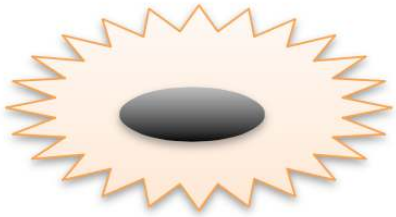
Кластеризация данных генной экспрессии при раке молочной железы

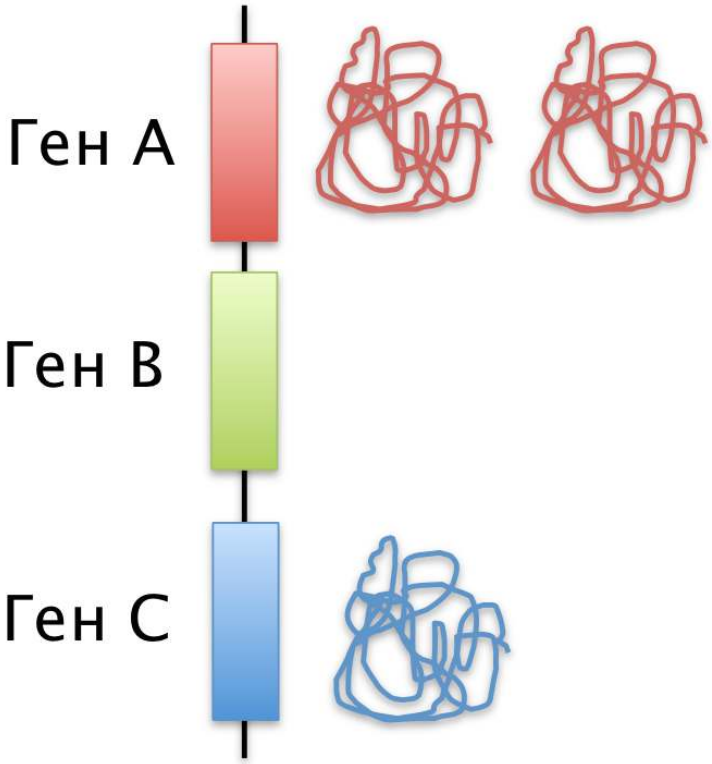
Дарья Вальтер,
НИУ ВШЭ, ФКН

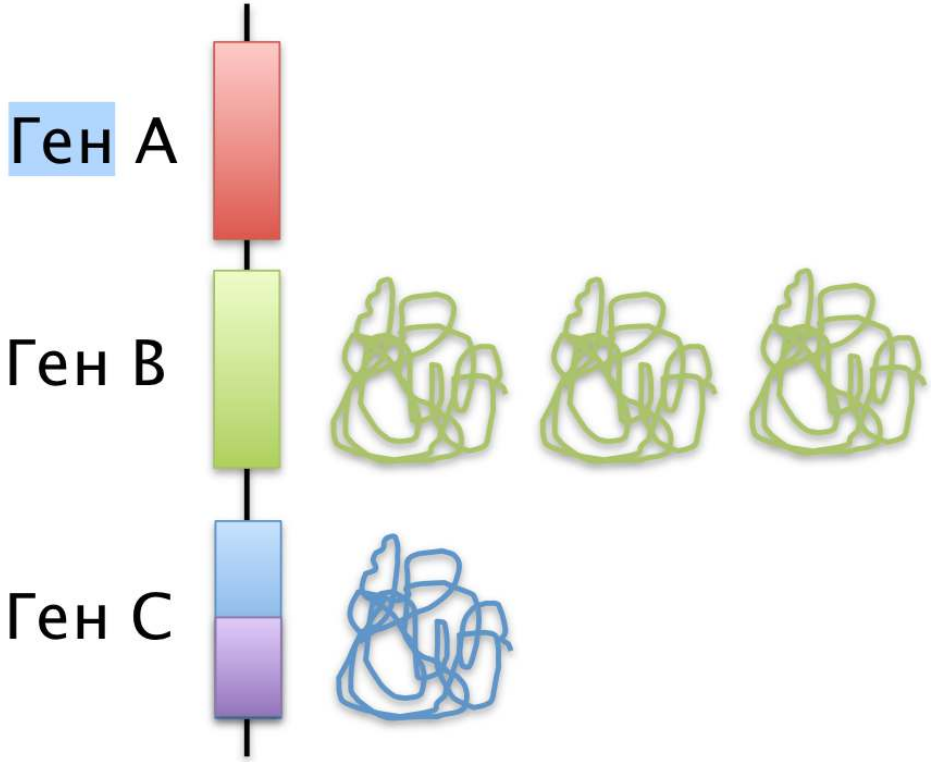
Экспрессия генов

Здоровая клетка



Раковая клетка





Что взять за биологические показатели?

Традиционная классификация рака молочной железы:

Subtype	These tumors tend to be*	Prevalence (approximate)
Luminal A	ER+ and/or PR+, HER2-, low Ki67	40%
Luminal B	ER+ and/or PR+, HER2+ (or HER2- with high Ki67)	20%
