Fehr A., Gray J., Luong K. et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 2009, vol. 323, no. 5910, pp. 133–138.
DOI: 10.1126/science.1162986
2. Korlach J., Bjornson K.P. et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Methods in Enzymology*, USA, California, Pacific Biosciences, Menlo Park, 2010, vol. 472, pp. 431–455.
DOI: 10.1016/S0076-6879(10)72001-2
3. Lundquist P.M., Zhong C.F., Zhao P. et al. Parallel confocal detection of single molecules in real time. *Optics Letters*, 2008, vol. 33, no. 9, pp. 1026–1028.
DOI: 10.1364/OL.33.001026
4. Lipman D., Pearson W. Rapid and sensitive protein similarity searches. *Science*, 1985, vol. 227, id. 4693, pp. 1435–1441. DOI: 10.1126/science.2983426
5. Pearson W., Lipman D. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 1988, vol. 85, no. 8, pp. 2444–2448.
DOI: 10.1073/pnas.85.8.2444
6. Vasil'eva E.K., Chubinskiy-Nadezhdin I.V. [Modeling of raw data from a single-molecule sequencer. Part I: simulating detection of emission from a zero-mode waveguide]. *Nauchnoe Priboorostroenie* [Scientific Instrumentation],

- 2026, vol. 36, no. 1, pp. 3–16. (In Russ).
URL: <http://iairas.ru/mag/2026/abst1.php#abst1>
7. Ventcel' E.S. *Teoriya veroyatnostej. Uchebnik dlya vuzov* [Probability Theory: A Textbook for Universities]. Moscow, Vysshaya shkola Publ., 1999. 576 p. (In Russ).
 8. Kalashnikova V.I., Kozodaev M.I. *Detektory elementarnykh chastic* [Elementary particle detectors]. Moscow, Nauka Publ., 1966. 408 p. (In Russ).
 9. Mattei A., Bailly N., Meissner A. DNA methylation: a historical perspective. *Trends in Genetics*, 2022, vol. 38, no. 7, pp. 676–707. DOI: 10.1016/j.tig.2022.03.010
 10. Sigurpalsdottir B.D., Stefansson O.A., Holley G. et al. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biology*, 2024, vol. 25, no. 1, id. 69. DOI: 10.1186/s13059-024-03207-9
 11. Flusberg B.A., Webster D.R., Lee J.H. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 2010, vol. 7, no. 6, pp. 461–465. DOI: 10.1038/nmeth.1459
 12. Clark T., Spittle K., Turner S., Korlach J. Direct Detection and Sequencing of Damaged DNA Bases. *Genome Integrity*, 2011, Vol. 2, id. 10. DOI: 10.1186/2041-9414-2-10
 13. Artemkin D.E., Barinov V.V., Ovechkin G.V., Stepanov I.M. *Osnovy komp'yuternogo modelirovaniya sistem* [Fundamentals of Computer Modeling of Systems]. Moscow, Laboratoriya Bazovykh Znanij Publ., 2004. 152 p. (In Russ).
 14. Schad E.E., Korlach J. et al. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Research*, 2013, vol. 23, pp. 129–141. DOI: 10.1101/gr.136739.111
 15. Buslenko N.P., Shrejder Yu.A. *Metod statisticheskikh ispytaniy (Monte-Karlo) i ego realizatsiya na CVM* [The statistical testing method (Monte Carlo) and its implementation on a computer]. Moscow, FIZMATLIT Publ., 1961. 226 p. (In Russ).
 16. Borodinov A.G., Manoilov V.V., Zarutsky I.V., Petrov A.I., Kurochkin V.E., Saraev A.S. [Machine learning in base-calling for next-generation sequencing methods]. *Informatika i avtomatizatsiya* [Informatics and automation], 2022, vol. 21, no. 3, pp. 572–603. DOI: 10.15622/ia.21.3.5 (In Russ).
 17. Andrade L., Manolagos E. Signal Background Estimation and Baseline Correction Algorithms for Accurate DNA Sequencing. *Journal of VLSI Signal Processing*, 2003, vol. 35, pp. 229–243. DOI: 10.1023/B:VLSI.0000003022.86639.1f

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

Article received by the editorial office on 23.03.2026

INTRODUCTION

The second part of the article concludes the description of a simulation model for fluorescence detection in zero-mode waveguides of a single-molecule sequencer. Real-time sequencing of single, non-clustered DNA molecules is a third-generation method aimed at obtaining reads thousands of nucleotide pairs long. Such reads with known characteristics can be obtained by simulating long sequences of nucleotide incorporation events into DNA. Such simulations for the SMRT (Single Molecule Real Time) method [1–3] can have important practical applications in a variety of applications, from developing and testing secondary data processing algorithms to optimizing chemical processes and improving sequencing accuracy.

