Научная статья

Дахова Елизавета

Май 2019

1 Введение

В биологических лабораториях часто используют ткани с различными типами клеток. Нередко возникает задача выделения определенного довольно редкого типа клеток. Использовать большие образцы ткани часто бывает дорого и затратно по времени. Однако если ограничиться образцами с небольшим количеством клеток, то из-за случайности в экспрессиях генов не получится выделить тип клеток как отдельный кластер. Поэтому хочется найти некоторое минимальное число клеток, при котором еще возможно решить эту задачу.

В контексте нашей задачи дадим определения следующих терминов. **РНК** – последовательность молекул (нуклеотидов), которая отвечает за производство белков. **Белок** – большая молекула, которая несет определенную функцию клетки. Благодаря разнообразию белков в клетках их возможно отличать друг от друга. **Ген** – структурная единица клетки, которая отвечает за производство РНК. **РНК секвенирование** или просто **секвенирование** – процесс получения РНК в лабораторных условиях. **Single cell секвенирование** – новый более дешевый и более быстрый метод секвенирования, который рассматривает экспрессии клеток в отдельности, тип клеток может быть не известен.

2 Постановка задачи

Требуется научиться выделять редкий подтип клеток в single cell данных. Для этого нужно понять, какое наименьшее количество клеток необходимо просеквенировать, чтобы обнаружить этот редкий тип в смеси с другими клетками.

Уточним задачу. Требуется найти N – минимальное количество клеток для секвенирования, чтобы клетки типа A образовывали отличимый от остальных клеток кластер. Далее будем называть клетки типа A как A-клетки. Величина α означает долю A-клеток, которая достаточно мала. Число G есть количество генов в аннотации. Также существует набор из P маркерных генов для A-клеток, который заранее известен. Маркерные гены отличаются тем, что экспрессия этих генах в данном типе клеток в несколько раз больше чем в большинстве других клеток. В таком случае нам нужно понять взаимосвязь между величинами P, α, G и минимальным значением N, для которого выполняются следующие условия:

- A-клетки образуют отдельный кластер;

- хотя бы 1 из P маркеров является маркером кластера.

Похожая задача рассматривается в статье [3]. Отличие состоит в том, что там обозреваются отдельные виды клеток с известными маркерными генами. Наша задача более общая – одним из ключевых пунктов в ней является генерация данных, для чего нужно задавать распределение экспрессий генов в клетках. До некоторого момента времени считалась, что в экспрессиях, полученные методами single cell преобладают нулевые значения. В статье [2] подробно объясняется, что это предположение неверно, а данные хорошо описываются отрицательным биномиальным распределением. В статье [1] подтверждается тот факт, что экспрессии генов подчиняются закону отрицательного биномиального распределения.

3 Описание экспериментов

Проведенные эксперименты состоят из следующих этапов.

- Генерация данных;
- Выбор метрики;
- Подбор метода кластеризации;
- Выявление зависимости между параметрами.

Далее подробно опишем каждый этап.

3.1 Генерация данных

Экспрессия каждого гена имеет отрицательное биномиальное распределение. Его параметры p и n лучше выразить через математическое ожидание μ и дисперсию σ^2 следующим образом

$$p = \frac{\mu}{\sigma^2}, \qquad n = \frac{\mu^2}{\sigma^2 - \mu}.$$

Такое представление обусловлено тем, что математическое ожидание и дисперсия лучше интерпретируемы. Математическое ожидание μ следует подобрать из равномерного распределения, так как нет информации о специфичности генов за исключением маркерных. Величину σ также возьмем из равномерного распределения. В силу положительности n на величину σ ставится ограничение снизу $\sigma > \sqrt{\mu}$. Так как в реальных образцах выборочная дисперсия зависит от среднего, то ограничение сверху на σ поставим $\mu/2$.

A-клетки можно выделить по P маркерным генам. Маркерные гены отличаются тем, что их экспрессия в клетках типа A в несколько раз больше чем в остальных клетках. Положим k – коэффициент отличия экспрессии, и для упрощения задачи установим его одинаковым для всех генов. Остальные гены во всех клетках экспрессируются по одному и тому же закону распределения.

