ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2023, том 33, № 2, с. 35–48

## МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ В ПРИБОРОСТРОЕНИИ

УДК 543.07

© В. В. Манойлов, А. Г. Бородинов, А. И. Петров, И. В. Заруцкий, В. Е. Курочкин, 2023

# АЛГОРИТМ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ДЛЯ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ПОСТРОЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ НУКЛЕОТИДОВ В СЕКВЕНАТОРЕ "НАНОФОР СПС"

Развитие информационных технологий и математических методов по обработке данных играет существенную роль для установления различных особенностей в анализируемых нуклеиновых кислотах и тенденций их изменений. Важным этапом в технологии массового параллельного секвенирования нуклеиновых кислот является процесс построения последовательности нуклеотидов по измеренным интенсивностям сигналов флуоресценции. В работе рассматривается алгоритм генерации обучающей выборки, которая используется для построения последовательности буквенных кодов нуклеотидов ДНК по интенсивностям сигналов флуоресценции, полученным непосредственно по результатам обработки изображений. В такие сигналы не были внесены корректировки, связанные с физическими и химическими особенностями проведения процесса секвенирования. В алгоритме использованы метод главных компонент и классификатор, основанный на методе k-средних. С помощью такого классификатора данные после преобразования методом главных компонент разделяются на четыре независимых класса по числу буквенных кодов нуклеотидов ДНК. С помощью обучающей выборки определяется принадлежность вектора, содержащего данные сигналов флуоресценции, одному из классов, а значит, его буквенный код. Алгоритм апробирован на тестовой выборке и показал высокую достоверность результатов.

*Кл. сл.*: секвенирование нуклеиновых кислот, математическая обработка и классификация многомерных данных, метод главных компонент, машинное обучение

## ВВЕДЕНИЕ

Важным элементом успешного развития геномного секвенирования является использование современных информационных технологий и математических методов по обработке данных для установления различных особенностей в анализируемых нуклеиновых кислотах. В Институте аналитического приборостроения РАН разработан аппаратно-программный комплекс (АПК) для расшифровки последовательности нуклеиновых кислот методом массового параллельного секвенирования ("Нанофор СПС").

Решение задачи по расшифровки генома в АПК разделяется на ряд этапов обработки исходных данных. Одним из важных первоначальных этапов обработки данных является оценка значений интенсивностей сигналов флуоресценции для различных длин волн на кадрах изображения проточной ячейки для нескольких циклов секвенирования методом синтеза. Такая оценка выполняется по программам обработки изображений, алгоритмы которых описаны в работах [1, 2]. Финальной частью такого предварительного анализа исходных данных является процесс построения последовательности буквенных кодов нуклеотидов.

Различные химические процессы, включенные в технологию секвенирования методом синтеза, вызывают изменения в значениях регистрируемых интенсивностей, включая эффекты фазирования / префазирования (phasing / prephasing), затухания сигнала и перекрестные помехи (cross-talk).

Применение машинного обучения (machine learning, ML) в задачах секвенирования ДНК включает в себя создание и оценку моделей, использующих алгоритмы, способные распознавать, классифицировать и прогнозировать определенные результаты на основе данных. Подходы ML подразделяются на обучение без учителя (unsupervised), обучение с частичным привлечением учителя (semi-supervised), обучение с учителем (supervised) [3]. Например, часто целью supervised ML, применяемого к данным секвенирования, является построение модели на основе обучающего набора собранных наблюдений с известной последовательностью нуклеотидов с целью прогнозирования нуклеотида для произвольного образца с неизвестным целевым значением типа определяемого нуклеотида [4, 5]. Входные переменные

часто при этом называют признаками (features), а соответствующие выборки — наблюдениями (observations).

В статье [6] был сделан обзор методов машинного обучения для решения задач построения последовательности нуклеотидов и рассмотрены несколько примеров применения машинного обучения для обработки данных секвенатора "Нанофор СПС". Данная работа является продолжением исследований, опубликованных в той статье.

В данной работе на основе метода главных компонент рассмотрен возможный подход машинного обучения (machine learning) для создания и оценки модели, реализующей этап построения последовательности нуклеотидов. На ряде данных секвенирования прибора "Нанофор СПС" показана перспективность метода машинного обучения для решения задачи построения последовательности нуклеотидов.

## ПОСТРОЕНИЕ МАТРИЦЫ ИНТЕНСИВНОСТЕЙ ДЛЯ ОБУЧАЮЩЕЙ ВЫБОРКИ

Для алгоритма машинного обучения из каждого объекта флуоресценции и его непосредственного фона извлекаются следующие признаки. Для фона (BG): max, mean, median и mode; для центральной зоны (FG): 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 (рис. 1).

Указанные признаки образуют строки матрицы **M**. В последнем столбце такой матрицы содержится label — буквенный код нуклеотида, полученный из данных заранее секвенированной последовательности, предварительно отображенной на известной (референтной) последовательности бактериофага Phix174. Фрагмент такой матрицы представлен ниже в табл. 1.

Матрица **M** имеет следующую структуру. Первый столбец содержит номер кластера в изображении сигналов флуоресценции. Столбцы со второго по пятый содержат FG-признаки интенсивностей канала A, а именно max, mean, pct90 и pct99. Столбцы с шестого по девятый содержат FG-признаки интенсивностей канала C. Столбцы с десятого по тринадцатый содержат FG-признаки интенсивностей канала G. Столбцы с четырнадцатого по семнадцатый содержат FG-признаки интенсивностей канала T. Столбцы с восемнадцатого по двадцать первый содержат BG-признаки фона канала A, а именно max, mean, median и mode. Столбцы с двадцать второго по двадцать пятый содержат BG-признаки фона канала C. Столбцы с двадцать шестого по двадцать девятый содержат BG-признаки фона канала G. Столбцы с тридцатого по тридцать третий содержат BG-признаки фона канала T. В последнем тридцать четвертом столбце содержится Id — буквенный код нуклеотида, совпадающий с буквенным кодом из последовательности в референтном геноме бактерии Phix174. Такой буквенный код был получен следующим образом. На основе собранных наблюдений определялись оценки интенсивностей сигналов флуоресценции в каждом канале. Затем эти оценки были скорректированы на влияние перекрестных помех, но исходные (некорректированные) оценки интенсивностей сохранялись в памяти.