Triple negative/basal-like	ER-, PR-, HER2-	15-20%
HER2 type	ER-, PR-, HER2+	10-15%

Молекулярные подтипы рака молочной железы

Другие интересные признаки:

- локализация
- степень тяжести
- **метастазы**
- **рецидивы**
- и др.

Наши задачи:

- Реализовать алгоритм кластеризации k-means
- Для осуществления эффективной кластеризации

поэкспериментировать:

  - с подбором параметра k
  - со способами нормализации данных

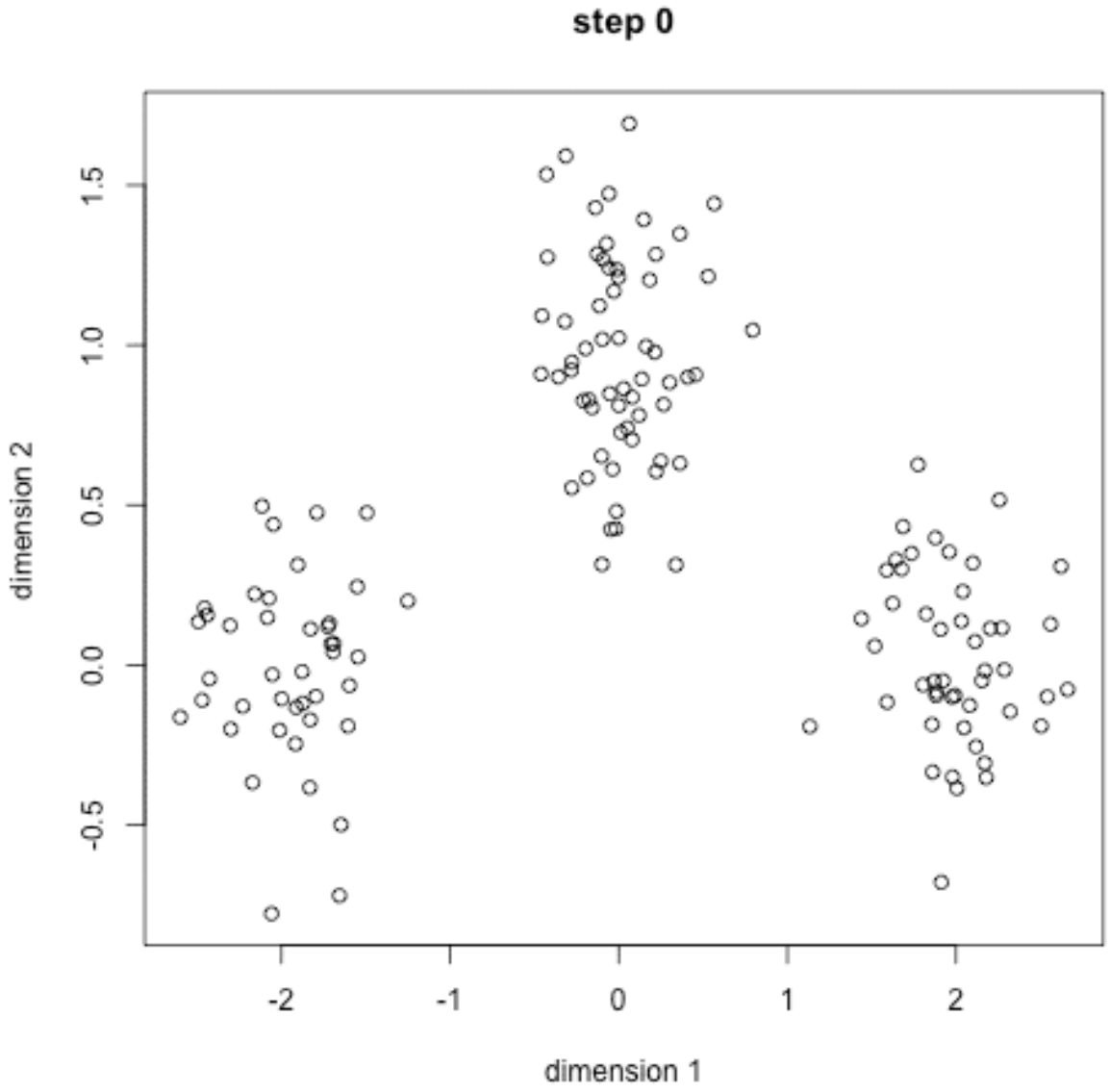
  И определить эффективный способ подбора количества кластеров разбиения.