3.2 Выбор метрики

Так как типы клеток в смеси не сбалансированы, то будем использовать метрику, которая не зависит от пропорций классов в данных – f1-score. [ОПРЕДЕЛЕНИЕ МЕТРИКИ ПОЧЕМУ ОНА]

3.3 Подбор алгоритма кластеризации

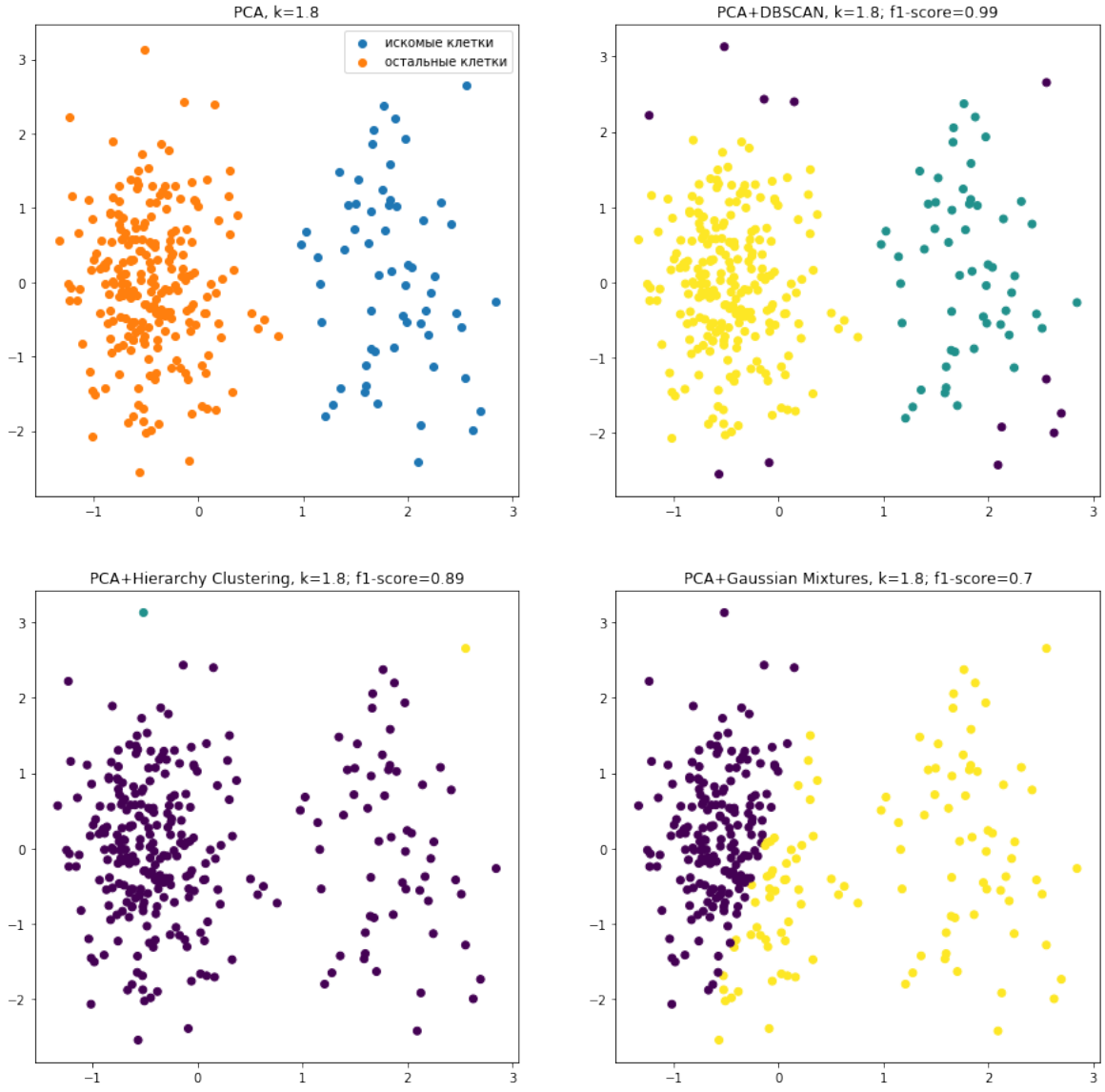


Рис. 1: Сравнение работы методов DBSCAN, Hierarchy Clustering, Gaussian Mixture Model для модельного примера при $N = 300$.