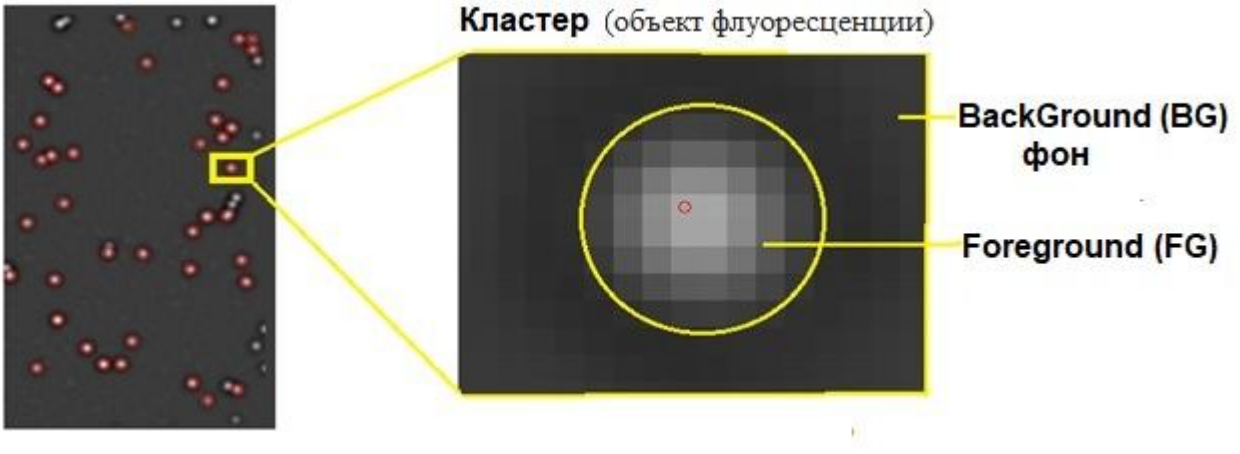


**Рис. 1.** Отбор исходных данных для машинного обучения

**Табл. 1.** Фрагмент матрицы **M** интенсивностей объектов флуоресценции

| Номер столбца матрицы **M** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | | 30 | 31 | 52 | 33 | 34 |
| № | FG — A | | | | FG — C | | | | FG — G | | | | FG — T | | | | BG — A | | | | | BG — T | | | | Id |
| 1 | 11 | 6 | 10 | 11 | 6 | 1 | 5 | 6 | 169 | 31 | 134 | 169 | 97 | 32 | 80 | 97 | 6 | 1 | 1 | 5 | | 2 | 6 | 0 | 0 | G |
| 2 | 9 | 4 | 8 | 9 | 58 | 7 | 48 | 58 | 2 | 0 | 2 | 2 | 1 | 1 | 1 | 1 | 5 | 0 | 1 | 4 | | 6 | 0 | 0 | 0 | C |
| 3 | 198 | 101 | 175 | 198 | 87 | 11 | 73 | 87 | 3 | 0 | 2 | 3 | 9 | 3 | 7 | 9 | 9 | 6 | 2 | 2 | | 7 | 0 | 0 | 0 | A |
| 4 | 6 | 4 | 5 | 6 | 9 | 1 | 6 | 9 | 193 | 51 | 162 | 193 | 66 | 27 | 56 | 66 | 4 | 0 | 0 | 3 | | 4 | 1 | 0 | 0 | G |
| 5 | 251 | 142 | 230 | 251 | 10 | 1 | 9 | 10 | 7 | 3 | 7 | 7 | 15 | 8 | 15 | 15 | 3 | 2 | 2 | 2 | | 3 | 1 | 0 | 0 | A |
| 6 | 6 | - 2 | 5 | 6 | 11 | 7 | 11 | 11 | 5 | 2 | 4 | 5 | 113 | 57 | 98 | 113 | 5 | 0 | 1 | 4 | | 7 | 0 | 1 | 6 | T |

После этого по скорректированным оценкам определялись буквенные коды нуклеотидов, которые сравнивались с буквенными кодами известной (референтной) последовательности нуклеотидов бактерии Phix 174. Если полученные буквенные коды совпадали с буквенными кодами референтной последовательности, то тогда исходные (некорректированные) оценки интенсивностей записывались в матрицу **M.**

Таким образом, матрица **M** содержит 34 столбца. Количество строк в матрице **M** определяется количеством кластеров объектов флуоресценции, информация о которых используется для построения обучающей выборки. В рассматриваемом ниже примере количество строк в матрице **M** равнялось 2000.

Пример фрагмента матрицы **M** представлен в табл. 1. Для сокращения места в тексте данной работы фрагмент матрицы **M** содержит двадцать шесть столбцов. Столбцы с двадцать второго по двадцать пятый, содержащие BG-признаки фона канала C, и столбцы с двадцать шестого по двадцать девятый, содержащие BG-признаки фона канала G, в табл. 1 не представлены.

Следует отметить, что данные об интенсивностях сигналов флуоресценции, представленные в матрице **M**, не подвергались коррекции на влияние различных химических процессов, включенных в технологию секвенирования, и перекрестные помехи, но буквенные коды, записанные в столбце Id, были получены с использованием указанных коррекций.

## ОПЕРАЦИИ ОБРАБОТКИ МАТРИЦЫ ИНТЕНСИВНОСТЕЙ

Операции обработки направлены на выделение из матрицы интенсивностей данных, принадлежащих определенному нуклеотиду, составление объединенной матрицы и сокращение размерности объединенной матрицы методом главных компонент.

Преобразуем описанную выше матрицу **M** в четыре матрицы, в каждой из которых будут содержаться интенсивности объектов флуоресценции, принадлежащие только определенному буквенному коду нуклеотидов. Таким образом, получаем матрицы $\mathbf{X_A}$, $\mathbf{X_C}$, $\mathbf{X_G}$ и $\mathbf{X_T}$ соответственно для каналов A, C, G и T.

В матрицы $\mathbf{X_A}$, $\mathbf{X_C}$, $\mathbf{X_G}$ и $\mathbf{X_T}$ включались числа из матрицы **M**, расположенные в столбцах со 2 по 33, т.е. номер кластера и соответствующий ему буквенный код нуклеотида в эти матрицы не включались. Для каждой из этих матриц из матрицы **M** выбиралось по 500 строк, соответствующих заданному буквенному коду. В результате получилось четыре матрицы, в каждой из которых содержится 500 строк и 32 столбца.

Составим теперь из матриц $\mathbf{X_A}$, $\mathbf{X_C}$, $\mathbf{X_G}$ и $\mathbf{X_T}$ объединенную матрицу $\mathbf{X_{ACGT}}$ таким образом, что ее первые 500 строк содержат данные из матрицы $\mathbf{X_A}$, строки с 501 по 1000 содержат данные из матрицы $\mathbf{X_C}$, строки с 1001 по 1500 содержат данные из матрицы $\mathbf{X_G}$ и строки с 1501 по 2000 содержат данные из матрицы $\mathbf{X_T}$. Таким образом, получается матрица $\mathbf{X_{ACGT}}$, состоящая из 2000 строк и 32 столбцов. Используем для преобразования матрицы $\mathbf{X_{ACGT}}$ метод главных компонент [7, 8]. В методе главных компонент данные разбиваются на компоненты, чтобы максимизировать линейную корреляцию между точками данных в матрице различий, задаваемых входными признаками. Посредством "преобразования координат" количество точек исходных данных с исходными координатами в многомерном пространстве заменяется данными с новыми полученными координатами, что снижает размерность набора данных за счет отбрасывания координат, которые могут не удовлетворять по ряду критериев, например по отношению полезного сигнал к шуму и т.п. Преобра-

зуем данные, содержащиеся в матрице $\mathbf{X}_{ACGT}$, методом главных компонент с помощью оператора преобразования PCA:

$$\mathbf{T}_{ACGT} = \text{PCA}\,(\mathbf{X}_{ACGT}). \qquad (1)$$

Полученная матрица $\mathbf{T}_{ACGT}$ называется матрицей счетов [7]. Количество строк в этой матрице равно количеству наблюдений, т.е. информации об интенсивностях сигналов флуоресценции 2000 кластеров. Количество столбцов в этой матрице равно 32. Числа по столбцам представляют собой новые координаты наших наблюдений. Такие координаты называются главными компонентами исходной матрицы $\mathbf{X}_{ACGT}$. При использовании метода главных компонент часто используют первые две главные компоненты в том случае, если собственные числа преобразованной методом главных компонент матрицы значительно убывают от первой к последующей компоненте [7]. Эти компоненты несут основную информацию из данных, представленных в исходной матрице. Для матрицы $\mathbf{T}_{ACGT}$ убывание собственных чисел от первой к третьей компоненте более чем в 4.5 раза дает основание использовать для представления наших данных первые две главные компоненты PC1 и PC2.