- Сравнить полученное разбиение на кластеры с биологическими классами: по метастазам и релапсам

Данные: из БД GEO

UNIT_ID	GeneSymbol	GSM441624	GSM441625	GSM441626	GSM441627	GSM441628
100	RPL9	4253,4	5712,27	6005,97	5706,55	5567,75
101	DDX5	1506,54	1963,1	1150,61	944439	2267,46
102	RPL6	2731,83	3905,98	4906,7	2989,54	3368,69
103	CTDNEP1	357261	324745	292929	273438	195271
104	RPL10A	3493,52	6193,67	4194,32	4876,14	3491,67
105	CBX3	418,38	592737	551273	893827	696828
106	RPL17	3867,25	5875,4	5656,68	5541,95	5797,03
107	PSMB2	595993	570399	494426	692732	906249
108	KHDRBS1	681552	1008,83	862269	850577	940,48
109	ATP6V1G2	879858	974609	914309	890648	804,74

Алгоритм кластеризации k-means



1) Инициализация начальных центров

2) Распределение точек по кластерам относительно центров

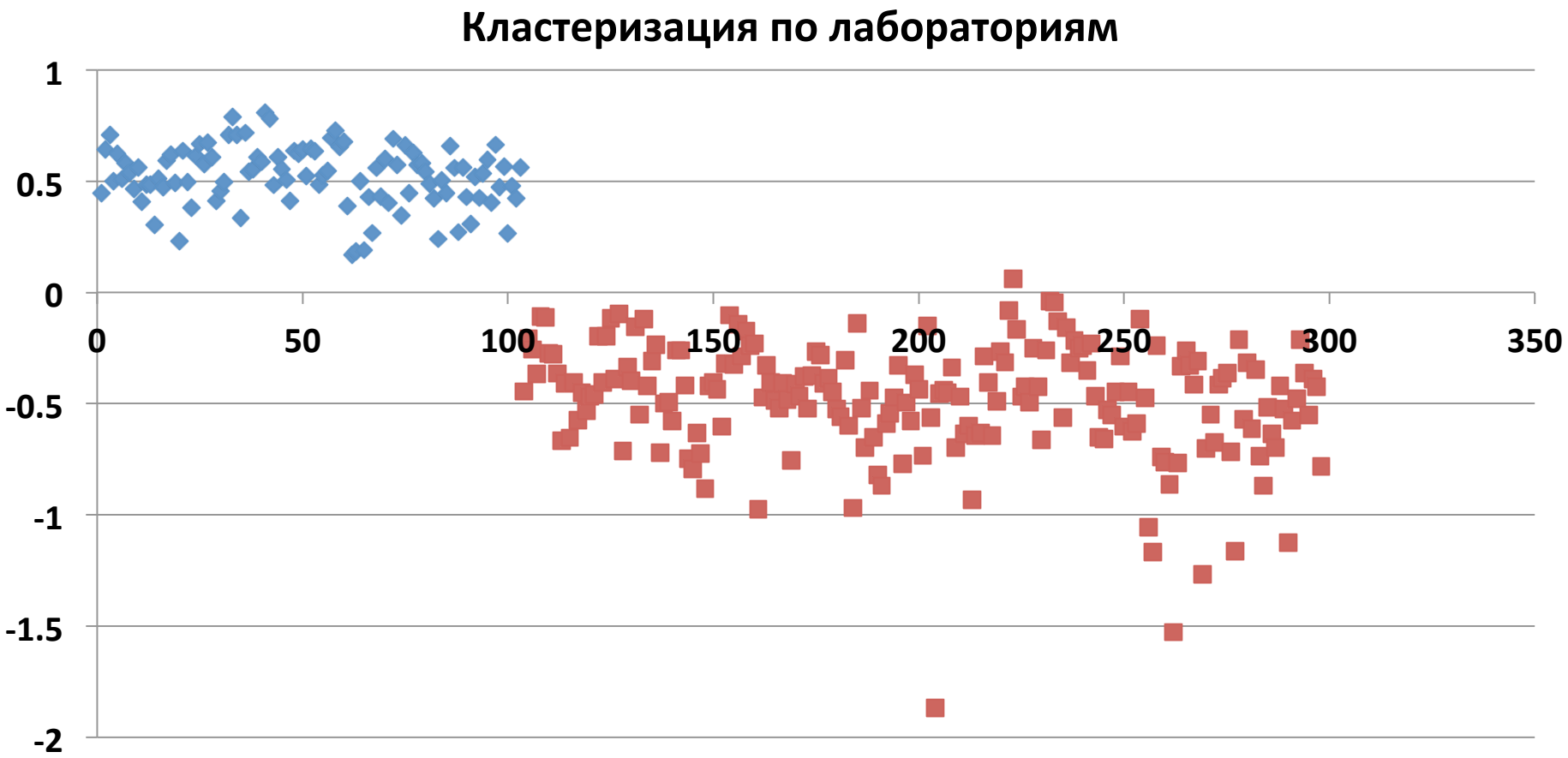
3) Новые центры – центры масс получившихся кластеров

Центры сходятся

Эксперименты…

- Со способами нормализации данных:
  1) Отдельно по каждому гену:

$$x = \frac{x - \min}{\max - \min}$$



Эксперименты

- Способы нормализации данных:

  2) Нормализация по консервативному гену

  - Практически все образцы отнеслись к одному кластеру: потеря информации

  3) Квантильная нормализация

  - Приведение распределений интенсивностей проб на микрочипах к одному распределению.
  - Наиболее эффективная

Эксперименты

- С подбором параметра k:

1. Эвристические методы

$$\textit{Average Inter Clusters} = \frac{\sum_{i<j, c(x_i)\neq c(x_j)} \rho(x_i, x_j)}{\sum_{i<j, c(x_i)\neq c(x_j)} 1} \longrightarrow \max$$

$$\textit{Average Within Cluster} = \frac{\sum_{i<j, c(x_i)= c(x_j)} \rho(x_i, x_j)}{\sum_{i<j, c(x_i)= c(x_j)} 1} \longrightarrow \min$$

- $c(x_i)$ – номер кластера,
- $\rho(x_i, x_j)$ – евклидово расстояние между точками

$$\frac{AvInterClusters}{AveragWithinClusters} \longrightarrow \max$$

Подбор параметра k

Average Silhouette Index = $\sum_i \frac{s(i)}{n}$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