Так как визуализация результатов имеет важное значение для исследования, то преобразуем данные с помощью метода главных компонент (PCA) для двух главных компонент, по которым в дальнейшем будем разделять данные по кластерам. Проверим несколько методов кластеризации – метод метрической кластеризации с шумом

DBSCAN, метод иерархической кластеризации (Hierarchy Clustering), а также кластеризацию на основе гауссовской смеси распределений (Gaussian Mixture Model). Метод k-means в данном случае не рассматривается, поскольку он является частным случаем гауссовских смесей, для шарообразных матриц ковариаций. Проведенные эксперименты показали, что лучше всех работает DBSCAN. На рисунке 1 проиллюстрирован типичный пример результата работы методов.

4 Полученные результаты

Зависимость между параметрами будем определять при помощи сетки значений параметров. Рассмотрим значения величины N в пределах от 50 до 500 клеток. Значения доли искомых клеток α возьмем равными 0.1 и 0.01. Количество генов в аннотации G рассмотрим равными 10^4 и 10^5. Количество маркерных генов P выберем равными 5 и 10 Значения метрики рассмотрим в зависимости от k – степени отличимости маркерных генов в искомых клетках. Полученные результаты представлены на рисунках 2 и 3.

Из представленных результатов можно заключить следующее эмпирические правила определения нужного числа клеток для эксперимента:

- Экспрессия маркерных генов должна отличаться хотя бы в 2.5 раза, чтобы клетки можно было отличить.
- Если экспрессия маркерных генов отличается хотя бы 2 раза, а доля искомых клеток примерно 0.1, то для эксперимента достаточно 200 клеток вне зависимости от значений P и G из нашего диапазона.
- Если доля искомых клеток примерно 0.01, то для адекватного результата потребуется не менее 300 клеток, а экспрессия маркерных генов должна отличается хотя бы 2 раза.
- Если доля искомых клеток примерно 0.01, то для получения хорошего результата должны выполняться следующие условия: количество клеток не меньше 500, экспрессия маркерных генов отличается хотя бы 5 раз, количество маркерных генов не менее 10 из 10^5 генов в аннотации.

5 Выводы и дальнейшие исследования

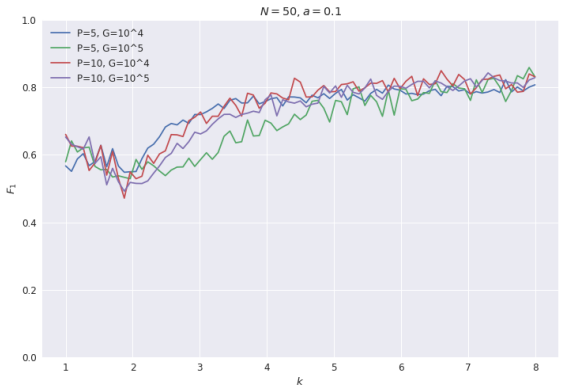
Список литературы

[1] *Count depth variation makes Poisson scRNA-seq data Negative Binomial.*

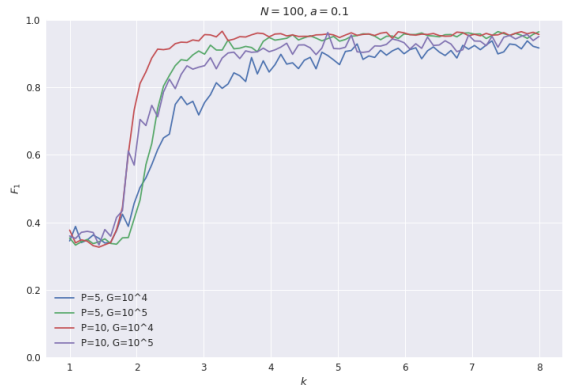
[2] *Droplet scRNA-seq is not zero inflated.*

[3] *Resolving Cell Types as a Function of Read Depth and Cell Number.* 10x Genomics, Inc, 2018.

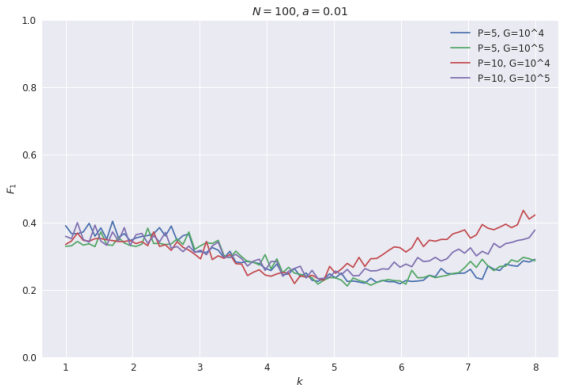
a)