Представим графически информацию, полученную для первых двух главных компонент каждой строки матрицы $\mathbf{T}_{ACGT}$. Для этого по горизонтальной оси будем откладывать значения первой главной компоненты — координаты x, а по вертикальной оси будем откладывать значения второй главной компоненты — координаты y. Таким образом, мы в двухмерном пространстве представляем основную информацию из исходной матрицы $\mathbf{X}_{ACGT}$, но отображенную в других координатах. Разделим полученные пары координат на четыре части по 500 пар в каждой части. В первой части содержится информация о нуклеотиде A, во второй о нуклеотиде C, в третьей о нуклеотиде G, в четвертой о нуклеотиде T. На рис. 2 эти части представлены "облаками" точек, помеченными индексами нуклеотидов A, C, G, T.

## КЛАССИФИКАЦИЯ ДАННЫХ ПО АЛГОРИТМУ K-MEANS (K-СРЕДНИХ)

Классификация k-средних — это метод разделения, который рассматривает наблюдения полученных данных как объекты, имеющие различные местоположения и расстояния друг от друга [9, 10]. Он разбивает объекты на k взаимоисключающих классов таким образом, чтобы объекты внутри каждого класса были как можно ближе друг к другу и как можно дальше от объектов в других классах. Каждый класс характеризуется своим центроидом, или центральной точкой.

Цель метода k-средних [10] состоит в том, чтобы сгруппировать выборки в определенное количество (k) непересекающихся подгрупп (классов) с использованием расстояний, рассчитанных между объектами, чтобы каждая точка данных принадлежала только к одной группе.

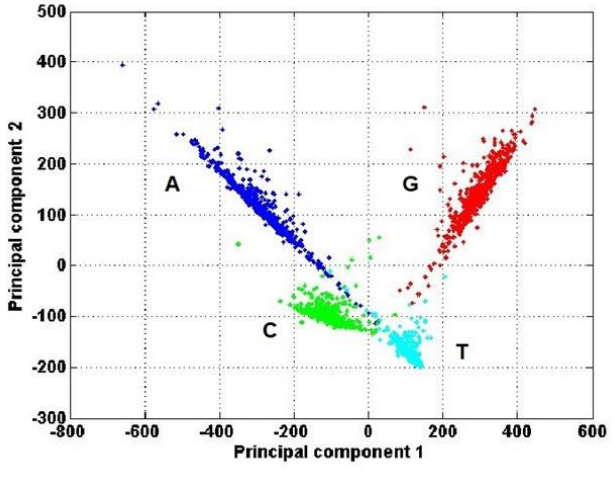


**Рис. 2.** Информация об интенсивностях сигналов флуоресценции после преобразования методом главных компонент

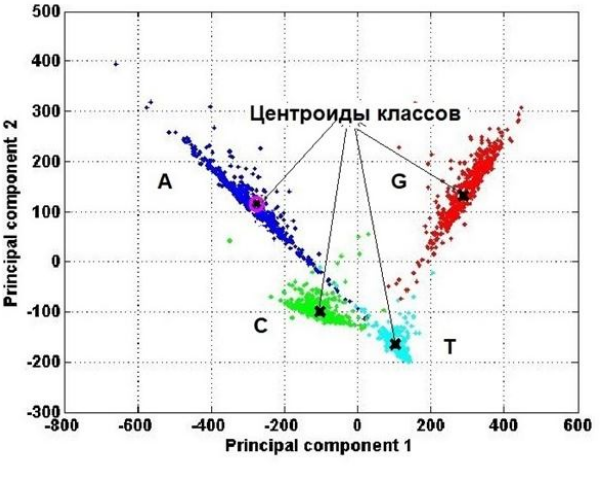


**Рис. 3.** Информация об интенсивностях сигналов флуоресценции после преобразования методом главных компонент и классификации методом k-средних

В настоящей работе для классификации использовался итеративный алгоритм k-средних, который объединяет исследуемые объекты в классы таким образом, чтобы сумма расстояний от каждого объекта до центроида его класса по всем классам была бы минимальной. Основная идея итерационного алгоритма заключается в том, что на каждой итерации вычисляется центр масс (центроид) для каждого класса, полученного на предыдущем шаге, затем данные разбиваются на классы вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения центров классов.

Для реализации этого алгоритма указывалось известное количество классов, на которые нужно разделить исследуемые объекты. В нашем случае это количество равно четырем, соответствующее числу нуклеотидов. В качестве меры близости исследуемых объектов к центроидам их классов использовалась сумма абсолютных разностей, т.е. расстояние L1. Выбор такой меры близости позволяет исключить влияние "выбросов" в исходных данных на точность оценки центроидов классов.

После проведения классификации по алгоритму k-средних получается набор из 4 пар чисел, которые содержат координаты центроидов каждого из четырех классов.

На рис. 3 представлена информация об интенсивностях сигналов флуоресценции после преобразования методом главных компонент и классификации методом k-средних. Координаты центроидов каждого из четырех классов показаны на рис. 3 указателем.

## ОПРЕДЕЛЕНИЕ ГРАНИЦ КЛАССОВ ДАННЫХ С ПОМОЩЬЮ ЭЛЛИПСОВ

После проведения классификации данных стали известны координаты центроидов для каждого из четырех классов. Следующим этапом обработки данных является определение границ каждого класса. Для определения этих границ построим эллипсы, координаты центров которых равны координатам центроидов найденных классов. Границы эллипсов определялись таким образом, чтобы внутри эллипсов находилось как можно больше представителей заданных классов. Для построения эллипсов проведем следующие операции:

1. Переместим данные класса А таким образом, чтобы координаты центроида перемещенных данных оказались бы нулевыми по горизонтальной и вертикальной осям так, как показано на рис. 4.

2. Произведем отбраковку "выбросов" таким образом, чтобы оставить для дальнейшей обработки данные, находящиеся в интервале от 1 до 99 процентиля.

3. Методом наименьших квадратов проведем линейную аппроксимацию перемещенных данных канала А и найдем угол наклона $\alpha$ аппроксимирующей прямой.

4. Повернем перемещенные в центр координат данные канала А на угол $\alpha$. Для этого сделаем преобразование данных по формуле (2):

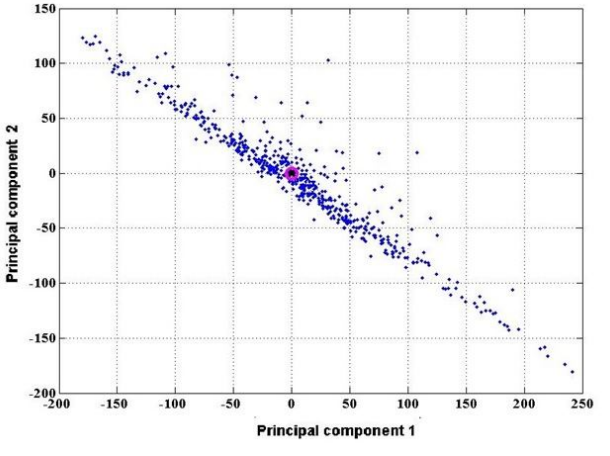


**Рис. 4.** Данные класса А после перемещения центроида в центр координат (PC1 – PC2)

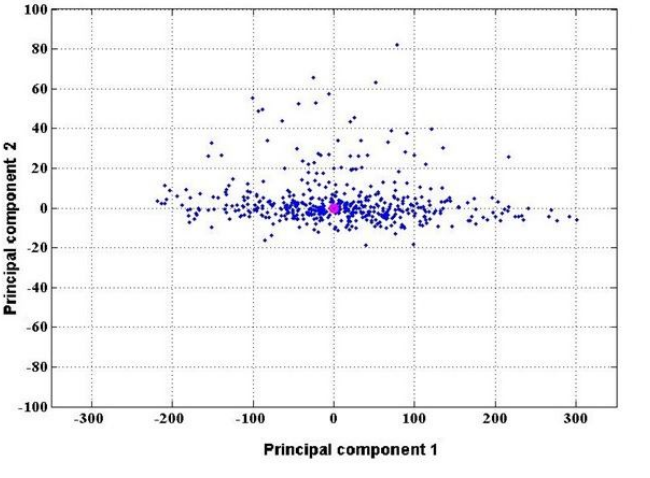


**Рис. 5.** Данные класса А после перемещения центроида в центр координат и поворота на угол $\alpha$

$$x_{\text{A}} = A_x \cdot \cos(-\alpha) - A_y \cdot \sin(-\alpha),$$
$$y_{\text{A}} = A_x \cdot \sin(-\alpha) + A_y \cdot \cos(-\alpha). \quad (2)$$