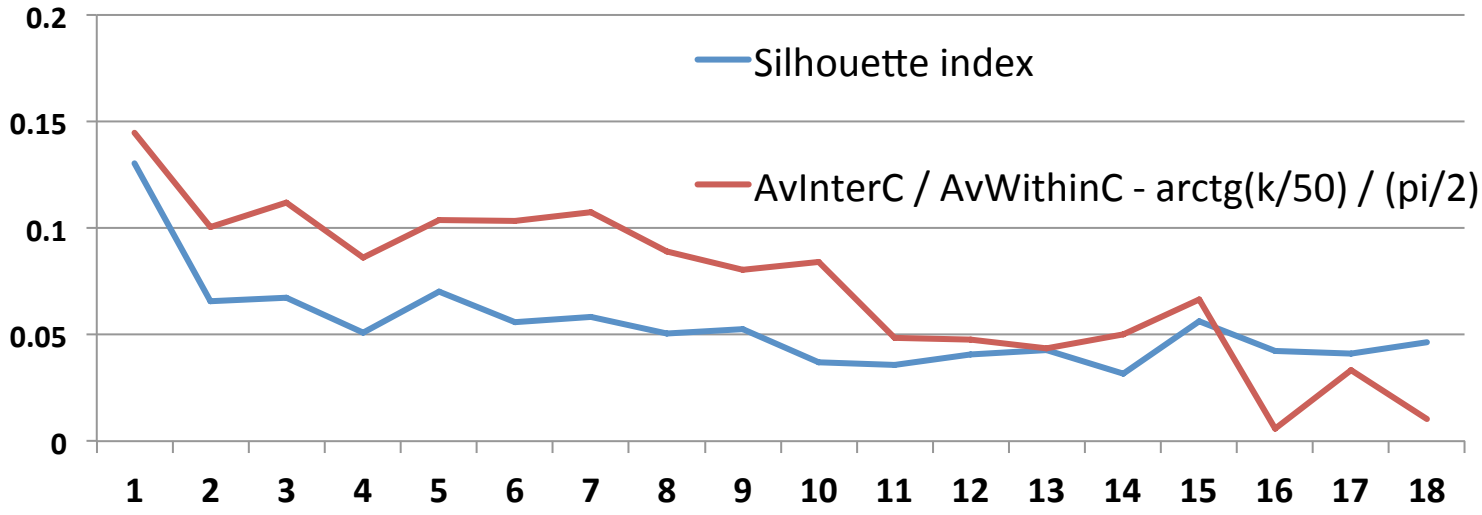
n – количество точек

a(i) – среднее расстояние между точкой i и остальными точками в том же кластере

b(i) – min(расстояний от i до центров других кластеров)

Характеризует выраженность кластеров:

- Удаленность друг от друга
- Компактность



Сравнение с биологической сутью

Сравниваем кластеры k-means и биологическое разбиение:

- По метастазам
- По релапсам
- По метастазам & релапсам

$$CRand = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} = \frac{\sum_{ij}\binom{n_{ij}}{2} - (\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2})/\binom{n}{2}}{\frac{1}{2}(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}) - (\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2})/\binom{n}{2}}$$