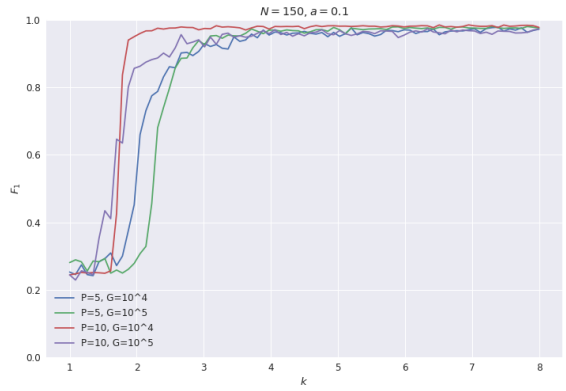
b)



c)

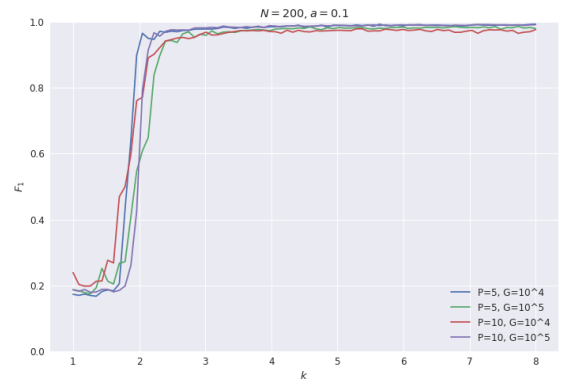


d)



e)

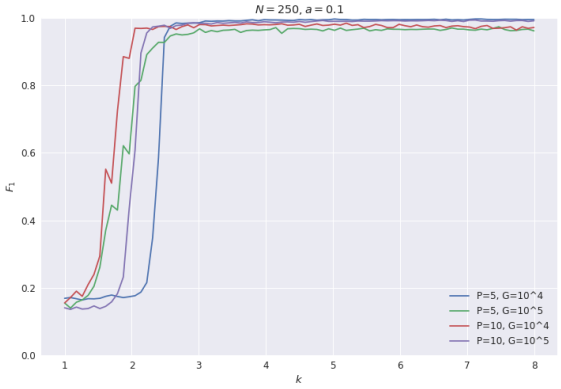
f)



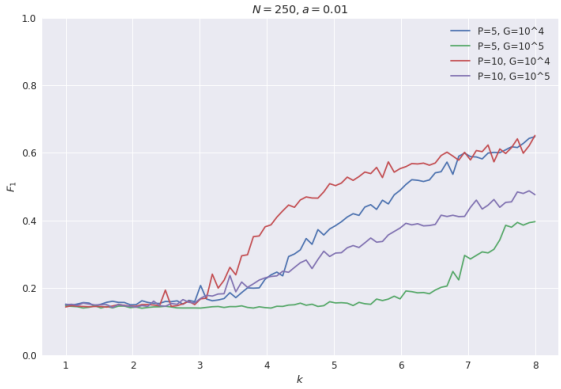
g)

h)

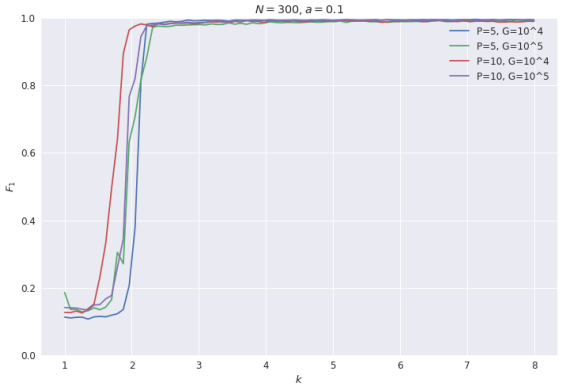
Рис. 2: Значения f1-score в зависимости от k и параметров N, α, P, G, где $P \in \{5, 10\}$, $G \in \{10^4, 10^5\}$: a) $N = 30, \alpha = 0.1$, b) $N = 50, \alpha = 0.1$ c) $N = 100, \alpha = 0.1$, d) $N = 100, \alpha = 0.01$, e) $N = 150, \alpha = 0.1$, f) $N = 150, \alpha = 0.01$, g) $N = 200, \alpha = 0.1$, h) $N = 200, \alpha = 0.01$