В этой формуле $A_x$ и $A_y$ — соответственно горизонтальные и вертикальные координатные оси в пространстве главных компонент (PC1 – PC2) для данных канала A после перемещения в центр; $x_{\text{A}}$ и $y_{\text{A}}$ — соответственно горизонтальные и вертикальные координаты данных канала A после поворота осей на угол $\alpha$. Данные после поворота представлены на рис. 5.

5. Построим эллипс с центром в нулевой точке, который имеет длины полуосей b и a.

Длина b малой полуоси искомого эллипса находится по формуле (3), а длина a большой полуоси — по формуле (4):

$$\text{b} = (\max(y_{\text{A}}) - \min(y_{\text{A}})) / \text{kb}, \quad (3)$$

$$\text{a} = (\max(x_{\text{A}}) - \min(x_{\text{A}})) / \text{ka}. \quad (4)$$

В этих формулах max и min означают соответственно максимумы и минимумы координат $x_{\text{A}}$ и $y_{\text{A}}$. Коэффициенты ka и kb определяются по итерационной процедуре таким образом, чтобы границы эллипсов каждого класса не пересекались и были бы на заданном минимальном расстоянии друг от друга.

6. Зная величины длин полуосей эллипса, теперь можно построить эллипс по формулам (5) с нулевым углом наклона к горизонтальной полуоси:

$$x_i^э = \text{a} \cdot \cos(t_i),$$
$$y_i^э = \text{b} \cdot \sin(t_i). \quad (5)$$

В этой формуле использована переменная $t_i$, которая является углом, изменяющимся от 0 до $2\pi$ радиан с заданным шагом, а $x_i^э$ и $y_i^э$ — соответственно горизонтальные и вертикальные координаты точек эллипса, т.е.:

$$t_i = 0 + (i - 1) \cdot \text{step}, \quad (6)$$

где $i$ — номер точки эллипса ($i \geq 1$), а step — шаг изменения переменной $t_i$. Например, если step = 0.05 радиан, то эллипс будет строиться по 126 точкам.

Эллипс, построенный по формуле (5), представлен на рис. 6.

На рис. 7 представлен эллипс с центром в точке с координатами, равными несмещенным координатам центроида для класса A, расположенный под углом $\alpha$.

Рассмотренная процедура построения границ данных для класса A была распространена для данных других классов: C, G и T. В результате были получены границы в виде эллипсов для всех классов.

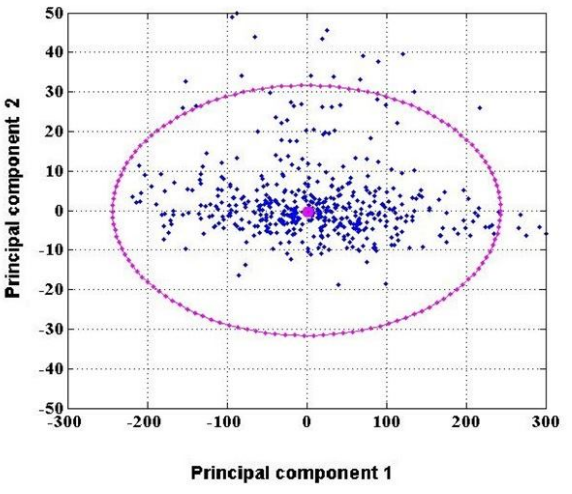


**Рис. 6.** Построение эллипса для данных класса A после перемещения в центр координат и поворота на угол $\alpha$

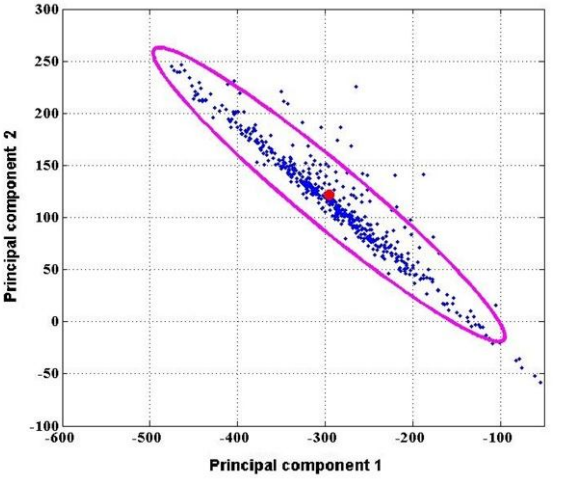


**Рис. 7.** Эллипс с центром в точке с координатами, равными координатам центроида для класса A, расположенный под углом $\alpha$

На рис. 8 показаны границы классов со значениями длин полуосей эллипсов, вычисленных по итерационной процедуре на основе формул (3) и (4).

Данные, оказавшиеся внутри полученных границ классов, составили обучающую выборку, которую можно использовать для решения задачи построения последовательности нуклеотидов (base-calling).