Developing robust base-calling algorithms, often based on machine learning, requires huge data sets with known true sequences. Simulations enable the generation of synthetic data in any volume, precisely controlling the sequence and signal parameters, which is impossible in real experiments. Testing the robustness of secondary processing algorithms is a critical task, for which realistic simulations allow us to verify how base-calling algorithms respond to various noises, artifacts, changes in polymerase speed, and omissions or prolongations of intervals between incorporation events, simulating real biological or instrumental failures.

The SMRT method not only reads the nucleotide sequence of a DNA strand but also reveals the kinetics of polymerase activity. Simulation allows us to accurately reproduce the distribution of interval durations between incorporated events and determine the optimal camera exposure time to accurately capture short fluorescence pulses while minimizing signal blurring or data loss.

Sequence-of-event simulation is only one step in a complete SMRT simulator. A key part of the complete data simulation is the integration of two steps—the nucleotide letter sequence simulation and the basic fluorescence emission event image simulation—where the event sequence is the input to the optical response simulation step. Simulating tens of thousands of zero-mode waveguides (ZMWs) requires an efficient and accurate event generator for each waveguide to simulate the operation of the entire sequencing cell.

SIMULATION OF THE EVENTS SEQUENCE FOR INCORPORATION OF LABELED NUCLEOTIDES INTO A DNA CHAIN

Parameters for primary data simulation

We define a *signal-event* as an event that consists of the emission of light in one of the spectral channels during the complementary incorporation of a fluores-

cently labeled nucleotide into a DNA chain. The second stage of the simulation is the formation of a signal-event sequence from basic signal counts, consistent with a given sequence of nucleotide letter designations. The duration of each such event should correspond to the duration of the fluorescent label's glow during incorporation of the labeled nucleotide.

All parameters used in the full simulation of primary data from a single-molecule sequencer can be divided into four groups according to the task being solved at each stage of data generation (Fig. 1, Table 1).

Fig. 1. Primary data simulation stages.

1 — generation or reading of the nucleotide sequence; 2 — generation of basic images and basic signals (samples) for one ZMW; 3 — generation of a sequence of signal events in four spectral channels for one ZMW; 4 — generation of a frame image series with data simulation for multiple ZMW

Tab. 1. Parameters of the simulation algorithm at different stages of data generation

At the first stage, the initial data is entered into the simulation algorithm — the nucleotide sequence is read. The nucleotide sequence in FASTA format [4, 5] can be obtained as a file from a genomic database or randomly generated directly in the simulator program.

Group 2 parameters are required to generate basic images and integral basic signals in the simulation domain for four spectral channels in accordance with the algorithm described in the first part of the article [6]. Simultaneously with the generation of this data and based on it, a general sequence of event-signals is constructed, traces and image files in movie format are generated and saved for each ZMW. Parameter Group 3 is used for this, regulating the nature of the occurrence of event-signals and their duration.

In the final stage, using Group 4 parameters, images of individual ZMW, the data for which was previously generated and saved to disk, are assembled into a frame, and then assembled into a series of multiple full-size image frames.

The sequence of signal events as a Poisson random process

The fluorescence of the label's fluorophore begins at the moment the nucleotide is fixed by the polymerase in the laser beam's range and continues as long as the label is retained in this region and does not lose its ability to fluoresce. According to various estimates, it takes from a few to hundreds of

milliseconds, which typically exceeds the exposure time of a single frame. During this time, the detection system records the fluorescence signal. After the label leaves the polymerase's retention zone, the video camera will record the background signal for the ZMW until a new labeled nucleotide approaches the polymerase. Thus, the sequencing process sequentially records relatively bright signal-events (pulses) and less bright background signals emitted by the ZMW between events (interpulses). The rate of the incorporation process, i.e., the frequency of occurrence of signal events, can vary across experiments.

Let the onset of a random signal event be the onset of fluorescence emission in the nucleotide incorporation reaction. The sequence of such events then forms a flow of reaction that starts. This flow, in turn, is generated by a Poisson flow of nucleotide uptake events by the polymerase. In the model under consideration, this flow is stationary.

According to [7, 8], the flow of random homogeneous events is a one-dimensional stationary Poisson flow and can be described by processes that obey the conditions of stationarity, ordinarity, and the absence of aftereffects. To describe and simulate epigenetic processes such as DNA methylation [9, 10], it will be necessary to use a non-stationary Poisson flow or a combination of two random processes with different event frequencies, since the durations of the IPD (interpulse duration) intervals between the events of incorporation of normal and methylated nucleotides differ significantly [11, 12].

In [8, 13], it has been shown that the distribution determining the periods between successive homogeneous events in a Poisson flow, i.e., predicting the lengths of the time intervals between them, can be described by an exponential function. In [14], it was noted that this approach is applicable to modeling the time intervals between SMRT fluorescence signals. When simulating the lengths of time intervals between events in a one-dimensional Poisson flow, a relationship linking random numbers with a given distribution law and numbers uniformly distributed over the interval from 0 to 1 can be used to obtain a sequence of random numbers with a given distribution law [7–9, 15]:

$$t_i = -\frac{1}{A} \ln(1 - R_i), \quad (1)$$

where t_i is the length of the i -th time interval, A is the average intensity of the flow of co-events, and R_i is a random value obtained over the interval from 0 to 1 using a random number generator.

Thus, an array of initial time instants is generated programmatically for intervals $t_1, t_2, \dots, t_i$, corresponding to the start times of nucleotide incorporation reactions. Simultaneously, an array of nucleotide letter

designations is generated or read from a file. Then, by simulating random event signal durations $PW_1, PW_2, \dots, PW_i$ (pulse width) for each nucleotide from the specified letter array, we can associate them with the obtained reaction start times (Fig. 2). By comparing the nucleotide designation, reaction start time, and reaction duration, we simulate basic images and signals using the algorithmic templates described in the first part of the article with an individual set of parameters for each nucleotide. In this case, we form five sequences—one general and four for the letters A, T, G, and C, each corresponding to its own nucleotide.

Fig. 2. Formation of a sequence of signal events

Event signal durations can be specified as constant or variable using normal or lognormal distributions. Nucleotides cannot be incorporated simultaneously, but they can be incorporated one after another so that the end of one reaction and the beginning of the next are contained within the duration of a single exposure. If the duration of the event PW_i , which began at time t_i , overlaps the start time of the next event $t_{(i+1)}$, the moment $t_{(i+1)}$ must be skipped and the next moment $t_{(i+2)}$ must be used (Fig. 3, a). If this moment is also overlapped by the duration of the PW_i event, the next start moment $t_{(i+3)}$ should be selected, and so on. A relatively long duration of event signals or a high event occurrence rate per second, for example, 5 events per second with an average duration of 200 ms, will automatically lead to an increase in the overall flow implementation time due to a decrease in the event occurrence rate while maintaining the Poisson nature of their occurrence.

Fig. 3. Features of the formation of a sequence of event signals

The start or end moment of an incorporation event falls randomly within the next sampling time interval; therefore, the duration of an event signal almost always cannot be sampled by an integer number of intervals equal to the exposure time (Fig. 3, b). For correct modeling, we assume that if the incorporation event began in the first half of the next exposure interval or ended in the second half, as in Fig. 3, events t_2 and t_4 (trailing edge) and t_5 (leading and trailing edges), then the specified fluorescence intensity of the basic image and signal are multiplied by a factor of 0.6. In other words, we assign the basic signal two-thirds of the intensity specified in the modeling, since most of the exposure interval under consideration is "overlapped" by the simulated pulse. If the incorpora-

tion event began in the second half of the next exposure interval or ended in the first half, as in Fig. 3, events t_3 (trailing edge) and t_4 (leading edge), then the specified fluorescence intensity of the basic image and signal are multiplied by a factor of 0.3. That is, if the signal-event duration interval is not a multiple of the exposure time, then the first and/or last signal-event sample, which obviously lasts less than the exposure time, decreases in amplitude. This is equivalent to the case when, during sequencing, the fluorescent label's emission occupies less than half the full exposure time, leading to a weakening of the total accumulated fluorescence signal during the exposure time.

The basic signals are used to plot the intensity of the labels' emission in a single spectral channel over time. On the dependence graph, each sampling interval equal to the exposure time can be assigned an integral signal simulating the signal accumulated on the video camera during that exposure time, thus forming a trace.

By generating basic images and calculating basic signals for four spectral channels, we gradually form the entire sequence of signal-events.

Traces specified digitally for four spectral detection channels can subsequently be used for base-calling when processing nucleotide sequence recognition tasks [16, 17].

Thus, a signal-event sequence can be formed as a series of single samples—basic signals—continuously following one another with a sampling interval equal to the exposure time. The sequence of such events itself represents realizations of a Poisson event flow, and variations in the time interval durations between these signals can be described as realizations of an exponential distribution law.

The sequence generation logic is shown in Figure 4.

Fig. 4. An example of generating traces and a movie for a short GCACG sequence

Fluorescent labels flash in polymerase-captured molecular complexes. Sequential flashes of different fluorochromes — in this case, red, dark red, green, dark red, and red — correspond in duration to event signals, shown in the same four colors. The overall pulse sequence is shown in gray on the time axis. The overall sequence shows that the last three event signals have merged and cannot be easily separated without information about the spectral channel in which they were recorded. Furthermore, we see that these are three optical signals with different spectral compositions. (The color argumentation can only be appreciated in the color format of the electronic version of the article.)

With a sampling rate equal to the exposure time, basic images are formed for all spectral channels. The number of basic images for each signal-event corresponds to the number of times the exposure time fits into the fluorescence time interval of the corresponding fluorochrome. The set of such images for a single ZMW represents movie-type data and is saved to disk as a graphic file. The corresponding basic signals form traces for each channel and the overall signal-event sequence. Traces for all four recording channels are saved to disk as a single file common to the ZMW.

Simulation of a series of frames

The problem of simulating a set of primary data in the form of full-size movie frames measuring 2048 by 2048 pixels is solved in two stages. In the first stage, basic images of the signal-event sequences are formed for all ZMW in the specified set for simulation. This data is saved to disk as movie-format image files and is generated for each ZMW individually according to specified parameters using the algorithms described previously.

In the second stage, frames are "assembled" from these pre-generated files. For this purpose, an image fragment corresponding to the data simulation for the next i -th recording time point (Fig. 5) is selected in each movie file for a single ZMW. This fragment must be transferred to the image of the i -th frame in the full simulation.

Fig. 5. Movie-type data for a single ZMW.

In (a), a fragment of the image with two incorporation events in the G channel (pixels with bright luminescence) is highlighted. In (b), the same fragment is shown at a larger scale with vertically highlighted sections simulating fluorescence responses formed at different moments of recording in four spectral channels. The first incorporation event is discretized into 5 time intervals ($n = 5$), equal to the exposure time during recording; the second incorporation event is discretized into 3 intervals ($n = 3$); between them are 4 fragments with a background signal

By sequentially copying image fragments for different single NMs into pre-calculated coordinates on the full-size image (Fig. 6), it is possible to form an image of any frame with multiple ZMW exhibiting fluorescent responses in different channels at the given time point.

Fig. 6. Movie-type data for multiple frames with different ZMW in the simulated data series

A series of a large number of such frames will simulate real raw images of a single-molecule sequencer obtained over a long period of time, where fragments $i_1, i_2 \dots, i_n$ in of the movie images for each ZMW are transferred to the corresponding positions on frames $i_1, i_2 \dots, i_n$. Thus, the number of exposure intervals will exactly correspond to the number of frames in the simulated dataset. These frames simulate the raw data that, during actual sequencing, is generated at the output of the detector's video camera. The raw data also includes traces, but in the actual detection process, these can only be obtained after processing movie-type data.

CONCLUSION

This paper proposes a stochastic simulation model for detecting the fluorescence of labeled nucleotides during single-molecule sequencing in four spectral channels. Multifactorial simulation of detector signals

is provided. A simulator program for raw digital data from a single-molecule sequencer has been developed, providing simulation of sequencing results in FASTA, trace, and movie formats.

Accurate software simulation allows one to assess how the signal-to-noise ratio and the frequency of incorporation events (and, consequently, the duration of interpulse intervals) affect base-calling efficiency and the overall accuracy of the sequence reading.

Simulation of the optical responses and the sequence of nucleotide incorporation events into the DNA strand are important components of a realistic simulator of raw data from a single-molecule sequencer. The model enables the generation of synthetic data for training and validating secondary processing algorithms, studying polymerase kinetic effects, and optimizing sequencing conditions. This could ultimately facilitate the decoding of genome fragments containing long, repetitive, and homologous regions.