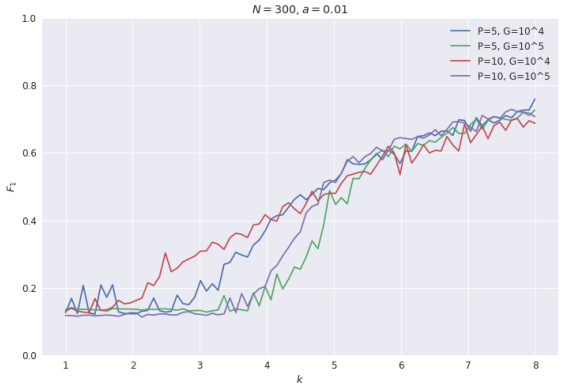
i)



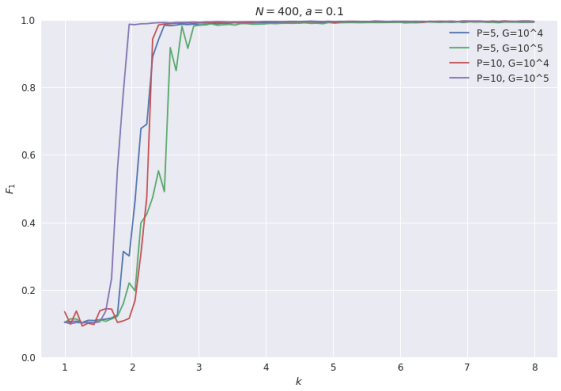
j)



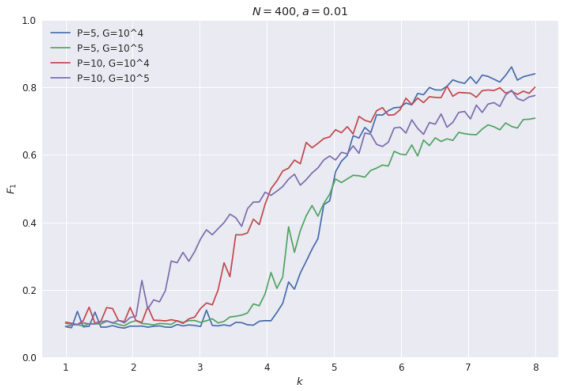
k)



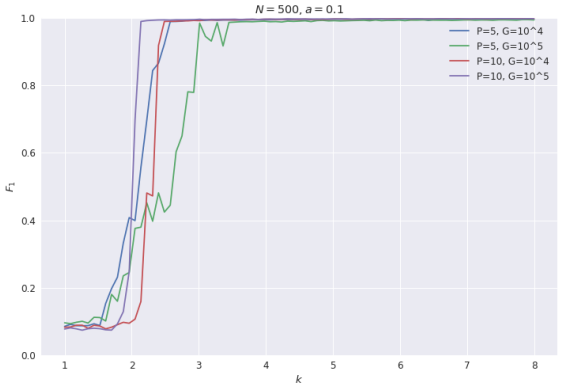
l)



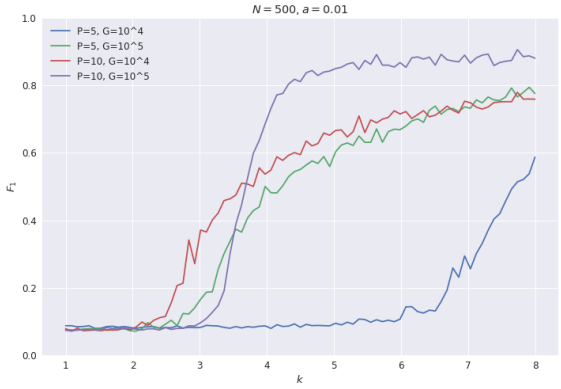
i)



j)



k)



l)

Рис. 3: Значения f1-score в зависимости от k и параметров N, α, P, G, где $P \in \{5, 10\}$, $G \in \{10^4, 10^5\}$: a) $N = 250, \alpha = 0.1$, b) $N = 250, \alpha = 0.01$ c) $N = 300, \alpha = 0.1$, d) $N = 300, \alpha = 0.01$, e) $N = 400, \alpha = 0.1$, f) $N = 400, \alpha = 0.01$, g) $N = 500, \alpha = 0.1$, h) $N = 500, \alpha = 0.01$