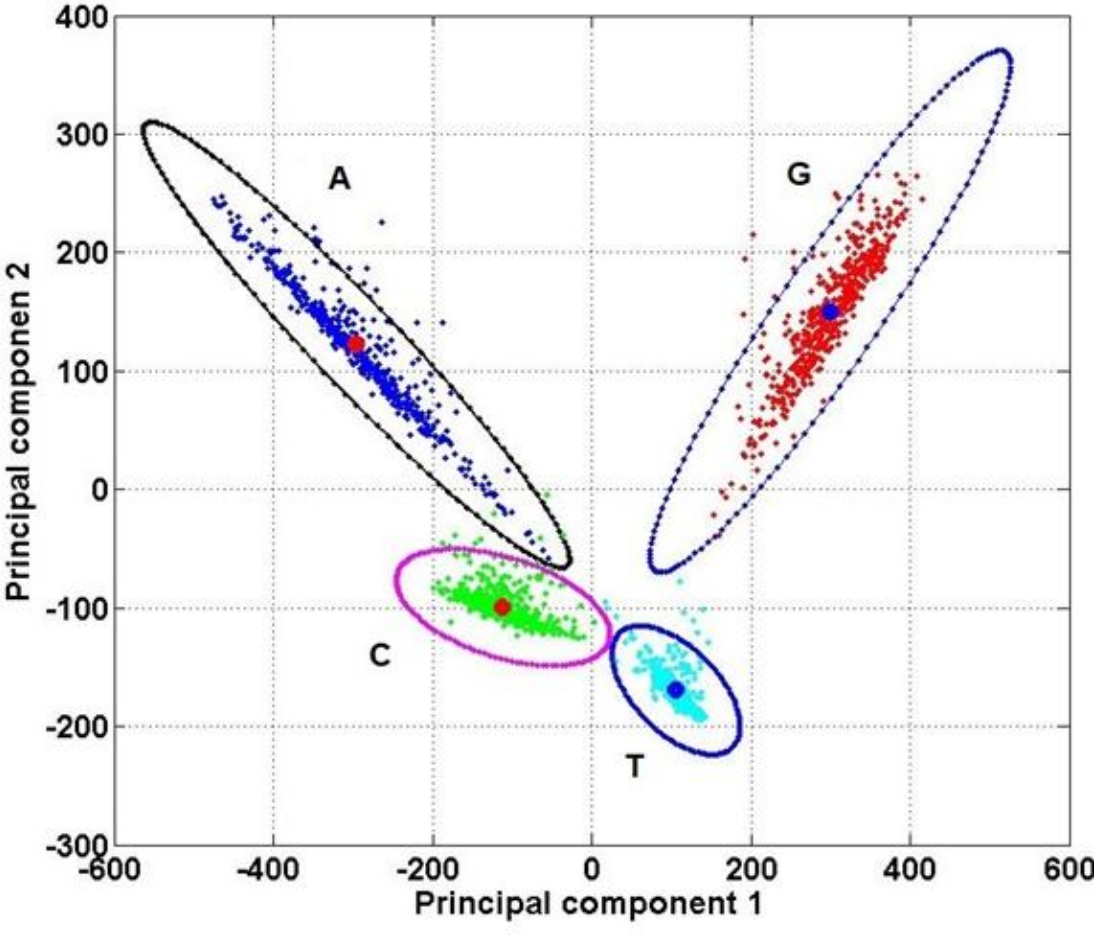


**Рис. 8.** Границы классов данных в виде эллипсов, параметры которых были вычислены по итерационной процедуре

Эффективность использования такого обучающего набора была проверена на тестовой выборке.

## ПОСТРОЕНИЕ ТЕСТОВОЙ ВЫБОРКИ И ВЫЯВЛЕНИЕ КОЛИЧЕСТВА ОШИБОК В ПОСЛЕДОВАТЕЛЬНОСТИ НУКЛЕОТИДОВ

Тестовая выборка, предназначенная для проверки возможности использования рассмотренного выше обучающего набора, составлялась из данных интенсивностей сигналов флуоресценции, которые не входили в матрицу **M**. Из таких данных составлялась матрица $\mathbf{M_1}$. Причем данные в матрице $\mathbf{M_1}$ также не подвергались коррекции, отмеченной выше для матрицы **M**. Структура матрицы $\mathbf{M_1}$ была подобна структуре рассмотренной матрицы **M**. Количество строк матрицы $\mathbf{M_1}$ равнялось 12 000. Из этой матрицы в программном цикле поочередно выбирались строки. Каждая из этих строк заменяла 1 строку в матрице $\mathbf{X_{ACGT}}$, например, с номером 500. После этого выполнялось преобразование модифицированной матрицы $\mathbf{X_{ACGT}}$ методом главных компонент по формуле (1) и получалась матрица $\mathbf{T_{ACGT}}$. Из этой матрицы выделялись первая и вторая главная компонента для строки с указанным номером. В результате получалась пара чисел. Иначе говоря, мы получали координаты точки в плоскости (PC1, PC2) (первая и вторая главная компонента). Допустим, что мы получили точку с координатами $(x, y)$.

Перед тем как переходить к следующей строке из матрицы $\mathbf{M_1}$, проводилась оценка, в какой из эллипсов попали координаты $(x, y)$ и таким образом определялся буквенный код нуклеотида, которому соответствовали интенсивности сигналов флуоресценции, записанные в анализируемой строке матрицы $\mathbf{M_1}$.

**Табл. 2.** Результат оценки количества ошибок по отдельным каналам

| Канал | Количество ошибок | Процент ошибочных кластеров (%) | Процент правильных кластеров (%) |
|---|---|---|---|
| A | 6 | 0.05 | 99.95 |
| C | 76 | 0.67 | 99.33 |
| G | 0 | 0 | 100 |
| T | 12 | 0.11 | 99.89 |

Этот буквенный код ставился в последнем столбце матрицы $\mathbf{M_1}$, аналогично столбцу Id в матрице **M**. Если точка с координатами $(x, y)$ не попадала ни в один из эллипсов, то в последнем столбце матрицы $\mathbf{M_1}$ ставился специальный знак. Далее анализировалась следующая строка матрицы $\mathbf{M_1}$.

Таким образом, были проанализированы 12 000 строк матрицы $\mathbf{M_1}$. Оценка количества ошибок в буквенных кодах нуклеотидов, полученных в результате анализа строк в матрице $\mathbf{M_1}$, проводилась в случае, когда величины полуосей эллипсов были подсчитаны по итерационной процедуре, результат которой показан на рис. 8.

Количество проверяемых кластеров — 12 000. Количество кластеров, которые не попали ни в один эллипс — 620, т.е. 5.1%. Количество кластеров, попавших в эллипсы — 11 380, т.е. 94.9%.

Результат оценки количества ошибок по отдельным каналам представлен в табл. 2.

## ВЫВОДЫ

1. Преобразование данных, содержащих информацию о сигналах флуоресценции, методом главных компонент и последующая класси фикация методом k-средних позволили создать обучающую выборку для идентификации буквенного кода нуклеотида.

2. Оценка эффективности обучающей выборки для идентификации букв нуклеотидов показала выявление более 99% правильных кластеров в тестовой выборке, объем которой в 6 раз превышал объем обучающей выборки.

3. Данные в тестовой выборке содержали информацию о сигналах флуоресценции, которая не подвергалась коррекции на влияние различных химических процессов, включенных 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), и тем не менее их обработка дала достаточно достоверный результат.

*Работа выполнена в ИАП РАН в рамках Государственного задания 075-01157-23-00 Министерства науки и высшего образования.*

## СПИСОК ЛИТЕРАТУРЫ

1. *Манойлов В.В., Бородинов А.Г, Заруцкий И.В, Петров А.И., Курочкин В.Е*. Алгоритмы обработки сигналов флуоресценции массового параллельного секвенирования нуклеиновых кислот // Труды СПИИРАН. 2019. Т. 18, № 4. С. 1010–1036. DOI: 10.15622/sp.2019.18.4.1010-1036
2. *Manoilov V.V, Borodinov A.G., Saraev A.S., Petrov A.I., Zarutskii I.V., Kurochkin V.E.* Algorithms for Image Processing in a Nanofor SPS DNA Sequencer // Technical Physics. 2022. Vol. 67, no. 4. P. 304–311. DOI: 10.1134/S1063784222050061
3. *Ghannam R.B., Techtmann S.M.* Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring // Computational and Structural Biotechnology Journal. 2021. Vol. 19. P. 1092–1107. DOI: 10.1016/j.csbj.2021.01.028
4. *Kircher M., Stenzel U., Kelso J.* Improved base calling for the Illumina Genome analyzer using machine learning strategies // Genome Biol. 2009. Vol. 10. Id. R83. DOI: 10.1186/gb-2009-10-8-r83
5. *Tegfalk E.* Application of machine learning techniques to perform base-calling in next-generation DNA sequencing. Thesis in degree project engineering physics KTH Royal Institute of Technology, Stockholm, Sweden, 2020. 53 p. URL: https://www.diva-portal.org/smash/get/diva2:1465444/FULLTEXT01.pdf
6. *Borodinov A., Manoilov V., Zarutsky I., Petrov A., Kurochkin V., Saraev A.* Machine learning in base-calling for next-generation sequencing methods // Informatics and Automation ('Trudy SPIIRAN'). 2022. Vol. 21, no. 3. P. 572–603. DOI: 10.15622/ia.21.3.5

7. *Померанцев А.* Метод главных компонент. (Сетевой ресурс) Российское хемометрическое общество. Учебники. URL: https://rcs.chemometrics.ru/ru/books
8. *Jolliffe I. T.* Principal Component Analysis. 2nd edition, Springer, 2002. 518 p.
9. *Martinez W.L., Martinez A.R.* Exploratory Data Analysis with MATLAB. A CRC Press Company Boca Raton, London, New York, Washington, D.C., 2005. 363 p.
10. *Kaufman L., Rousseeuw, P.J.* Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, New Jersey, John Wiley & Sons Inc, 2009. 342 p.

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

Контакты: *Манойлов Владимир Владимирович*, manoilov_vv@mail.ru

Материал поступил в редакцию 16.03.2023

# MACHINE LEARNING ALGORITHM FOR THE CONSTRUCTION OF A NUCLEOTIDE SEQUENCE IN THE NANOFOR SPS SEQUENCER USING THE PRINCIPAL COMPONENT ANALYSIS

**V. V. Manoilov, A. G. Borodinov, A. I. Petrov, I. V. Zarutsky, V. E. Kurochkin**

*Institute for Analytical Instrumentation of RAS, Saint Petersburg, Russia*

The development of information technologies and mathematical methods for data processing plays an essential role in establishing various features in the analyzed nucleic acids and trends in their modifications. An important stage in the technology of *massively parallel sequencing* of nucleic acids is the process of constructing a nucleotide sequence based on the measured intensities of fluorescence signals. The paper considers an algorithm for generating a training sample, that is used to construct a sequence of letter codes of DNA nucleotides via the intensities of fluorescence signals obtained directly from the results of image processing. These signals were not corrected for the physical and chemical characteristics of the sequencing process. The algorithm uses principal component analysis and a k-means classifier. With the help of such a classifier, the data after transformation using the method of principal components is separated into four independent classes according to the number of letter codes of DNA nucleotides. With the help of the training sample, the class to which the vector containing the fluorescence signal data belongs, and hence its letter code, are determined. The algorithm's performance on a test sample revealed great outcome reliability.

*Keywords*: nucleic acid sequencing, mathematical processing and classification of multivariate data, principal component analysis, machine learning

## INTRODUCTION

An important element of the successful development of genomic sequencing is the use of modern information technologies and mathematical methods for data processing to establish various features in the analyzed nucleic acids. The Institute for Analytical Instrumentation of the Russian Academy of Sciences has developed a computer appliance (Нанофор СПС) for decoding nucleic acid sequences using the method of mass parallel sequencing.

The solution to the problem of genome decoding in a computer appliance is divided into a number of stages of processing the initial data. One of the important initial steps in data processing is the evaluation of fluorescence signal intensities for various wavelengths in the flow-through cell image frames for several cycles of synthetic sequencing. Such an assessment is performed using image processing programs, the algorithms of which are described in [1, 2]. The final part of such a preliminary analysis of the initial data is the process of constructing a sequence of letter codes for nucleotides.

Various chemical processes involved in synthesis sequencing technology cause changes in the recorded intensities, including the effects of phasing / prephasing, signal decay, and crosstalk.

The application of machine learning (ML) to DNA sequencing tasks involves building and evaluating models that use algorithms that can recognize, classify, and predict certain outcomes from data. ML approaches are subdivided into learning without a teacher (unsupervised), learning with partial involvement of a teacher (semi-supervised), and learning with a teacher (supervised) [3]. For example, often the goal of supervised ML applied to sequencing data is to build a model based on a training set of collected observations with a known nucleotide sequence in order to predict a nucleotide for an arbitrary sample with an unknown target value of the type of nucleotide to be determined [4, 5]. The input variables are often referred to as features, and the corresponding samples are observations.

In the article [6], a review of machine learning methods for solving the problem of constructing a nucleotide sequence was made, and several examples of the use of machine learning for processing sequencer Нанофор СПС data were considered. This work is a continuation of the research published in that article.

In this paper, based on the PCA, we consider a possible machine learning approach for creating and evaluating a model that implements the stage of constructing a nucleotide sequence. A series of sequencing data from the Нанофор СПС instrument shows the promise of the machine learning method for solving the problem of constructing a nucleotide sequence.

## CONSTRUCTION OF THE INTENSITY MATRIX-FOR TRAINING SAMPLE

For the machine learning algorithm, the following features are extracted from each fluorescence object and its immediate background. For the background (BG): max, mean, median, and mode; for the foreground zone (FG): max, mean, pct90, and pct99, where max is the maximum intensity value, mean is the arithmetic mean, mode is the most common value, and pct90 and pct99 are the 90th and 99th percentiles, respectively (Fig. 1).

**Fig. 1.** Selection of initial data for machine learning

**Tab. 1.** Fragment of the matrix **M** of object fluorescence intensities

These features form the rows of the matrix **M**. The last column of this matrix contains a label, the letter code of the nucleotide obtained from the data of a preformed sequence previously mapped onto the known (reference) sequence of the bacteriophage Phix174. A fragment of such a matrix is presented below in Tab. 1.

The matrix **M** has the following structure. The first column contains the number of clusters in the fluorescence signal image. Columns from the second to the fifth contain FG-features of channel A intensities, namely max, mean, pct90 and pct99. Columns six through nine contain the FG features of the C channel intensities. Columns ten through thirteen contain the FG-features of the G channel intensities. Columns fourteen through seventeen contain the FG-features of the T channel intensities. Columns eighteen through twenty-one contain the BG-features of the channel A background, namely max, mean, median, and mode. Columns twenty-two to twenty-five contain the C channel BG-features. Columns twenty-six to twenty-nine contain the G channel BG-features. Columns thirty through thirty-three contain the T channel BG-features. The last thirty-fourth column contains Id — the nucleotide letter code that matches the letter code from the sequence in the reference genome of the bacterium Phix174. Such a letter code was obtained as follows. Based on the collected observations, estimates of the intensities of fluorescence signals in each channel were determined. These estimates were then corrected for the effects of crosstalk, but the original (uncorrected) intensity estimates were kept in memory.

After that, according to the corrected estimates, the nucleotide letter codes were determined, which were compared with the letter codes of the known (reference) nucleotide sequence of the bacterium Phix 174. If the obtained letter codes coincided with the letter codes of the reference sequence, then the original (non-corrected) intensity estimates were written in the **M** matrix.

Thus, the matrix **M** contains 34 columns. The number of rows in the matrix **M** is determined by the number of clusters of fluorescence objects, information about which is used to build a training sample. In the example below, the number of rows in the matrix **M** was 2000.

An example of a fragment of the matrix **M** is presented in Tab. 1. To reduce the space in the text of this work, the fragment of the matrix **M** contains twenty-six columns. In Tab. 1, columns twenty-two to twenty-fifth, containing the BG-features of the C channel, and columns twenty-six to twenty-ninth, containing the BG-features of the G channel, are not shown.

It should be noted that the data on the intensities of fluorescence signals presented in the matrix **M** were not corrected for the influence of various chemical processes included in the sequencing technology, and crosstalk, but the letter codes written in the Id column were obtained using the indicated corrections.

## INTENSITY MATRIX PROCESSING

Processing operations are aimed at extracting the data belonging to a certain nucleotide, from the intensity matrix, compiling the combined matrix, and reducing the dimensionality of the combined matrix using principal component analysis.

Let us transform the matrix **M** described above into four matrices, each of which will contain the intensities of fluorescence objects belonging only to a certain letter code of nucleotides. Thus, we obtain matrices $\mathbf{X_A}$, $\mathbf{X_C}$, $\mathbf{X_G}$ and $\mathbf{X_T}$, respectively, for channels A, C, G and T.

The matrices $\mathbf{X_A}$, $\mathbf{X_C}$, $\mathbf{X_G}$ and $\mathbf{X_T}$ included numbers from the matrix **M**, located in columns 2 to 33, i.e., the cluster number and the corresponding letter code of the nucleotide were not included in these matrices. For each of these matrices, 500 rows corresponding to a given letter code were selected from the matrix **M**. The result is four matrices, each containing 500 rows and 32 columns.

Let us now compose the combined $\mathbf{X_{ACGT}}$ matrix from matrices $\mathbf{X_A}$, $\mathbf{X_C}$, $\mathbf{X_G}$ and $\mathbf{X_T}$ in such a way that its first 500 rows contain data from the $\mathbf{X_A}$ matrix, rows 501 to 1000 contain data from the $\mathbf{X_C}$ matrix, rows 1001 to 1500 contain data from the $\mathbf{X_G}$ matrix, and rows 1501 to 2000 contain data from the $\mathbf{X_T}$ matrix. Thus, an $\mathbf{X_{ACGT}}$ matrix consisting of 2000 rows and 32 columns is obtained. We use the method of principal components to transform the $\mathbf{X_{ACGT}}$ matrix [7, 8]. In the principal component analysis, the data is split into components to maximize the linear correlation between data points in the matrix of differences given by the input features. Through "coordinate transformation", the number of original data points with original coordinates in multidimensional space is replaced with data with newly obtained coordinates, which reduces the volume of the data set by discarding coordinates that may not satisfy a number of criteria, for example, the ratio of useful signal to noise, etc. Let's transform the data contained in the $\mathbf{X_{ACGT}}$ matrix using principal component analysis and the PCA transformation operator:

$$\mathbf{T_{ACGT}} = \mathrm{PCA}\ (\mathbf{X_{ACGT}}). \qquad (1)$$

The resulting $\mathbf{T_{ACGT}}$ matrix is called the score matrix [7]. The number of rows in this matrix is equal to the number of observations, that is, information about the intensities of the fluorescence signals from 2000 clusters. The number of columns in this matrix is 32. The numbers in the columns represent the new coordinates of our observations. Such coordinates are called the principal components of the original $\mathbf{X_{ACGT}}$ matrix. When using the principal component analysis, the first two principal components are often used, in the event that the eigenvalues of the matrix transformed by the principal component analysis, decrease significantly from the first to the next component [7].

These components carry the basic information from the data presented in the original matrix. For the $\mathbf{T_{ACGT}}$ matrix, the decrease in eigenvalues from the first to the third component is more than 4.5 times, which gives reason to use the first two main components, PC1 and PC2, to represent our data.

Let's graphically represent the information obtained for the first two principal components of each row of the $\mathbf{T_{ACGT}}$ matrix. To do this, we will plot the values of the first principal component, the x coordinates, along the horizontal axis, and the values of the second principal component, the y coordinates, along the vertical axis. Thus, in two-dimensional space, we represent the basic information from the original $\mathbf{X_{ACGT}}$ matrix, but display it in other coordinates. We divide the obtained pairs of coordinates into four parts, with 500 pairs in each part. The first part contains information about nucleotide A, the second about nucleotide C, the third about nucleotide G, the fourth about nucleotide T. In Fig. 2, these parts are represented by "clouds" of dots marked with nucleotide indices A, C, G, T.

**Fig. 2.** Information on fluorescence signal intensities after principal component transformation

## DATA CLASSIFICATION USING THE K-MEANS ALGORITHM

The k-means classification is a separation method that considers the observations of the received data as objects having different locations and distances from each other [9, 10]. It splits the objects into k mutually exclusive classes in such a way that the objects within each class are as close to each other as possible and as far from the objects in other classes as possible. Each class is characterized by its centroid, or central point.

The purpose of the k-means method [10] is to group the samples into a certain number (k) of non-overlapping subgroups (classes) using distances calculated between objects so that each data point belongs to only one group.

In this paper, for classification, an iterative k-means algorithm was used, which combines the objects under study into classes in such a way that the sum of the distances from each object to the centroid of its class would be minimal over all classes. The main idea of the iterative algorithm is that at each iteration, the center of mass (centroid) is calculated for each class obtained in the previous step, and then the data is split into classes again in accordance with which new centers turned out to be closer according to the chosen metric. The algorithm terminates when the class centers do not change at some iteration.

To implement this algorithm, a known number of classes were specified, according to which the objects under study should be distributed. In our case, this number is four, corresponding to the number of nucleotides. As a measure of the proximity of the studied objects to the centroids of their classes, the sum of absolute differences, i.e., distance L1, was used. The choice of such a proximity measure makes it possible to exclude the influence of "outliers" in the initial data on the accuracy of class centroid estimation.

After classification using the k-means algorithm, a set of 4 pairs of numbers is obtained, that contain the coordinates of the centroids of each of the four classes.

Fig, 3 provides information on the intensities of fluorescence signals after principal component transformation and k-means classification. The centroid coordinates of each of the four classes are shown in Fig. 3 with a pointer.

**Fig. 3.** Information on fluorescence signal intensities after principal component transformation and k-means classification

## DEFINING DATA CLASSES BOUNDARIES BY MEANS OF ELLIPSES

After the data classification, the coordinates of the centroids for each of the four classes became known. The next stage of data processing is to determine the boundaries of each class. To determine these boundaries, we construct ellipses, the coordinates of the centers of which are equal to the coordinates of the centroids of the found classes. The boundaries of the ellipses were determined in such a way that there were as many representatives of the given classes as possible inside the ellipses. To construct ellipses, we will carry out the following operations.

1. Move class A data in such a way that the coordinates of the centroid of the moved data would be zero on the horizontal and vertical axes, as shown in Fig. 4.

**Fig. 4.** Class A data after moving the centroid to the center of coordinates (PC1 – PC2)

2. We will reject the "outliers" in such a way as to leave the data in the range from 1 to 99 percentile for further processing.

3. Using the least squares method, we will carry out a linear approximation of the displaced data of channel A and determine the slope angle $\alpha$ of the approximating straight line.

4. Rotate the channel A data, moved to the center of coordinates, through the angle $\alpha$. To do this, we will transform the data according to the formula (2):

$$\begin{aligned} x_{\text{A}} &= A_x \cdot \cos(-\alpha) - A_y \cdot \sin(-\alpha), \\ y_{\text{A}} &= A_x \cdot \sin(-\alpha) + A_y \cdot \cos(-\alpha). \end{aligned} \tag{2}$$

In this formula, $A_x$ and $A_y$ are respectively the horizontal and vertical coordinate axes for channel A data after moving to the center; $x_{\text{A}}$ and $y_{\text{A}}$ are respectively the horizontal and vertical coordinates of channel A data after rotating the axes through the angle $\alpha$. The data after rotation are shown in Fig. 5.

**Fig. 5.** Class A data after moving the centroid to the center of coordinates and rotating through the angle $\alpha$

5. Let's build an ellipse centered at the zero point, that has the lengths of the semiaxes b and a. The length b of the minor semiaxis of the desired ellipse is found by formula (3), and the length a of the major semiaxis is found by formula (4):

$$\text{b} = (\max(y_{\text{A}}) - \min(y_{\text{A}})) / \text{kb}, \tag{3}$$

$$\text{a} = (\max(x_{\text{A}}) - \min(x_{\text{A}})) / \text{ka}. \tag{4}$$

In these formulas, max and min imply, respectively, the maxima and minima of the coordinates $x_{\text{A}}$ and $y_{\text{A}}$. The coefficients ka and kb are determined using an iterative procedure in such a way that the boundaries of the ellipses of each class do not intersect and are at a given minimum distance from each other.

6. Knowing the lengths of the semiaxes of the ellipse, it is now possible to construct an ellipse according to formulas (5) with a zero angle of inclination to the horizontal semiaxis:

$$\begin{aligned} x_i^э &= \text{a} \cdot \cos(t_i), \\ y_i^э &= \text{b} \cdot \sin(t_i). \end{aligned} \tag{5}$$

This formula uses the variable $t_i$, which is an angle ranging from 0 up to $2\pi$ radians with a given step, and $x_i^э$ and $y_i^э$ are, respectively, the horizontal and vertical coordinates of the points of the ellipse, that is:

$$t_i = 0 + (i - 1) \cdot \text{step}, \tag{6}$$

where $i$ is the number of the ellipse point ($i \geq 1$), and step is the step of changing the variable $t_i$. For example, if step = 0.05 radians, then the ellipse will be built using 126 points.

The ellipse constructed according to formula (5) is shown in Fig. 6.

**Fig. 6.** Constructing an ellipse for class A data after moving to the center of coordinates and rotating through an angle $\alpha$

Fig. 7 shows an ellipse centered at a point with coordinates equal to the unshifted coordinates of the centroid for class A, and located at an angle $\alpha$.

**Fig. 7.** Ellipse centered at a point with coordinates equal to the coordinates of the centroid for class A, located at an angle α

The considered procedure for constructing data boundaries for class A was extended to the data of other classes: C, G and T. As a result, boundaries in the form of ellipses were obtained for all classes. Fig. 8 shows the class boundaries with the values of the lengths of the semiaxes of the ellipses calculated using the iterative procedure based on formulas (3) and (4). The data that ended up inside the obtained class boundaries formed a training set that can be used to solve the problem of constructing a nucleotide sequence (base calling).

**Fig. 8.** Data class boundaries in the form of ellipses, the parameters of which were calculated using an iterative procedure

On a test set, the utility of such a training set was evaluated.

## CONSTRUCTION OF A TEST SAMPLE AND DETECTING THE NUMBER OF ERRORS IN THE SEQUENCE OF NUCLEOTIDES

The test sample, intended to check the possibility of using the training set considered above, was composed of the given fluorescence signal intensities, which were not included in the matrix **M**. The matrix $\mathbf{M_1}$ was compiled from such data. Moreover, the data in the $\mathbf{M_1}$ matrix were also not subjected to the correction noted above for the **M** matrix. The $\mathbf{M_1}$ matrix structure was similar to that of the considered matrix **M**. The number of $\mathbf{M_1}$ matrix rows was 12 000. Rows were selected in turn from this matrix in the program cycle. Each of these rows replaced 1 row in the $\mathbf{X_{ACGT}}$ matrix, for example, with the number 500. After that, the modified $\mathbf{X_{ACGT}}$ matrix was transformed using the principal component analysis according to formula (1) and the $\mathbf{T_{ACGT}}$ matrix was obtained. From this matrix, the first and second principal components were extracted for the row with the specified number. The result was a pair of numbers. In other words, we obtained the coordinates of a point in the plane (PC1, PC2) (the first and second principal components). Let's say that we have a point with coordinates ($x$, $y$).

Before proceeding to the $\mathbf{M_1}$ matrix next row, an assessment was made in which of the ellipses the coordinates ($x$, $y$) fell into, and thus the letter code of the nucleotide was determined, which corresponded to the fluorescence signal intensities recorded in the analyzed row of the $\mathbf{M_1}$ matrix.

This letter code was placed in the last column of the matrix $\mathbf{M_1}$, similar to the column Id in the matrix **M**. If the point with coordinates ($x$, $y$) did not fall into any of the ellipses, then a special sign was placed in the last column of the matrix $\mathbf{M_1}$. Then the next row of the matrix $\mathbf{M_1}$ was analyzed.

Thus, $\mathbf{M_1}$ matrix 12 000 rows were analyzed. Estimating the number of errors in the letter codes of the nucleotides obtained as a result of the analysis of raws in the matrix $\mathbf{M_1}$ was carried out in the case when the values of the semiaxes of the ellipses were calculated using an iterative procedure, the result of which is shown in Fig. 8.

The number of checked clusters is 12 000. The number of clusters that did not fall into any of the ellipses is 620, i.e., 5.1%. The number of clusters that fell into ellipses is 11 380, i.e. 94.9%.

The result of estimating the number of errors for individual channels is presented in Tab. 2.

**Tab. 2.** The result of estimating the error rate for individual channels

## CONCLUSIONS

1. The transformation of data containing information about fluorescence signals using principal component analysis and subsequent k-means classification made it possible to create a training set for identifying the letter code of the nucleotide.

2. Evaluation of the effectiveness of the training sample for identifying letters of nucleotides showed the identification of more than 99 percent of the correct clusters in the test sample, the volume of which was 6 times larger than the volume of the training sample.

3. The data in the test sample contained information about fluorescence signals, which was not corrected for the influence of various chemical processes included in the synthesis sequencing technology, causing shifts in the values of the recorded intensities. including the effects of phasing / prephasing, signal decay, and crosstalk. Nevertheless, their processing gave a fairly reliable result.

## REFERENCES


1. Manoilov V.V., Borodinov A.G., Zarutsky I.V., Petrov A.I., Kurochkin V.E. [Algorithms of Processing Fluorescence Signals for Mass Parallel Sequencing of Nucleic Acids]. *Trudy SPIIRAN* [Informatics and Automation (SPIIRAS Proceedings)], 2019, vol. 18, no. 4, pp. 1010–1036. DOI: 10.15622/sp.2019.18.4.1010-1036 (In Russ.).
2. Manoilov V.V., Borodinov A.G., Saraev A.S., Petrov A.I., Zarutskii I.V., Kurochkin V.E. Algorithms for image processing in a Nanofor SPS DNA sequencer. *Technical Physics*, 2022, vol. 67, no. 4, pp. 304–311. DOI: 10.1134/S1063784222050061
3. Ghannam R.B., Techtmann S.M. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*, 2021, vol. 19, pp. 1092–1107. DOI: 10.1016/j.csbj.2021.01.028
4. Kircher M., Stenzel U., Kelso J. Improved base calling for the Illumina Genome analyzer using machine learning strategies. *Genome Biol*., 2009, vol. 10, id. R83. DOI: 10.1186/gb-2009-10-8-r83
5. Tegfalk E. *Application of machine learning techniques to perform base-calling in next-generation DNA sequencing*. *Thesis in degree project engineering physics KTH Royal Institute of Technology, Stockholm, Sweden*. KTH Royal Institute of Technology School of Engineering Sciences, 2020. 53 p. URL: https://www.diva-portal.org/smash/get/diva2:1465444/FULLTEXT01.pdf
6. Borodinov A., Manoilov V., Zarutsky I., Petrov A., Kurochkin V., Saraev A. Machine learning in base-calling for next-generation sequencing methods. *Informatics and Automation ('Trudy SPIIRAN')*, 2022, vol. 21, no. 3, pp. 572–603. DOI: 10.15622/ia.21.3.5
7. Pomerantsev A. *Metod glavnykh komponent.* (Setevoi resurs) Rossiiskoe khemometricheskoe obshchestvo. Uchebniki [Method of main components. (network resource) Russian Chemometric Society. Textbooks]. URL: https://rcs.chemometrics.ru/ru/books (In Russ.).
8. Jolliffe I.T. *Principal Component Analysis.* 2nd edition. Springer, 2002. 518 p.
9. Martinez W.L., Martinez A.R. *Exploratory Data Analysis with MATLAB*. A CRC Press Company Boca Raton, London, New York, Washington, D.C., 2005. 363 p.
10. Kaufman L., Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, New Jersey, John Wiley & Sons Inc, 2009. 342 p.


Contacts: *Manoilov Vladimir Vladimirovich,*
manoilov_vv@mail.ru

Article received by the editorial office on 16.03.2024