# ПРЕДСКАЗАНИЕ ЗАБОЛЕВАНИЙ РАКА ГРУДИ ПО ГЕНОМНЫМ ДАННЫМ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ СНИЖЕНИЯ РАЗМЕРНОСТИ

**А. А. Горбунова**

*Белорусский государственный университет, г. Минск;*
*anastasia.gorbunova.so@yandex.ru;*
*науч. рук. – Н. Н. