УДК 543.07 **DOI** 10.15622/ia.21.3.5

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

МАШИННОЕ ОБУЧЕНИЕ В ЗАДАЧАХ BASE-CALLING ДЛЯ МЕТОДОВ СЕКВЕНИРОВАНИЯ НОВОГО ПОКОЛЕНИЯ

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.**

Аннотация. Развитие технологий секвенирования следующего поколения (NGS) внесло существенный вклад в тенденции снижения затрат и получения массивных данных секвенирования. В Институте аналитического приборостроения РАН разрабатывается аппаратно-программный комплекс (АПК) для расшифровки последовательности нуклеиновых кислот методом массового параллельного секвенирования (Нанофор СПС). Алгоритмы обработки изображений, входящие в состав АПК, играют существенную роль в решении задач расшифровки генома. Финальной частью такого предварительного анализа сырых данных является процесс base-calling. Base-calling — это процесс определения нуклеотидного основания, которое генерирует соответствующее значение интенсивности в каналах флуоресценции для различных длин волн на кадрах изображения проточной ячейки для различных циклов секвенирования методом синтеза. Приведен обширный анализ различных подходов к решению задач base-calling и сводка распространенных процедур, доступных для платформы Illumina. Рассмотрены различные химические процессы, включенные в технологию секвенирования методом синтеза, вызывающие смещения в значениях регистрируемых интенсивностей, включая эффекты фазирование / префазирование (phasing/prephasing), затухания сигнала (signal decay) и перекрестные помехи (cross-talk). Определена обобщённая модель, в рамках которой рассматриваются возможные реализации. Рассмотрены возможные подходы машинного обучения (machine learning) для создания и оценки моделей, реализующих этап обработки base-calling. Подходы ML принимают различные формы, включая обучение без учителя (unsupervised), обучение с ча-стичным привлечением учителя (semi-supervised), обучение с учителем (supervised). В работе показана возможность применения различных алгоритмов машинного обучения на основе платформы Scikit-learn. Отдельной важной задачей является оптимальное выделение признаков, выделенных в обнаруженных кластерах на проточной ячейке для машинного обучения. Наконец, на ряде данных секвенирования для приборов MiSeq Illumina и Нанофор СПС показана перспективность метода машинного обучения для решения задачи base-calling.

Ключевые слова: секвенирование нового поколения, base-calling, биоинформатика, машинное обучение.

1. Введение. Последние достижения в технологии высокопроизводительного секвенирования позволяют одновременно секвенировать миллионы фрагментов ДНК, создавая огромную пропускную способность, что в свою очередь требует более эффективных и точных методов анализа [1]. Поскольку стоимость секвенирования продолжает быстро снижаться, перспективы геномики расширяются. В то же время необходимо решить ряд проблем,

связанных с получением данных и статистическим анализом. Например, получение определенной последовательности нуклеотидов из данных интенсивностей флуоресценции усложнено из-за химических процессов, а также из-за многочисленных оптических аппаратных искажений. Переход от зашумленных данных интенсивностей флуоресценции, как результата процесса обработки изображений к последовательностям нуклеотидных оснований получил название base-calling. Точность определения каждого нуклеотида (base-call) представлена в виде показателя качества (quality score) и соотносится с каждым определенным нуклеотидным основанием. Под ридом (read, прочтение) понимается отдельная последовательность нуклеотидов, полученная в результате секвенирования. Эти статистические показатели далее используются в процессе выравнивания (alignment) полученных ридов с известным эталонным геномом в процессе таких видов анализа, как анализ ChIP-seq [2] или RNA-seq [3]. Следовательно, точность и качество процедуры base-calling могут напрямую влиять на дальнейший анализ последовательностей ДНК. В работе представлен обзор решений задачи base-calling и апробация нового алгоритма base-calling, основанного на методах машинного обучения.

2. Основы метода

2.1. Биологические основы. Хромосома эукариот образуется из единственной и чрезвычайно длинной молекулы ДНК, которая содержит линейную последовательность множества генов. Хромосомы состоят из более мелких субъединиц, называемых нуклеотидами, каждая из которых содержит пентозный сахар, фосфатную группу и одно из четырех азотистых оснований (A, C, G, T). Общепринято использование букв A, C, G и T для обозначения нуклеотида, который содержит соответствующее азотистое основание. Важно отметить, что A связывается с T и C связывается с G, поэтому без потери общности одна нить ДНК определяет другую (комплементарна ей).

2.2. Высокопроизводительное секвенирование (NGS). Во-первых, следует получить несколько образцов генома из организма (стадия сбора или acquisition). Поскольку геномы многих организмов, как правило, очень длинные, проанализировать последовательность от начала до конца не представляется возможным. Поэтому отобранная ДНК разрезается на более мелкие фрагменты и делается несколько копий для усиления полезного сигнала. Эти шаги называются фрагментация (fragmentation) и амплификация (amplification), соответственно. Предыдущие шаги образуют стадию подготовки

геномного образца (genomic sample preparation) или подготовки библиотеки (library preparation) [4].

Основания нуклеотидов флуоресцентно помечены, и отдельные изображения в различных каналах флуоресценции получаются для каждого из четырех оснований A, C, G и T. Каналы флуоресценции отличаются друг от друга длинами волн возбуждающего сигнала. Изображения далее обрабатываются для получения значений интенсивности для каждого нуклеотида [5]. Каждый из нуклеотидов генерирует сигнал флуоресценции на определенной длине волны. Base-calling — это процесс определения нуклеотидного основания, которое генерирует соответствующее значение интенсивности в каналах A, C, G и T. Существует неопределенность при проведении процедуры base-calling, поэтому соответствующая оценка качества присваивается каждому выбранному основанию. Этот показатель качества (quality score) является функцией вероятности неправильного выбора нуклеотидного основания. В таблице 1 приведен список распространенных методов base-calling. Поскольку произошла фрагментация цепочки ДНК, информация о позициях расположения ридов теряется. Так как нас интересует весь геном или очень большие части генома, мы должны попытаться собрать фрагменты вместе, чтобы получить исходные позиции нуклеотидов в полной цепочке ДНК. Если существует эталонный геном, он может помочь нам определить местоположение секвенированных участков ДНК [6].

Несмотря на всю кажущуюся вариабельность методов base-calling наиболее широко используемым базовым алгоритмом является Bustard, несколько алгоритмов были построены с использованием Bustard в качестве отправной точки. Алгоритм Bustard основан на параметрической модели и применяет алгоритм Маркова для определения вероятности моделирования переходной матрицы фазинга/префазинга (phasing/prephasing), и матрицы cross-talk. Алгоритм Bustard предполагает, что матрица перекрестных помех постоянна для данного цикла секвенирования, и что эффекты фазинга одинаковы относительно всех нуклеотидных оснований.

В методах BayesCall and naiveBayesCall для оценки всех неизвестных параметров используется алгоритм максимума правдоподобия (EM), а полученная максимальная апостериорная вероятность используется в base-calling. При этом naiveBayesCall существенно выигрывает в оптимизации скорости вычислений. Основной мотивацией для модификации OnlineCall является создание вычислительно эффективного алгоритма с базовой моделью BayesCall. Оценка параметров выполняется с помощью unsupervised онлайн-

алгоритма EM, а полученные апостериорные вероятности используются для base-calling. В алгоритме Softy используется также обобщенная модель base-calling, но апостериорные вероятности, используемые для процедуры base-calling, получаются либо с помощью алгоритма прямого-обратного прохода (FB), либо алгоритма soft-output Витерби (SOVA).

Таблица 1. Сводка распространенных процедур base-calling

Наименование	Год	Формат ввода	Мера качества	Тип модели
BlindCall [7]	2014	RTA	None	Blind deconvolution
3Dec [8]	2017	RTA	Phred	Parametric
freelbis [9]	2013	IPAR, Firecrest, Bustard reads	Phred	Nonparametric SVM
Softy [10]	2013	Firecrest	Probability	Parametric
AYB [11]	2012	RTA	Phred	Nonparametric
OnlineCall [12]	2012	Firecrest	Probability	Parametric
BM-BC [13]	2012	Firecrest	Unknown	Parametric
ParticleCall [14]	2012	Firecrest	Probability	Parametric
TotalReCaller [15]	2011	Firecrest, RTA	None	Parametric
naiveBayesCall [16]	2010	Firecrest	Probability	Parametric
Srfim [17]	2009	IPAR, Firecrest	Phred	Parametric
BayesCall [18]	2009	Firecrest	Probability	Parametric
Ibis [19]	2009	IPAR, Firecrest, Bustard reads	Phred	Nonparametric SVM
Rolexa [20]	2008	IPAR, Firecrest	Phred	Parametric
Alta-Cyclic [21]	2008	Firecrest с Bustard reads	Unknown	Nonparametric SVM

Метод Rolexa предлагает новый алгоритм base calling, использующий кластеризацию на основе параметрических моделей и статистические оценки для выявления неоднозначных нуклеобаз и кодирования их символами IUPAC (International Union of Pure and Applied Chemistry). Используются также оптимальные вложенные теги (sub-tags), используя оценку, основанную на информационном содержании, чтобы удалить неопределенные основания ближе к концу прочтения.

All Your Base (AYB) base-caller — это еще один метод, который полностью отличается от Bustard и семейства модификаций Bustard. Он основан на модели, но не предполагает наличия определенных

распределений наблюдаемых интенсивностей, что делает его непараметрическим методом. Истинная матрица нуклеобаз получается с помощью алгоритма Витерби, который находит наиболее вероятную последовательность нуклеотидов, а алгоритм прямого/обратного распространения (Forward/Backward) используется для получения апостериорных вероятностей, что приводит к оценкам показателей качества.

Метод улучшенной базовой идентификации (Improved Base Identification System, Ibis) для проведения base-calling использует схему статистического обучения. Этот подход использует метод опорных векторов (SVM) для поиска закономерностей в данных. SVM используют полиномиальные ядра с входными данными текущего, предыдущего и последующего циклов. Этап обучения находит оптимальную гиперплоскость, которая может разделить паттерны между интенсивностями четырех каналов. Для обучения SVM требуются достаточно объемные результаты выравнивания прочтений относительно референтного генома.

Новейший метод base-calling 3Dec учитывает эффект cross-talk между ближайшими кластерами.

3. Base-calling в секвенировании методом синтеза (Sequencing-by-synthesis). Чтобы прояснить процедуру base-calling обсудим принципиально новые методы химической подготовки анализируемых образцов к проведению генетического анализа., используемую в платформе Illumina и называемую секвенированием методом синтеза (Sequencing-by-synthesis) [22].

После получения образцов генома двухцепочечная ДНК случайным образом фрагментируется посредством обработки ультразвуком. Короткие известные последовательности, называемые адаптерами, лигированы к концам двухцепочечных фрагментов. Адаптеры используются для прикрепления фрагментов к поверхности, на которой располагаются короткие последовательности ДНК. Чтобы генерировать достаточное количество ДНК-материала для последовательности используют стадии полимеразной цепной реакции для создания нескольких дубликатов молекул ДНК / адаптера. Эти молекулы денатурируются, превращаясь в одноцепочечные, а затем связываются с поверхностью стеклянной проточной ячейки, содержащей плотное множество (газон) прикрепленных олигопраймеров. Проточные кюветы (Flowcells), как правило, содержат 8 дорожек (lanes), а меньшие субъединицы, называемые плитками (tiles), составляют одну дорожку. Количество плиток варьируется между версиями Illumina, но обычно содержит до 100

плиток. Олигопраймеры комплементарны одному концу одноцепочечной цепи матрицы. Полимеразный фермент завершит построение комплементарной цепи, и исходный шаблон будет смыт. Затем происходит "стыковочная" амплификация (bridge amplification) с целью несколько раз скопировать шаблоны, чтобы получить почти 1000 копий идентичных одиночных нитей и достаточно плотный кластер. Последовательности синтезируются по одному основанию за раз параллельно во всей проточной ячейке. Процесс присоединения одного азотистого основания (нуклеотида) будем называть циклом. Для каждого цикла добавляется ДНК-полимераза и молекулы, состоящие из флуоресцентно меченных оснований с присоединенным обратимым терминатором, которые позволяют последовательно проводить процесс, предотвращая присоединение более чем одного основания. После присоединения одного основания к цепочке лазер возбуждает кластер, генерируя излучение флуоресценции на определенной длине волны.

В системе параллельного секвенирования прибора «Нанофор СПС» используются четыре видеокамеры по числу типов нуклеотидов. Каждая из видеокамер настроена на регистрацию одного из типов нуклеотидов: **«A»**, **«C»**, **«G»** или **«T»**. Сигнал флуоресценции возбуждается двумя лазерами в определенном диапазоне излучения видимого света. Регистрируемое излучение пропускается через различные светофильтры, соответствующие длинам волн флуоресценции каждого из четырех красителей, которыми специфично помечены нуклеотиды. Таким образом, каждая из видеокамер регистрирует изображения кластеров молекул ДНК, на конце которых расположены нуклеотиды определенной «буквы».

Четыре набора изображений фиксируются через четыре различных фильтра, один для каждого из красителей, используемых соответственно для каждого из оснований. Терминаторы удаляются, чтобы обеспечить включение основания в следующий цикл секвенирования. Это продолжается до тех пор, пока полное прочтение (read) не будет полностью выполнено.

Обработка изображения в заданном цикле дает множество из четырех интенсивностей, где каждое из четырех значений представляет интенсивность, считанную через определенный оптический фильтр. Выходные данные этапа обработки изображений содержат значения интенсивности для всех прочтений (ридов) и циклов, (x,y) -координаты позиций кластера на проточной ячейке и, наконец, нуклеотидные основания, получаемые как результат процедуры base-calling. На рисунке 1 показан пример значений

интенсивности для одного прочтения. Самый элементарный метод base-calling выдал бы последовательность на основе значения максимальной интенсивности, и в этом случае это будет последовательность GGAAAATGAG. В некоторых циклах максимум очевиден, но в других случаях самые большие и вторые по величине значения интенсивностей достаточно близки, см. циклы 7 и 3 соответственно. На соотношениях максимальной и остальных интенсивностей в различных каналах строится построение фильтров, отсекающих кластеры недостаточного качества (chastity, purity). Только статистические методы помогут моделировать эти сложные ситуации и разрешить коллизии. Несмотря на то, что технология секвенирования бурно развивается в последние годы, особенности в химии протекания процессов в проточной кювете продолжают вызывать сложности в процессе определения нуклеотидов из данных зафиксированных интенсивностей.

```
> intensities
      1     2     3      4      5      6     7      8      9     10
A -17.7  16.5 847.7 1077.6 1044.7 1039.9  17.4   55.6 1015.9   63.5
C   9.2  34.8 651.8  835.4  754.6  708.4  38.1   50.8  736.1   36.5
G 1121.5 955.8  -6.4   15.4    9.9    3.9  37.2 1146.9   37.4 1234.4
T  588.9 494.9  14.8    3.6    5.4   25.6 639.2  647.4   30.6  670.7
```

Рис. 1. Пример значений интенсивности для первых 10 циклов чтения. Строки представляют интенсивности в определенных каналах, а столбцы представляют циклы. Элементарный base-caller выдал бы последовательность для прочтения GGAAAATGAG

Химические процессы, включенные в технологию секвенирования методом синтеза, вызывают некоторые смещения в значениях регистрируемых интенсивностей, включая эффекты фазирование / префазирование (phasing/prephasing), затухания сигнала (signal decay) и перекрестные помехи (cross-talk) [23]. В любом данном кластере возможно, что небольшая часть синтезированных цепочек будет отставать в синхронизации присоединения нуклеотидов по сравнению с остальными цепями. Например, ферменты могут не сработать, в результате чего основание не включится в цепочку ДНК. Нить ДНК может просто отставать, перегонять планируемый процесс присоединения по одному основанию или полностью становиться неактивной, что приводит к неточным показаниям интенсивностей в цикле. Это явление называется фазированием (phasing или lagging). Префазирование (prephasing или leading) происходит, когда небольшая часть цепей забегает вперед и включает сразу два нуклеотида. Prephasing имеет те же последствия, что и phasing. Как phasing, так и prephasing могут привести к неправильному определению

нуклеотидного основания. На рисунке 2 показано, как phasing и prephasing могут выглядеть с точки зрения анализа интенсивностей. Интенсивности канала A для одного кластера отображаются в зависимости от цикла, а в нижней части каждой строки показано основание. Следовательно, более высокие значения интенсивности соответствуют нуклеотидному основанию, полученному из стандартной процедуры base-calling от Illumina. Фиолетовая стрелка указывает на то, что эффект phasing произошел, поскольку после нуклеотида A наблюдается более высокий сигнал, а черная стрелка указывает на то, что эффект prephasing произошел, поскольку существует более высокий сигнал от нуклеотидных оснований, не являющихся A. Во время процесса секвенирования проточная ячейка промывается несколько раз, и возможно, что сам секвенируемый материал также будет смыт. Кроме того, неспособность ферментов приводит к неактивности в некоторых нитях секвенируемого ДНК. Такие потери вызывают снижение интенсивности сигнала и увеличение шума в процессе секвенирования. Это явление известно как затухание сигнала (signal decay), и, как очевидно, существует корреляция между длиной циклов секвенирования и количеством потерянного материала. Затухание сигнала можно увидеть на рисунке 3. Этот график показывает максимальную и минимальную из четырех интенсивностей, которые мы можем рассматривать как сигнал и шум соответственно, за несколько прочтений и циклов. Наклон сигнала и шума показывают тенденцию к уменьшению сигнала и увеличению шума по всему циклу. Флуоресцентные красители, используемые в платформе Illumina, имеют частично перекрывающуюся частоту излучения, что приводит к некоторой корреляции показаний интенсивностей. На рисунке 4 показана диаграмма интенсивностей по каналам A и C. Красное, синее и серое облако точек показывают корреляцию интенсивностей A/C для оснований A, C, G и T соответственно. По мере увеличения интенсивности A интенсивность C также увеличивается для тех оснований, которые были названы A, и, когда интенсивность C увеличивается, интенсивность A также увеличивается для тех оснований, которые названы C. Корреляция между каналами интенсивности A/C для оснований A и C не обязательно совпадают.

Частота излучения используемых красителей частично перекрывается, что приводит к корреляции показаний интенсивностей. Это приводит к тому, что с ростом интенсивности A растет и интенсивность C, и наоборот. Подобное явление происходит и с каналами G и T.

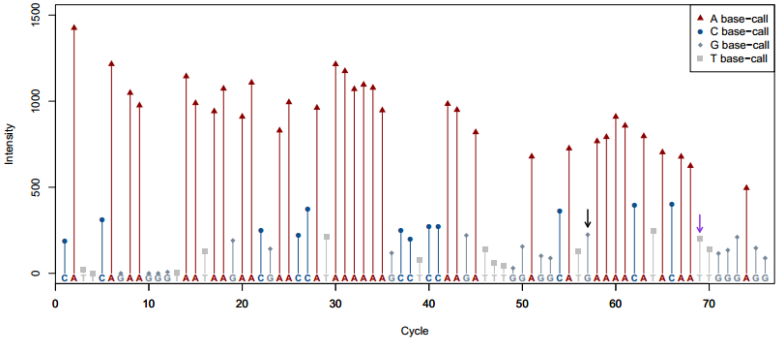

Рис. 2. На этом рисунке показано, где могут происходить phasing и prephasing и как могут выглядеть при этом значения интенсивностей. Phasing может произойти в цикле 69 (фиолетовая стрелка), потому что интенсивность в канале A составляет около 200, даже если соответствующий этому циклу нуклеотид был T. Следует обратить внимание, что предыдущий нуклеотид был A, и аналогично в цикле 57 интенсивность в канале также составляет около 200, даже если соответствующее каналу основание было G. Это показывает, что процесс prephasing мог произойти, поскольку нуклеотидом в следующем цикле является A [23]

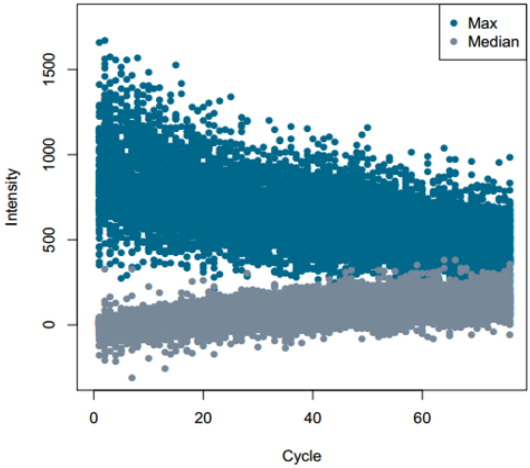

Рис. 3. Максимум из четырех интенсивносей и медиана оставшихся трех значений нанесены на график зависимости от цикла секвенирования. Тенденция к снижению сигнала и тенденция к увеличению шума показывают, как затухание сигнала влияет на значения интенсивности в разных циклах [23]

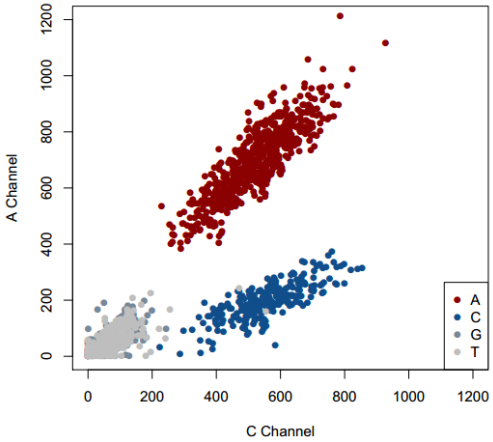

Рис. 4. Интенсивности каналов A и C нанесены для множества прочтений определенном цикле 30. Цвет показывает нуклеотид, поэтому мы можем видеть три облака точек; красный - для A, синий - для C, а серые - для оснований G и T. Поскольку интенсивность A увеличивается для оснований A, интенсивность C также увеличивается. Эта корреляция значений интенсивности является результатом процесса cross-talk [23]

4. Модель Base-calling

4.1. Нотация. Сначала определим математические обозначения для модели base-calling. В таблице 2 даны обозначения, которые будут использоваться для описания модели base-calling. В таблице 3 приведен пример того, как четверки интенсивностей представлены в математических обозначениях. Затем определим обобщеннную модель base-calling.

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

i		Индекс прочтений (read index), i = 1,2, ..., N
j		Индекс циклов (cycle index), j = 1, 2, ..., J
k		Индекс каналов (signal channel index), k = A,C,G,T
$\mathbf{Z_i}$	**4 x J**	Массив интенсивностей после коррекции Illumina
B	**4 x J**	Массив коррекции фона (background correction)
$\mathbf{Y_i}$	**4 x J**	Массив на Наблюдаемые интенсивности
M	**4 x 4**	Cross-talk матрица
$\mathbf{X_i}$	**4 x J**	Массив истинных интенсивностей
P	**J x J**	Phasing/prephasing матрица
D	**J x J**	Signal decay матрица
$\mathbf{E_i}$	**4 x J**	Error term матрица

Таблица 3. Пример представления интенсивностей согласно введенной нотации

Наблюдаемые интенсивности Y_i

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	Y_{i1A}	Y_{i2A}	Y_{i3A}	Y_{i4A}	Y_{i5A}	Y_{i6A}	Y_{i7A}	Y_{i8A}	Y_{i9A}	Y_{i10A}
C	Y_{i1C}	Y_{i2C}	Y_{i3C}	Y_{i4C}	Y_{i5C}	Y_{i6C}	Y_{i7C}	Y_{i8C}	Y_{i9C}	Y_{i10C}
G	Y_{i1G}	Y_{i2G}	Y_{i3G}	Y_{i4G}	Y_{i5G}	Y_{i6G}	Y_{i7G}	Y_{i8G}	Y_{i9G}	Y_{i10G}
T	Y_{i1T}	Y_{i2T}	Y_{i3T}	Y_{i4T}	Y_{i5T}	Y_{i6T}	Y_{i7T}	Y_{i8T}	Y_{i9T}	Y_{i10T}

↓

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	-17.7	16.5	847.7	1077.6	1044.7	1039.9	17.4	55.6	1015.9	63.5
C	9.2	34.5	651.8	835.4	754.6	708.4	38.1	50.8	736.1	36.5
G	1121.5	955.8	-6.4	15.4	9.9	3.9	37.2	1146.9	37.4	1234.4
T	588.9	494.9	14.8	3.6	5.4	25.6	639.2	647.4	30.6	670.7

4.2. Обобщённая модель base-calling. Все методы base-calling работают со значениями интенсивности, полученными во время секвенирования. После рассмотрения различных методов base-calling становится ясно, что в методах, используемых для моделирования интенсивностей, есть немало общего. Эти методы варьируются от параметрических до непараметрических и статистических моделей, основанных на полностью эмпирических методах машинного обучении. Попытаемся определить общую модель, которая объединяет подавляющее большинство методов base-calling. Классификация методов base-calling приведена на рисунке 5. Сначала определим следующую обобщенную модель для base-calling как:

$$Z_i - B = Y_i = MX_i PD + E_i \quad (1)$$

В формуле (1) использованы обозначения согласно таблице 2, содержащей краткие описания параметров, а также предполагаемые размеры матриц. Использованы следующие индексы; прочтение (read) / кластер **i** для **i = 1, 2, ..., N**, цикл **j** для **j = 1,2, ..., J** и сигнальный канал **k** для **k** из **{A, C, G, T}**. Обычно индексы **1**, **2**, **3** и **4** используются взаимозаменяемо с **A**, **C**, **G** и **T** соответственно. $\mathbf{Z_i}$ - это интенсивности перед коррекцией фона Illumina B. $\mathbf{Y_i}$ - это наблюдаемые интенсивности. Наш интерес заключается в получении значений $\mathbf{X_i}$, интенсивности для данного нуклеотидного основания. Например, если истинный нуклеотид представляет собой **G**, то соответствующий столбец $\mathbf{X_i}$ будет $\mathbf{(0,0,1,0)}^{\mathbf{T}}$. Строки и столбцы как $\mathbf{Y_i}$, так и $\mathbf{X_i}$ представляют каналы и циклы соответственно. Cross-talk, phasing/prephasing и signal decay моделируются матрицами **M**, **P** и **D**

соответственно. Матрица $\mathbf{E_i}$ предназначена для представления статистической ошибки.

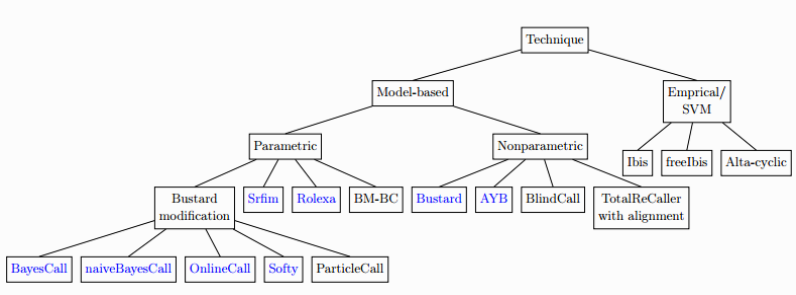

Рис. 5. Различные методы моделирования base-calling, используемые в настоящее время

Рассмотрим уравнение (1) с матрицей **M**:

$$M = \begin{pmatrix} 1 & m_{12} & m_{13} & m_{14} \\ m_{21} & 1 & m_{23} & m_{24} \\ m_{31} & m_{32} & 1 & m_{34} \\ m_{41} & m_{42} & m_{43} & 1 \end{pmatrix}, \quad (2)$$

где каждый элемент $\mathbf{m_{rs}}$ матрицы перекрестных помех **M** указывает величину наблюдаемой интенсивности в канале s, генерируемую сигналом от нуклеотида **r**; для каждого **r**, **s** = 1,2,3,4. Элементы **M** оцениваются посредством итеративного процесса для оценки элементов матрицы, предложенного в работе [24]. Необработанные интенсивности флуоресценции рассматриваются как линейные комбинации вклада флуоросцентных красителей и каналов. Подход состоит в том, чтобы оценивать каждый элемент $\mathbf{m_{rs}}$ и $\mathbf{m_{sr}}$, сначала рассматривая только эти каналы, **r** и **s**. Производится разбиение множества точек на бины по квантилям. Для тех значений, чей **r**-й компонент попадает в данный интервал квантилей, возьмем пару интенсивностей в каналах, имеющую минимальное значение в **s**-компоненте. По этим выбранным парам точек строится регрессия по норме L1 и получаем оценку наклона кривой. Наклон является оценкой $\mathbf{m_{rs}}$. Поменяем местамикомпоненты интенсивностей и повторим процедуру, чтобы получить оценку $\mathbf{m_{sr}}$. Сделаем это для всех пар **r** и **s** с $\mathbf{r \neq s}$, чтобы получить оценки элементов матрицы cross-talk (2).

Phasing and prephasing моделируются с помощью матрицы вероятности перехода **Q (J x J),** отслеживающей положение

обратимых терминаторов. Напомним, что Phasing происходит, когда позиция терминатора отстает, а prephasing происходит, когда позиция терминатора опережает небольшую долю в ДНК составе шаблона. Таким образом, элементы Q в терминах **u** (текущей) и **v** (следующей) позиции терминатора, могут быть смоделированы как:

$$Q_{uv} = \begin{cases} p & v = u \\ 1-p-q & v = u+1 \\ q & v = u+2 \\ 0 & \textit{в ином случае} \end{cases}, \qquad (3)$$

где **p** – вероятность phasing, **q** – вероятность prephasing, а **1 – p – q** – вероятность нормального включения нуклеотида. Поскольку считается, что phasing и prephasing происходит не только в границах одного цикла от текущего цикла, и что эти эффекты могут сохраняться в течение нескольких циклов до и после, рассматривается матрица вероятностей **t**-шагового перехода, $\mathbf{Q}^{\mathbf{t}}$. При этом, **(u, v)**-й элемент $\mathbf{Q}^{\mathbf{t}}$ представляет вероятность того, что в данной цепочке ДНК шаблона в цикле **u** переместится в цикл **v** после **t** циклов. Таким образом, **(v, t)** -й элемент матрицы phasing / prephasing уравнения (1) принимает вид:

$$p * P_{v-1,t} + (1-p-q) * P_{v-1,t-1} + q * P_{v-1,t-2}, \qquad (4)$$

для **v = 2, ..., J** и **t = 1, 2, ..., J** с $\mathbf{P_n = 1 - p - q}$, $\mathbf{P_{12} = q}$, а остальные столбцы первой строки равны 0. Параметры **p** и **q** оцениваются по возрастающей корреляции интенсивностей в течение первых нескольких циклов секвенирования. Потери сигнала оцениваются диагональной матрицей:

$$\begin{gathered} D = \left[diag\left(\frac{\overline{W_1}}{\overline{W_1}}, \frac{\overline{W_1}}{\overline{W_2}}, ..., \frac{\overline{W_1}}{\overline{W_j}} \right) \right]^{-1}, \\ \overline{W_j} = \sum\nolimits_{i=1}^{N} (Y'_{ijA} + Y'_{ijC} + Y'_{ijG} + Y'_{ijT}), \\ Y'_{ij} = M^{-1} Y_{ij}, \end{gathered} \qquad (5)$$

где диагональные элементы представляют собой результат перенормировки концентраций путем взятия среднего значения интенсивностей с поправкой на перекрестные помехи и использования его в качестве нормализующего фактора.

5. Машинное обучение в задаче base-calling. Введение в задачи ДНК секвенирования прикладного машинного обучения (machine learning, ML) включает в себя создание и оценку моделей, использующих алгоритмы, способные распознавать, классифицировать и прогнозировать определенные результаты на основе данных. Подходы ML принимают различные формы, включая обучение без учителя (unsupervised), обучение с частичным привлечением учителя (semi-supervised), обучение с учителем (supervised) (рисунок 6) [25]. Например, часто целью supervised ML, применяемого к данным секвенирования, является построение правила принятия решения (т.е. модели) из набора собранных наблюдений для прогнозирования метки ответа немеченого образца с использованием набора измерений. Входные переменные часто при этом называют признаками (features), а соответствующие выборки – наблюдениями (observations).

Принципиальное различие между unsupervised (USML) и supervised машинным обучением (SML) заключается в том, что в USML образцы разделяются с использованием функций без какой-либо ссылки на метки ответов и строится прогноз к какому кластеру может принадлежать ответ, тогда как SML строит гиперповерхности, которая разделяет базовое векторное пространство на наборы, по одному для каждого класса [26] (рисунок 7).

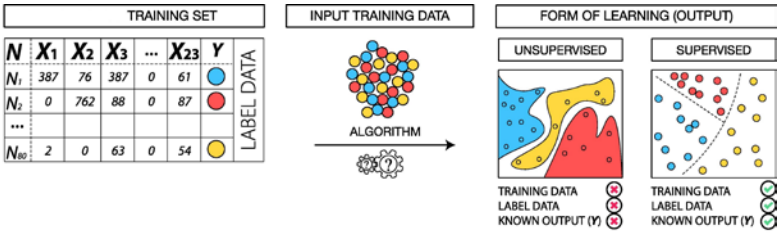

Рис. 6. Схематическое представление unsupervised и supervised форм ML и нескольких конкретных методов, предсказывающих три метки условного ответа (синий/красный/желтый) [25]. Рисунок изображает общую матрицу, содержащую наблюдения или образцы (**N**), признаки (**X1, …, X23**) и метки нескольких классов (**Y**). Входные данные обрабатываются, чтобы либо предсказать к какому множеству принадлежит то или иное наблюдение (unsupervised), либо найти наилучшую границу, разделяющую множества (supervised). При этом следует понимать отличие множества от класса, хотя оба содержат объекты, близкие по своим свойствам. Классы и их свойства задаются априорно, в то время как множества формируются исключительно на основе близости значений признаков объектов, а свойства выясняются в процессе их содержательной интерпретации

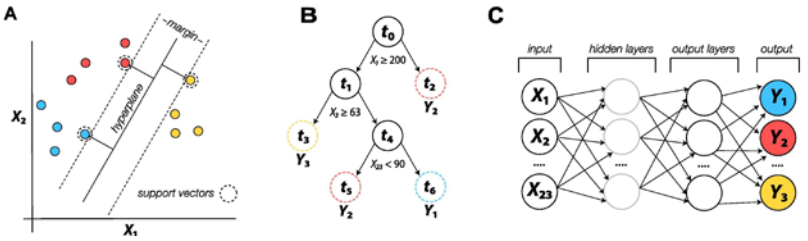

Рис. 7. Различные методы машинного обучения [25]. (A) Линейный классификатор метода опорных векторов SVM, демонстрирующий разделение между метками классов, где гиперплоскость максимизирует расстояние между ближайшими точками на множестве данных для обучения (training set). Опорные векторы относятся к трем векторам положения, проведенным из начала точек выборки (пунктирная окружность) с целью максимизации расстояния между оптимальной гиперплоскостью и опорными векторами с целью проведения границы решения. (B) Дерево решений, построенное для классификации выборок в Y на основе значений входных признаков. Деревья начинаются с корневого узла (t0) и растут до различных листовых узлов (заштрихованные круги), чтобы закончиться конечным узлом, так что агрегированные прогнозы по конечным узлам усредняются по k-деревьям для наилучших прогнозов $\hat{Y}$. (C) Нейронная сеть, отображающая структуру последовательных слоев. Входные значения X передаются на следующий скрытый уровень, который передает взвешенные соединения на выходной уровень для предсказания $\hat{Y}$

5.1. Обучение без учителя (USML). Методы обучения без учителя часто используются для первоначального исследовательского анализа многомерных данных секвенирования и для выработки гипотез с целью последующего анализа, поскольку они помогают в визуализации и прояснении структуры данных, которые не имеют заранее определенных меток ответов, присвоенных наблюдениям. Эти методы работают с целью выявления однородных подгрупп путем кластеризации данных или для обнаружения аномалий путем поиска закономерностей с помощью методов уменьшения размерности (DR).

5.1.1. Метод k-средних (K-means clustering). Цель метода k-средних [27] состоит в том, чтобы сгруппировать выборки в определенное количество (**k**) непересекающихся подгрупп (кластеров) с использованием расстояний, рассчитанных между объектами, чтобы каждая точка данных принадлежала только к одной группе. Этот метод назначает точки данных кластеру таким образом, чтобы сумма квадратов расстояний между точками данных и центроидом (среднее значение всех точек данных, представленных геометрическим центром кластера) была минимизирована. За счет уменьшения

внутрикластерных вариаций точки данных располагаются так, чтобы построить кластер, который принимает сферическую форму, окружающую центр тяжести, и это позволяет различным подгруппам данных оставаться как можно дальше друг от друга. Недостатком K-средних является то, что они не могут хорошо строить кластеры на точках данных, которые образуют данные более сложной формы нежели круговые. Дополнительным ограничением является то, что требуется предварительно определить определенное количество кластеров, что иногда является серьезным ограничением.

5.1.2. Метод главных координат (Principal Coordinates Analysis, PCoA). В анализе PCoA [28] данные разбиваются на компоненты, чтобы максимизировать линейную корреляцию между точками данных в матрице различий, задаваемых входными признаками. Посредством «преобразования координат» количество точек данных x заменяется новыми полученными координатами y, что снижает размерность набора данных за счет отбрасывания координат, которые могут не удовлетворять заданному порогу дисперсии данных секвенирования. Этот метод сохраняет глобальную структуру данных, проецируя их на пространство с меньшим количеством измерений. Точки размещаются так, чтобы попарные расстояния между ними в новом пространстве как можно меньше отличались от эмпирически измеренных расстояний в пространстве признаков изучаемых объектов. Метод главных координат PCoA, или многомерное шкалирование (MDS, multidimensional scaling), во многом похож на метод главных компонент PCA [29] -, но вместо корреляционной матрицы выполняет вычисление собственных значений и собственных векторов симметричной матрицы расстояний. Так можно компенсировать некоторые отклонения от предпосылок в отношении статистического распределения данных, принятых для корреляционного анализа, но одновременно возникает проблема выбора подходящей метрики дистанции.

5.1.3. Стохастическое вложение соседей с t-распределением (t-distributed Stochastic Neighbor Embedding, t-SNE). Стохастическое вложение соседей с t-распределением — это метод машинного обучения визуализации данных, разработанный Лоренсом ван дер Маатеном и Джефри Гинтоном [30]. Это удобный метод нелинейного снижения размерности путем вложения многомерных данных в двух- или трехмерное пространство для дальнейшей визуализации. В частности, он отображает каждую точку многомерного пространства в двух или трехмерную точку евклидового пространства так, что подобные объекты располагаются рядом, а непохожие объекты

соответствуют удаленным точкам с высокой вероятностью. Алгоритм t-SNE состоит из двух основных этапов. Первоначально, t-SNE создает распределение вероятностей по парам многомерных объектов таким образом, что подобные объекты имеют высокую вероятность быть выбранными, в то время как непохожие точки имеют очень малую вероятность быть выбранными вместе. Далее, t-SNE определяет подобное распределение вероятностей для точек в карте низкомерного пространства и минимизирует разногласия по расстоянию Кульбака-Лейблера между двумя распределениями по месту расположения точек на карте.

5.2. Машинное обучение с учителем (SML). Машинное обучение с учителем (SML) — это более сложная форма изучения наборов данных секвенирования, поскольку, в отличие от неконтролируемых методов, метки ответов (**Y**) назначаются каждому образцу в наборе данных, группируя их в значимые категории. При этом более целенаправленное исследование данных может быть достигнуто, поскольку модель обучается на специальном наборе признаков (**X**) (training set) для создания правил, в которых они могут служить предикторами явлений или результатов. Другими словами, с помощью некоторой функции происходит установление соответствие между набором признаков (**X**) и метками ответов (**Y**). После обучения эта модель может принимать новые немаркированные образцы с аналогичными признаками (testing set) и прогнозировать их результат (**Y**) на основе того, что она узнала из обучающего набора. SML можно использовать с непрерывными числовыми выходными данными или категориальными выходными данными. В следующем разделе представлен обзор некоторых из наиболее распространенных алгоритмов SML для задач прогнозирования на основе данных секвенирования.

5.2.1. Случайные леса (Random Forests, RF). Случайные леса [31] широко используются для решения различных задач биоинформатики. Этот метод строит несколько лесов, состоящих из деревьев решений, используя информацию, содержащуюся во входных функциях, для последовательного разделения выборок на основе присвоенных им значений (**Y**). Леса управляются начальной загрузкой и критерием разделения узлов, который использует информацию, содержащуюся в случайном подмножестве признаков. Тот факт, что в каждом лесу строятся сотни или тысячи деревьев решений с использованием подмножества как выборок, так и признаков, позволяет получить совокупное среднее значение прогнозов, сделанных в каждом конечном узле. Таким образом, RF является

идеальной основой для последовательного выявления «истинных эффектов» в сложных и разнородных данных. Дополнительными факторами, которые делают RF привлекательным на практике, является то, что они являются готовыми, поддающимися вычислительной обработке и высокопроизводительными классификаторами, устойчивыми к выбросам, зашумленным и нелинейным данным и ошибкам в метках ответов. Метод RF менее подвержен эффекту переобучения (overfitting), чем другие методы SML, что способствует его привлекательности.

5.2.2. Градиентный бустинг (Gradient Boosting, GB). Градиентный бустинг [26] или метод повышения градиента, когда он используется для деревьев решений, представляет собой ансамблевый метод, в котором используется процесс повышения (бустинг) для последовательного объединения отдельных алгоритмов обучения (деревьев решений) с целью получения более удачного решения (learner). Деревья с усилением градиента отличаются от деревьев RF тем, что каждое дерево решений строится последовательно в попытке уменьшить ошибки предыдущего дерева, а не параллельно. Кроме того, каждое дерево, построенное в GB, имеет фиксированный размер и соответствует исходным данным, а не выборкам начальной загрузки, как это делается в RF. Подобно RF, можно использовать как числовые, так и категориальные признаки, но на практике может быть сложнее найти оптимальные параметры настройки для хорошей подгонки модели, такие как количество оценщиков дерева.

5.2.3. Метод опорных векторов (Support Vector Machines, SVM). Целью метода SVM [32] является нахождение наилучшего обобщенного линейного разделения меток ответа (Y) гиперплоскостью, которая максимизирует разницу между различными значениями Y (или каждым классом) в помеченных метками данных. Граница решения находится таким образом, что каждый класс отделяется при максимально возможном расстоянии от ближайших выборок (называемых опорными векторами и определяющими эту границу решения). SVM относятся к категории линейных дискриминантных методов SML. Алгоритм основан на допущении, что чем больше расстояние между параллельными разделяющими гиперплоскостями, тем меньше будет средняя ошибка классификатора. Эти модели могут работать с различными типами объектов, но по своей природе их трудно интерпретировать, поскольку они не дают прямых оценок вероятности в полученной оценке.

5.2.4. Логистическая регрессия. В отличие от обычной регрессии метод логистической регрессии [33] не предсказывает

значение числовой переменной на основе выборки начальных значений. Вместо этого значение функции представляет собой вероятность того, что данное исходное значение принадлежит определенному классу. Основная идея логистической регрессии заключается в том, что пространство начальных значений может быть разделено линейной границей (т.е. прямой) на две области, соответствующие классам. При этом оптимально использовать регуляризацию, как метод, используемый для уменьшения переобучения. Гребневая регрессия или ридж-регрессия [34] удовлетворяет модели, которая уменьшает дисперсию без увеличения систематической ошибки и это достигается путем наложения ограничений на сложность параметров. Этот метод добавляет штрафной член к функции потерь, позволяя ограничить сложность параметра. Гребневая регрессия может использоваться как для классификации, так и для регрессии, но может быть весьма требовательной к вычислительным ресурсам в случае большого входного пространства признаков.

5.2.5. Нейронные сети (Neural Networks). Нейронные сети [35] используют архитектуру построения иерархической модели, в которой несколько структурированных сетей взаимосвязанных узлов (нейронов) строятся с весами, соответствующими каждому ребру сети, чтобы обеспечить сопоставление входных данных **X** с ответами **Y**. Сети связаны между собой через механизм распространения с прямой связью (feed-forward propagation), где каждый нейрон получает входные данные от предыдущих нейронов. Сеть начинается с входных слоев, которые связаны с каждым нейроном в одном или нескольких скрытых слоях, которые используют алгоритм обратного распространения (backpropagation algorithm) для максимизации весов, размещенных на каждом нейроне, для улучшения прогнозирования. Этот процесс является итеративным, когда последний скрытый слой встречается с выходным слоем для получения прогнозируемого вывода ответа (**Y**). Нейронные сети очень перспективны в своей способности идентифицировать сложную структуру в многомерных и сложных наборах данных. Нейронные сети часто называют методами «черного ящика», поскольку бывает сложно интерпретировать способы принятия решения.

6. Применение методов машинного обучения для решения задачи base-calling. Для применения различных алгоритмов машинного обучения используется платформа Scikit-learn [36], поддерживая простой в использовании интерфейс, тесно интегрированный с языком Python. Основным объектом при этом

является estimator, который реализует метод fit, отвечающий за подгонку модели на предоставленной обучающей выборке. Классы модели, принадлежащие классу методов обучения с учителем, например, SVM, также реализуют метод predict, обеспечивающий составление прогноза по обучающей выборке. Кроме того, некоторые наследники класса estimator, т.н. transformer-классы (как, например, PCA) имеют метод transform, который позволяет менять входные данные модели. Класс estimator в общем случае реализует метод score, который позволяет посчитать показатель выбранной метрики качества. Другим важным классом Scikit-learn является cross-validation iterator, предоставляющий разнообразные методы скользящего контроля.

Последним этапом процесса секвенирования NGS является анализ изображения, суть которого состоит в идентификации кластеров с последующим проведением операции base-calling. В работе [5] подробно описывались подходы для выполнения задачи обнаружения кластеров. Положения и оценённые радиусы такого кластера затем используются для извлечения ряда характеристик из каждого кластера и его непосредственного фона. Эти функции затем используются в качестве атрибутов, то есть служат входными данными для алгоритмов машинного обучения. Извлечение признаков выполняется путем изучения интенсивности света каждого пикселя определенной прямоугольной области кластера и некоторого окружающего фона. Затем вычисляются восемь статистических показателей: четыре для фона (background) и четыре для с сердцевины кластера (foreground). Этот процесс показан на рисунке 8.

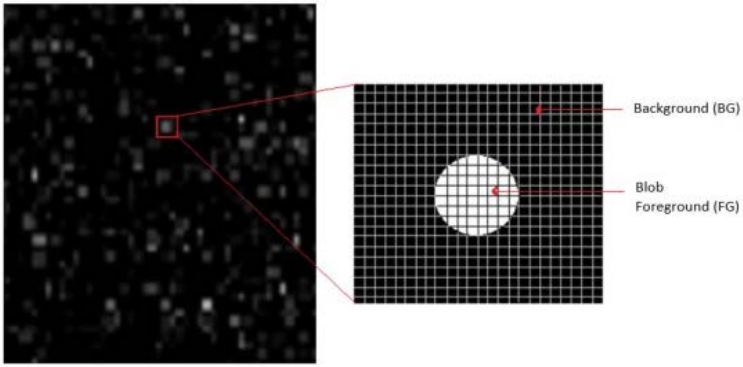

Рис. 8. Отбор признаков из “сырых” данных секвенирования NGS [37]

Извлекаются следующие статистические показатели: для фона (background) - max, mean, median и mode, для центральной зоны кластера (foreground) - max, mean, pct90 и pct99, где max — максимальное значение интенсивности, mean — среднее арифметическое значение, mode — наиболее часто встречающееся значение и pct90, и pct99 — 90-й и 99-й процентили соответственно. Каждый набор изображений, используемых для процесса base-calling одного цикла, состоит из четырех изображений, по одному для каждого канала флуоресценции. Поскольку из каждого такого изображения извлекается восемь статистических показателей, для каждого кластера имеется в общей сложности тридцать два статистических показателя, которые используются в качестве атрибутов в алгоритмах машинного обучения.

Одним из наиболее фундаментальных рабочих принципов секвенирования NGS является распараллеливание, что означает, что ДНК фрагментируется на короткие риды, которые секвенируются одновременно. При этом в каждом кластере будут разные фрагменты ДНК. Однако для целей данного проекта машинного обучения мы используем заранее секвенированные последовательности, предварительно отображенные (мэппированные) на известные референтные последовательности. Иногда это достаточно короткие индексные последовательности, короткие последовательности ампликонов (искусственно синтезированные известные последовательности нуклеотидов), иногда – короткие риды из ДНК последовательности бактериофага Phix174. Это позволяет сформировать обучающую выборку с известным нуклеотидным составом коротких прочтений.

После того, как данные надлежащим образом исследованы и преобразованы (в частности, нормализованы), набор данных необходимо случайным образом разделить на две части: обучающий и тестовый набор. Обычно около восьмидесяти процентов, выбирается в качестве обучающих данных, а оставшиеся данные составляют тестовый набор.

7. Результаты машинного обучения в задаче base-calling. Различные методы машинного обучения были апробированы на разных наборах данных секвенирования, таких как, набор индексов, небольшие ампликоны, данные секвенирования эталонного референтного генома Phix174. Обрабатываемые данные были получены на приборе MiSeq фирмы Illumina (США) и на опытном образце отечественного прибора «Нанофор СПС».

Для машинного обучения на данных секвенирования использовались следующие методы:

- логистическая регрессия (logit model) с различными видами регуляризации;
- метод опорных векторов (support vector machine);
- деревья принятия решений (decision tree);
- случайные леса (random forest);
- ансамблевые методы, в частности, бэггинг классификаторы с голосованием.

Типичный отчет со сравнением точности предсказания правильной буквы в нуклеотидной последовательности выглядит следующим образом (где box-plot диаграмма в удобной форме показывает медиану, среднее, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы) (рисунок 9) [38].

В нашем случае каждый из четырех каналов для определения нуклеотидов имеет 8 признаков. Четыре на полезный сигнал и четыре на фон. Соответствующая box-plot диаграмма для 32 признаков (признаки 0 – 7 для канала A, 8 – 15 для канала C, признаки 16 – 23 для канала G, признаки 24 – 31 для канала T) представлена на рисунке 10. Последовательность признаков для каждого канала: foreground - max, mean, pct90 и pct99; background - max, mean, median и mode.

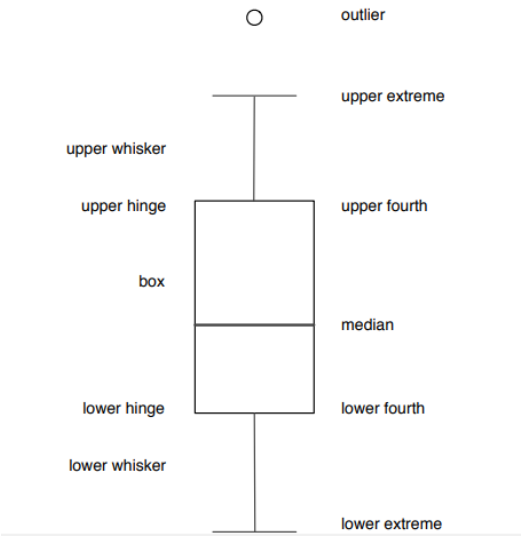

Рис. 9. Структурная схема box-plot диаграммы

Вероятности правильного предсказания для данных, полученных на приборе MiSeq фирмы Illumina (США) и на опытном образце отечественного прибора «Нанофор СПС оказались примерно одинаковыми. Наибольшую вероятность правильного предсказания показал метод логистической регрессий. Для логистической регрессии для повышения точности предсказания на тестовом наборе была использована L1-регуляризация (lasso regularization).

Распределение распределение всех нуклеотидов по каналам для образца Phix 174 примерно одинаковое и приблизительно равно 25%, что очевидно из рисунка 10. Диаграммы типа box-plot, подобные представленному на рисунке 10, дают наглядное представление о совокупности распределения интенсивностей кластеров и шумов в ближайшем окружении кластеров по всем каналам нуклеотидов и позволяют выбрать набор признаков, отвечающих за разделение интенсивностей в кластерах по каналам.

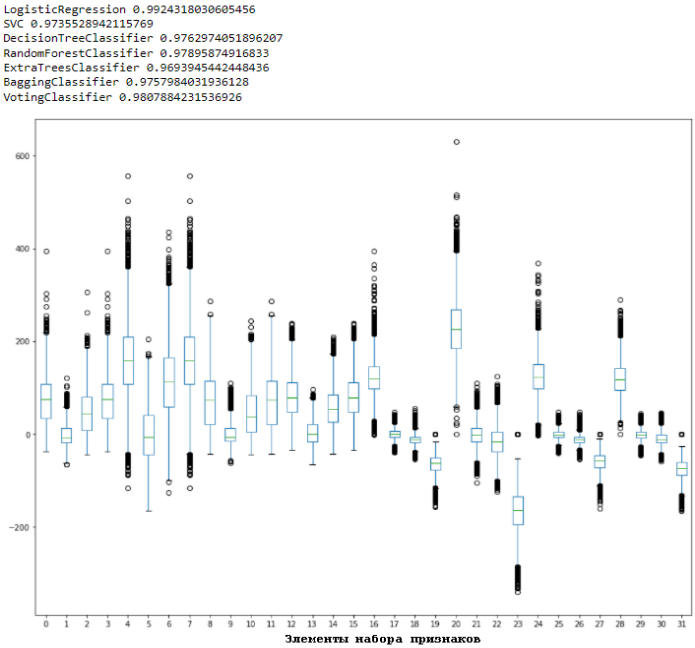

Рис. 10. Типичный отчет со сравнением точности предсказания методами машинного обучения с сооответствующей box-plot диаграммой. В левом верхнем углу представлены вероятности правильного предсказания для различных методов машинного обучения. По оси x представлен набор признаков

Прогноз, полученный методом Decision Tree, позволяет путем анализа узлов дерева решений выделить наиболее значащие для прогнозирования признаки и снизить количество обрабатываемой информации.

В частности, такой анализ позволил для большинства методов машинного обучения снизить количество признаков для прогнозирования с 32 до 8, ограничившись только двумя первыми характеристиками для foreground и background. При этом точность прогноза даже повысилась.

Анализ box-plot диаграмм отдельно для каждого канала изображений и для пар каналов, вносящих основной вклад в эффект cross-talk позволяет судить о тех комбинациях признаков, которые определяют работу методов обучения.

Наконец, ожидаемый выигрыш от применения ансамблевых методов пока не был получен либо из-за неоптимального выбора соответствующих гиперпараметров методов, либо просто из-за недостаточности объема выборки для обучения.

Таким образом, использование методов машинного обучения в задаче построения последовательностей нуклеотидов на основе измеренных интенсивностей сигналов флуоресценции показало возможность получения достоверных результатов без использования традиционных процедур первичного анализа данных, таких как коррекция влияния перекрестных помех (cross-talk), фильтрация амплитуд кластеров на основе сравнения интенсивностей сигналов в различных каналах (purity и chastity) и других.

8. Выводы.

1. Приведенный обзор методов машинного обучения для обработки информации при построении последовательностей нуклеотидов на основе статистистической модели base-calling позволяет раскрыть вычислительные алгоритмы этой стадии обработки данных в приборах для генетического анализа. В виду особенностей задачи base-calling (типичная задача классификации по четырем классам нуклеотидов) особо перспективным кажется применение таких методов машинного обучения, как логистическая регрессия и нейронные сети. Для выделения особо значимых признаков перспективны также методы на основе деревьев решений (Decision Tree) и метод главных компонент.

2. Данный обзор является одним из первых в русскоязычной литературе, в котором обсуждаются вычислительные алгоритм обработки информации на этапе base-calling и особенности их реализации. В тоже же время многократно возросшие вычислительные

мощности раскрывают новые горизонты применения искусственного интеллекта.

3. Приведенные результаты применения разнообразных моделей машинного обучения к решению задачи base-calling показали, что все они на тестовой выборке дают результаты построения последовательностей нуклеотидов, совпадающие на 97…99.5% с последовательностями нуклеотидов в референтных геномах. При этом обучающая выборка была достаточно ограниченной по объему.

4. Ожидается, что дальнейшее применение более объемной обучающей выборки даст еще более лучшие результаты в решении задачи base-calling. На следующей стадии обработки важное значение имеют алгоритмы для оценки показателей качества определения как отдельных нуклеотидов, так и ридов.

5. Методы машинного обучения типа деревьев поиска и уменьшение размерности пространства признаков, по которым проводится обучение позволяет выделить параметры, определяющие вероятность статистической ошибки при высокопроизводительном секвенировании.

6. Безусловно, полезным при накоплении результатов секвенирования на отечественном полногеномном секвенаторе является сравнение результатов машинного обучения на этапе base-calling с данными, полученными на секвенаторе Illumina.

7. Ожидается, что дальнейшее применение более объемной обучающей выборки даст еще более лучшие результаты в решении задачи base-calling. На следующей стадии обработки важное значение имеют алгоритмы для оценки показателей качества определения как отдельных нуклеотидов, так и ридов [38]. Методы машинного обучения типа деревьев поиска и уменьшение размерности пространства признаков, по которым проводится обучение, позволяют выделить параметры, определяющие вероятность статистической ошибки при высокопроизводительном секвенировании.

Для успешного внедрения рассмотренной практики машинного обучения предполагается стадия обучения при известной циклограмме эксперимента на некотором эталонном образце с известным геномным составом (например, Phix174). Далее полученная обученная модель может быть использована для решения задачи base-calling уже на любых образцах при сходных циклограммах эксперимента. Поскольку для обучения в качестве обучающих признаков взяты параметры изображений в одном цикле, предполагается, что такая модель сможет учесть эффекты decay и cross-talk. При включении в модель признаков

предыдущих и последующих циклов возможно учесть эффекты phasing/prephasing.

Литература

1. Бородинов А. Г., Манойлов В. В., Заруцкий И. В., Петров А. И., Курочкин В. Е. Поколения методов секвенирования ДНК (ОБЗОР) // Научное приборостроение. 2020. т. 30. № 4. С. 3—20
2. Wenxiu Ma, Wing Hung Wong The analysis of ChIP-Seq data // Methods Enzymol. 2011. vol. 497. pp. 51-73.
3. Zhong Wang, Mark Gerstein, Michael Snyder RNA-Seq: a revolutionary tool for transcriptomics // Nat Rev Genet. 2009. vol.10. no. 1. pp. 57-63.
4. Syed, F., Grunenwald, H. & Caruccio, N. Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition // Nat Methods. 2009. vol. 6. pp. i–ii.
5. Манойлов В. В., Бородинов А. Г., Заруцкий И. В., Петров А. И., Курочкин В. Е. Алгоритмы обработки сигналов флуоресценции массового параллельного секвенирования нуклеиновых кислот // Труды СПИИРАН. 2019. т. 18. № 4. С. 1010–1036.
6. Schilbert H.M., Rempel A., Pucker B. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data // Plants. 2020. vol. 9. p. 439.
7. Ye C., Hsiao C., Corrada-Bravo H. BlindCall: ultra-fast base-calling of high-throughput sequencing data by blind deconvolution // Bioinform. 2014. vol. 30. no. 9. pp. 1214–1219.
8. Wang B, Wan L, Wang A, Li L.M. An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters // Sci Rep. 2017. vol. 7.
9. Renaud G., Kircher M., Stenzel U., Kelso J. FreeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers // Bioinformatics. 2013. vol. 29. pp. 1208–1209.
10. Das S., Vikalo H. Base calling for high-throughput short-read sequencing: dynamic programming solutions // BMC Bioinformatics. 2013. vol. 14. p. 129.
11. Massingham T., Goldman N. All your base: a fast and accurate probabilistic approach to base calling // Genome Biol. 2012. vol. 13. p. R13.
12. Das S., Vikalo H. OnlineCall: fast online parameter estimation and base calling for illumina's next-generation sequencing // Bioinformatics. 2012. vol. 28. no. 13. pp. 1677–1683.
13. Ji Y., Mitra R., Quintana F., Jara A., Mueller P., Liu P., Lu Y., Liang S. BM-BC: a Bayesian method of base calling for Solexa sequence data // BMC Bioinformatics. 2012. vol. 13. p. S6.
14. Shen X., Vikalo H. ParticleCall: A particle filter for base calling in next-generation sequencing systems // BMC Bioinformatics. 2012. vol. 13. p. 160.
15. Menges F., Narzisi G., Mishra B. TotalReCaller: improved accuracy and performance via integrated alignment and base-calling // Bioinformatics. 2011. vol. 27. no. 17. pp. 2330-2337.
16. Kao W.C., Song Y.S. naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing // J Comput Biol. 2011. vol.18. no. 3. pp. 365-377.
17. Corrada-Bravo H., Irizarry R.A. Model-based quality assessment and base-calling for second-generation sequencing data // Biometrics. 2009. vol. 3. pp. 665–674.
18. Kao W.C., Stevens K., Song Y.S. BayesCall: a model-based basecalling algorithm for high-throughput short-read sequencing // Genome Res. 2009. vol. 19. pp. 1884–1895.

19. Kircher M., Stenzel U., Kelso J. Improved base calling for the Illumina Genome analyzer using machine learning strategies // Genome Biol. 2009. vol. 10. pp. R83.1–.9.
20. Rougemont J., Amzallag A., Iseli C. Probabilistic base calling of Solexa sequencing data // BMC Bioinformatics. 2008. vol. 9. p. 431.
21. Erlich Y., Mitra P.P., Delabastide M., et al. Alta-cyclic: a self-optimizing base caller for next-generation sequencing // Nat Methods. 2008. vol. 5. pp. 679–682.
22. Зубов В. В., Чемерис Д. А., Василов Р. Г., Курочкин В. Е., Алексеев Я. И. Краткая история методов высокопроизводительного секвенирования нуклеиновых кислот // Биомика. 2021. т. 13. № 1. С. 27-46.
23. Cacho A. Base-Calling of High-Throughput Sequencing Data Using a Random Effects Mixture Model // UC Riverside. 2016. 91 p.
24. Li L., Speed T. An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing // Electrophoresis. 1999. vol. 20. pp. 1433–1442.
25. Ghannam R., Techtmann S. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring // Computational and Structural Biotechnology Journal. 2021. vol. 19. pp. 1092-1107.
26. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction // Springer Science & Business Media. 2009. 745 p.
27. Forgy E.W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications // Biometrics. 1965. vol. 21. pp. 768–769.
28. Mohammadi S.A., Prasanna B.M. Review and Interpretation Analysis of Genetic Diversity in Crop Plants —Salient Statistical Tools // Crop Science. 2003. vol. 43. pp. 1235-1248.
29. Jackson J.E. A User's Guide to Principal Components // John Wiley & Sons. 1991.
30. Van der Maaten L., Hinton G. Visualizing Data using t-SNE // Journal of Machine Learning Research. 2008. vol. 9. pp. 2579-2605.
31. Breiman L. Random forests // Machine Learn. 2001. vol. 45. no. 1. pp. 5–32.
32. Suykens J.A., Vandewalle J. Least squares support vector machine classifiers // Neural Process Letters. 2004. vol. 9. no. 3. pp. 293–300.
33. Tolles J, Meurer W.J. Logistic Regression: Relating Patient Characteristics to Outcomes // JAMA. 2016. vol. 316. no. 5. pp. 533-534.
34. Hoerl A.E., Kennard R.W. Ridge regression: biased estimation for nonorthogonal problems // Technometrics. 1970. vol. 12. no. 1. pp. 55–67.
35. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. 2015. vol. 521. pp. 436–444.
36. About us — scikit-learn 0.20.1 documentation. URL: https://scikit-learn.org. (дата обращения 18.03.2022).
37. Tegfalk E. Application of machine learning techniques to perform base-calling in next-generation DNA sequencing // KTH, SCI. 2020.
38. Wickham H., Stryjewski L. 40 years of boxplots. URL: https://vita.had.co.nz/papers/boxplots.pdf. (дата обращения 23.03.2022).

Бородинов Андрей Геннадьевич — канд. физ.-мат. наук, начальник, сектор информационных проектов, АО "Научные приборы". Область научных интересов: математическая статистика, проблемы анализа, обработки и представления данных, искусственный интеллект. Число научных публикаций — 15. borodinov@gmail.com; улица Ивана Черных, 31-33, 198095, Санкт-Петербург, Россия; р.т.: +7(911)212-0895.

Манойлов Владимир Владимирович — д-р техн. наук, доцент, заведующий лабораторией, лаборатория автоматизации измерений и цифровой обработки сигналов, Институт аналитического приборостроения Российской академии наук (ИАП РАН).

Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 76. manoilov-vv@mail.ru; улица Ивана Черных, 31-33, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0750.

Заруцкий Игорь Вячеславович — канд. техн. наук, старший научный сотрудник, лаборатория автоматизации измерений и цифровой обработки сигналов, Институт аналитического приборостроения Российской академии наук (ИАП РАН). Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 45. igorzv@yandex.ru; улица Ивана Черных, 31-33, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0720.

Петров Александр Иванович — канд. техн. наук, заведующий сектором электроники и программного обеспечения, лаборатория методов и приборов иммунного и генетического анализа, Институт аналитического приборостроения Российской академии наук (ИАП РАН). Область научных интересов: представление и обработка сигналов и изображений в аналитических приборах. Число научных публикаций — 25. fataip@mail.ru; улица Ивана Черных, 31-33, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0720.

Курочкин Владимир Ефимович — д-р техн. наук, профессор, заведующий лабораторией, лаборатория методов и приборов иммунного и генетического анализа, Институт аналитического приборостроения Российской академии наук (ИАП РАН). Область научных интересов: исследования и оптимизация электромиграционных методов анализа, развитие аналитических методик для капиллярного электрофореза, исследование оп-тических методов детектирования, разработка методов и приборов для ДНК анализа, разработка методик подготовки проб и специализированных реактивов. Число научных публикаций — 210. lavrovas@yandex.ru; улица Ивана Черных, 31-33, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0719.

Сараев Алексей Сергеевич — инженер 2 категории, лаборатория методов и приборов иммунного и генетического анализа, Институт аналитического приборостроения Российской академии наук (ИАП РАН). Область научных интересов: моделирование процессов в аналитических приборах. Число научных публикаций — 1. alex.niispb@yandex.ru; улица Ивана Черных, 31-33, 198095, Санкт-Петербург, Россия; р.т.: +7(812)363-0720.

Поддержка исследований. Работа выполнена в рамках государственного задания Министерства науки и высшего образования номер гос. регистрации 122032300337-4 от 23.03.22.

DOI 10.15622/ia.21.3.5

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

MACHINE LEARNING IN BASE-CALLING FOR NEXT-GENERATION SEQUENCING METHODS

Borodinov A., Manoilov V., Zarutsky I., Petrov A., Kurochkin V., Saraev A. **Machine Learning in Base-Calling for Next-Generation Sequencing Methods.**

Abstract. The development of next-generation sequencing (NGS) technologies has made a significant contribution to the trend of reducing costs and obtaining massive sequencing data. The Institute for Analytical Instrumentation of the Russian Academy of Sciences is developing a hardware-software complex for deciphering nucleic acid sequences by the method of mass parallel sequencing (Nanofor SPS). Image processing algorithms play an essential role in solving the problems of genome deciphering. The final part of this preliminary analysis of raw data is the base-calling process. Base-calling is the process of determining a nucleotide base that generates the corresponding intensity value in the fluorescence channels for different wavelengths in the flow cell image frames for different synthesis sequencing runs. An extensive analysis of various base-calling approaches and a summary of the common procedures available for the Illumina platform are provided. Various chemical processes included in the synthesis sequencing technology, which cause shifts in the values of recorded intensities, are considered, including the effects of phasing / prephasing, signal decay, and crosstalk. A generalized model is defined, within which possible implementations are considered. Possible machine learning (ML) approaches for creating and evaluating models that implement the base-calling processing stage are considered. ML approaches take many forms, including unsupervised learning, semi-supervised learning, and supervised learning. The paper shows the possibility of using various machine learning algorithms based on the Scikit-learn platform. A separate important task is the optimal selection of features identified in the detected clusters on a flow cell for machine learning. Finally, a number of sequencing data for the MiSeq Illumina and Nanofor SPS devices show the promise of the machine learning method for solving the base-calling problem.

Keywords: next-generation sequencing, base-calling, bioinformatics, machine learning.

Borodinov Andrew — Ph.D., Head of sector, Sector of information projects, Scientific Instruments Joint Stock Company. Research interests: mathematical statistics, problems of analysis, processing and presentation of data, artificial intelligence. The number of publications — 15. borodinov@gmail.com; 31-33, Ivana Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(911)212-0895.

Manoilov Vladimir — Ph.D., Dr.Sci., Associate Professor, Head of laboratory, Laboratory of automation of measurements and digital signal processing, Institute for Analytical Instrumentation Russian Academy of Sciences (IAI RAS). Research interests: representation and processing of signals and images in analytical devices. The number of publications — 76. manoilov-vv@mail.ru; 31-33, Ivana Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0750.

Zarutsky Igor — Ph.D., Senior researcher, Laboratory of automation of measurements and digital signal processing, Institute for Analytical Instrumentation Russian Academy of Sciences (IAI RAS). Research interests: representation and processing of signals and images in

analytical devices. The number of publications — 45. igorzv@yandex.ru; 31-33, Ivana Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0720.

Petrov Alexander — Ph.D., Head of the sector of electronics and software, Laboratory of methods and instruments for immune and genetic analysis, Institute for Analytical Instrumentation Russian Academy of Sciences (IAI RAS). Research interests: representation and processing of signals and images in analytical devices. The number of publications — 25. fataip@mail.ru; 31-33, Ivana Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0720.

Kurochkin Vladimir — Ph.D., Dr.Sci., Professor, Head of the laboratory, Laboratory of methods and instruments for immune and genetic analysis, Institute for Analytical Instrumentation of the Russian Academy of Sciences (IAI RAS). Research interests: research and optimization of electromigration analysis methods, the development of analytical methods for capillary electrophoresis, the study of optical methods of detection, the development of methods and instruments for DNA analysis, the development of methods for preparing samples and specialized reagents. The number of publications — 210. lavrovas@yandex.ru; 31-33, Ivana Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0719.

Saraev Aleksey — 2nd category engineer, Laboratory of methods and devices for immune and genetic analysis, Institute for Analytical Instrumentation of the Russian Academy of Sciences (IAI RAS). Research interests: modeling of processes in analytical instruments. The number of publications — 1. alex.niispb@yandex.ru; 31-33, Ivana Chernykh St., 198095, St. Petersburg, Russia; office phone: +7(812)363-0720.

Acknowledgements. This research was performed within the framework of the state number registration 122032300337-4 dated 03/23/22, Ministry of Science and Higher Education of the Russian Federation.

References

1. Borodinov A.G., Manoilov V.V., Zarutsky I. V., Petrov A. I., Kurochkin V. E. [Generations of DNA sequencing methods (REVIEW)]. Nauchnoe priborostroenie - Nauchnoe priborostroenie. 2020. vol. 30. no. 4. pp. 3-20. (In Russ.).
2. Wenxiu Ma, Wing Hung Wong. The analysis of ChIP-Seq data. Methods Enzymol. 2011. vol. 497. pp. 51-73.
3. Zhong Wang, Mark Gerstein, Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009. vol.10. no. 1. pp. 57-63.
4. Syed, F., Grunenwald, H. & Caruccio, N. Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition. Nat Methods. 2009. vol. 6. pp. i–ii.
5. Manoilov V.V., Borodinov A.G., Petrov A.I., Zarutsky I.V., Kurochkin V.E. [Algorithms of processing fluorescence signals for mass parallel sequencing of nucleic acids]. Trudy SPIIRAN - SPIIRAS Proceedings. 2019. vol. 18. no. 4. pp. 1010–1036.
6. Schilbert H.M., Rempel A., Pucker B. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants. 2020. vol. 9. p. 439.
7. Ye C., Hsiao C., Corrada-Bravo H. BlindCall: ultra-fast basecalling of high-throughput sequencing data by blind deconvolution. Bioinform. 2014. vol. 30. no. 9. pp. 1214–1219.
8. Wang B, Wan L, Wang A, Li L.M. An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters. Sci Rep. 2017. vol. 7.

9. Renaud G., Kircher M., Stenzel U., Kelso J. FreeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. Bioinformatics. 2013. vol. 29. pp. 1208–1209.
10. Das S., Vikalo H. Base calling for high-throughput short-read sequencing: dynamic programming solutions. BMC Bioinformatics. 2013. vol. 14. p. 129.
11. Massingham T., Goldman N. All your base: a fast and accurate probabilistic approach to base calling. Genome Biol. 2012. vol. 13. p. R13.
12. Das S., Vikalo H. OnlineCall: fast online parameter estimation and base calling for illumina's next-generation sequencing. Bioinformatics. 2012. vol. 28. no. 13. pp. 1677–1683.
13. Ji Y., Mitra R., Quintana F., Jara A., Mueller P., Liu P., Lu Y., Liang S. BM-BC: a Bayesian method of base calling for Solexa sequence data. BMC Bioinformatics. 2012. vol. 13. p. S6.
14. Shen X., Vikalo H. ParticleCall: A particle filter for base calling in next-generation sequencing systems. BMC Bioinformatics. 2012. vol. 13. p. 160.
15. Menges F., Narzisi G., Mishra B. TotalReCaller: improved accuracy and performance via integrated alignment and base-calling. Bioinformatics. 2011. vol. 27. no. 17. pp. 2330-2337.
16. Kao W.C., Song Y.S. naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. J Comput Biol. 2011. vol.18. no. 3. pp. 365-377.
17. Corrada-Bravo H., Irizarry R.A. Model-based quality assessment and base-calling for second-generation sequencing data. Biometrics. 2009. vol. 3. pp. 665–674.
18. Kao W.C., Stevens K., Song Y.S. BayesCall: a model-based basecalling algorithm for high-throughput short-read sequencing. Genome Res. 2009. vol. 19. pp. 1884–1895.
19. Kircher M., Stenzel U., Kelso J. Improved base calling for the Illumina Genome analyzer using machine learning strategies. Genome Biol. 2009. vol. 10. pp. R83.1–.9.
20. Rougemont J., Amzallag A., Iseli C. Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics. 2008. vol. 9. p. 431.
21. Erlich Y., Mitra P.P., Delabastide M., et al. Altacyclic: a selfoptimizing base caller for next-generation sequencing. Nat Methods. 2008. vol. 5. pp. 679–682.
22. Zubov V.V., Chemeris D.A., Vasilov R.G., Kurochkin V.E., Alekseev Ya.I. [Brief history of high-throughput nucleic acid sequencing methods]. Biomika - Biomics. 2021. vol. 13. no. 1. pp. 27-46. (In Russ.)
23. Cacho A. Base-Calling of High-Throughput Sequencing Data Using a Random Effects Mixture Model. UC Riverside, 2016. 91 p.
24. Li L., Speed T. An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing. Electrophoresis. 1999. vol. 20. pp. 1433–1442.
25. Ghannam R., Techtmann S. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Computational and Structural Biotechnology Journal. 2021. vol. 19. pp. 1092-1107.
26. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009. 745 p.
27. Forgy E.W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965. vol. 21. pp. 768–769.
28. Mohammadi S.A., Prasanna B.M. Review and Interpretation Analysis of Genetic Diversity in Crop Plants —Salient Statistical Tools. Crop Science. 2003. vol. 43. pp. 1235-1248.
29. Jackson J.E. A User’s Guide to Principal Components. John Wiley & Sons, 1991.
30. Van der Maaten L., Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008. vol. 9. pp. 2579-2605.
31. Breiman L. Random forests. Machine Learn. 2001. vol. 45. no. 1. pp. 5–32.

32. Suykens J.A., Vandewalle J. Least squares support vector machine classifiers. Neural Process Letters. 2004. vol. 9. no. 3. pp. 293–300.
33. Tolles J, Meurer W.J. Logistic Regression: Relating Patient Characteristics to Outcomes. JAMA. 2016. vol. 316. no. 5. pp. 533-534.
34. Hoerl A.E., Kennard R.W. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970. vol. 12. no. 1. pp. 55–67.
35. LeCun Y., Bengio Y., Hinton G. Deep learning. Nature. 2015. vol. 521. pp. 436–444.
36. About us — scikit-learn 0.20.1 documentation. Available at: https://scikit-learn.org. (accessed 18.03.2022).
37. Tegfalk E. Application of machine learning techniques to perform base-calling in next-generation DNA sequencing. KTH, SCI, 2020.
38. Wickham H., Stryjewski L. 40 years of boxplots. Available at: https://vita.had.co.nz/papers/boxplots.pdf. (accessed 23.03.2022).