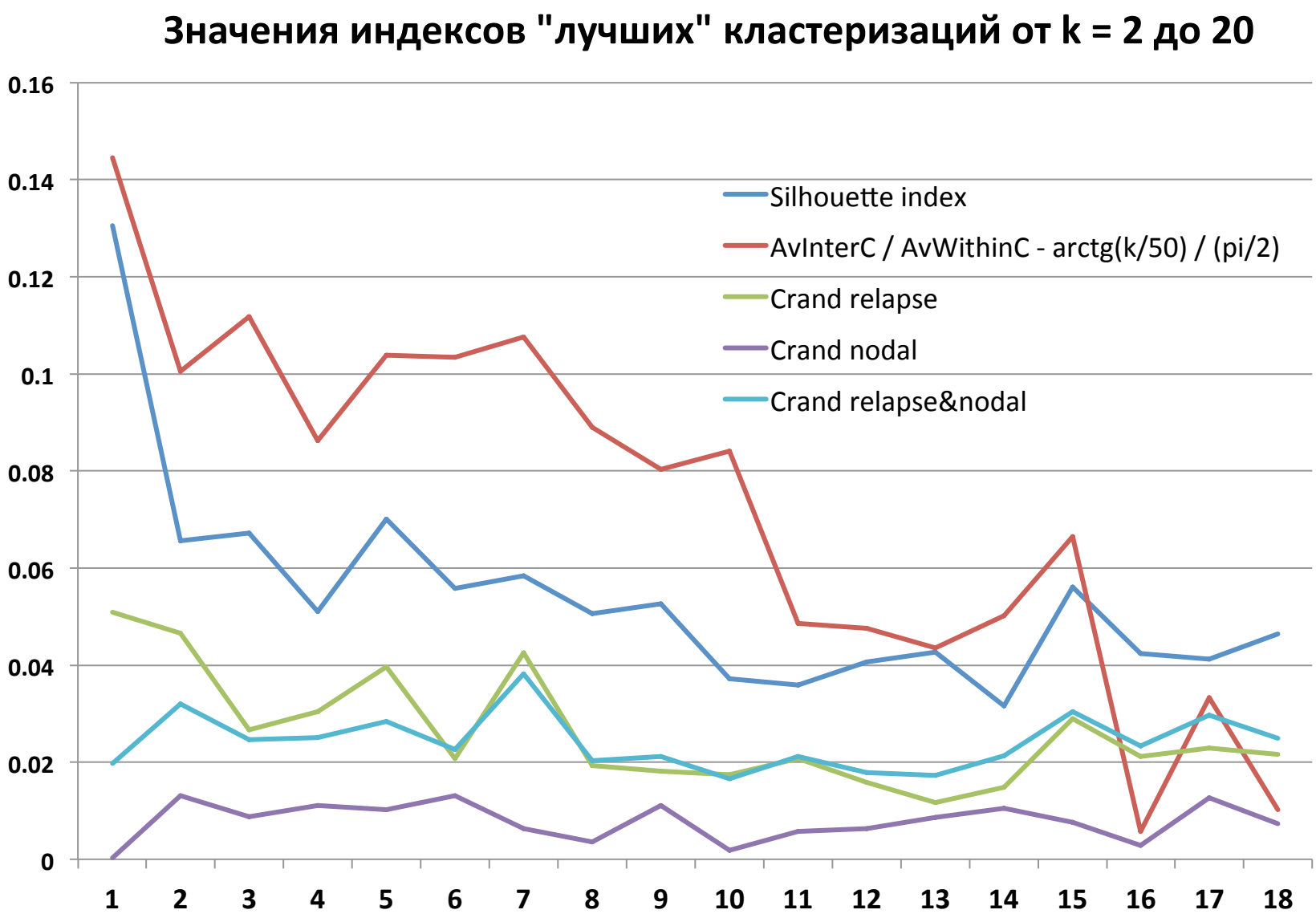
X\Y	Y_1	Y_2	$\cdots$	Y_s	Sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
Sums	b_1	b_2	$\cdots$	b_s	

$X_1, \ldots, X_r, Y_1, \ldots, Y_s$ – кластеры двух кластеризаций

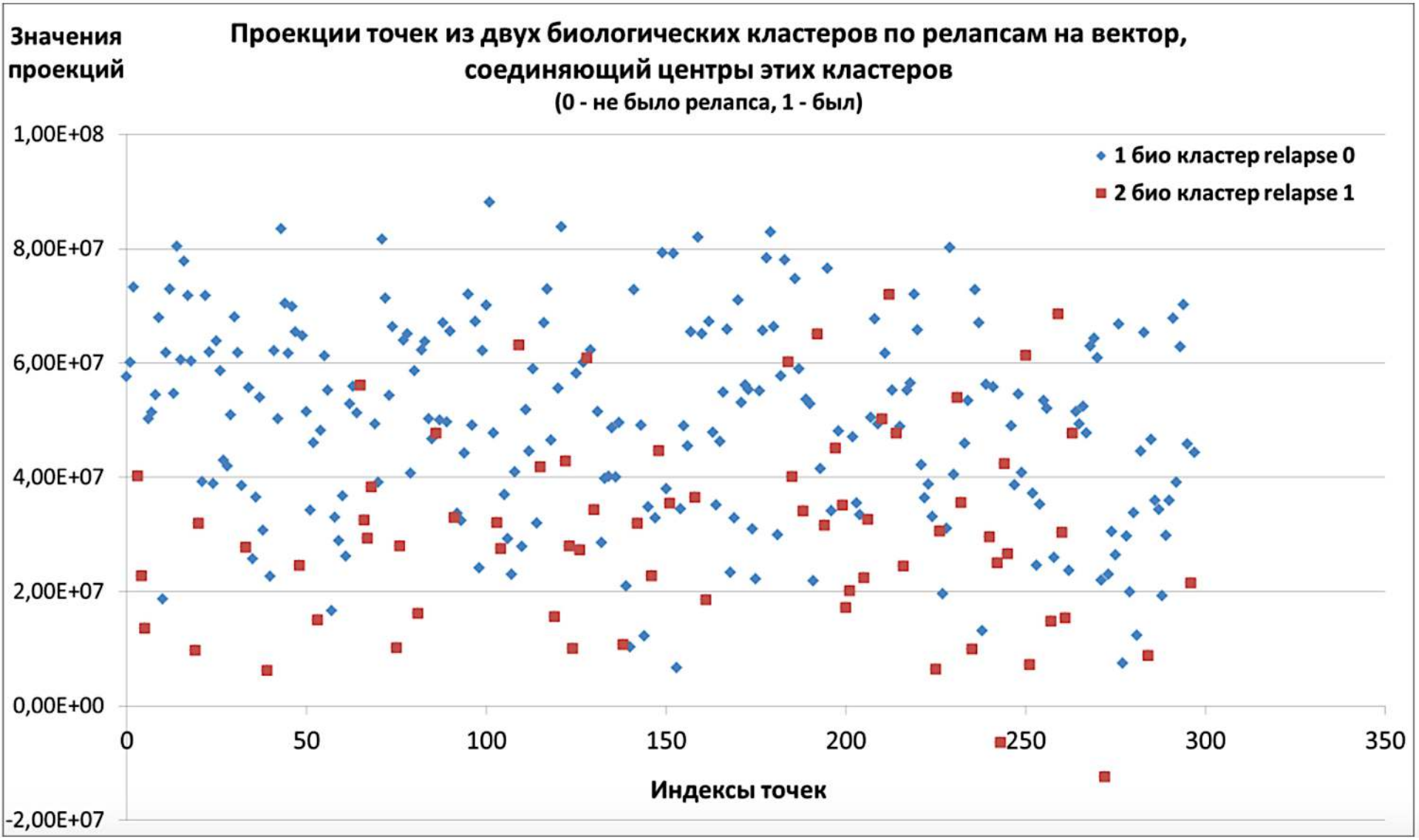
n_{ij} – количество совпадающих элементов в кластерах X_i, Y_j

a_i, b_i – количество элементов в кластере

В результате



Наблюдение: визуализация биологических кластеров по релапсам



Некоторые выводы

- Данные очень плохо кластеризуются
- Не разбиваются на классы по взятым биологическим характеристикам

...it is what it is

-Unknown

Следствие:

в нашей выборке на основании данных о раковых транскриптах нельзя сделать вывод о наличии метастаз и рецидивов у пациентов

Дальнейший путь развития

- Другие алгоритмы кластеризации,

  Другие пространственные метрики вместо Евклидовой,

  Другие способы нормализации данных
- Кластеризация не по пробам, а по генам!
- Enrichment analysis топа наиболее/ наименее экспрессированных генов для разных кластеров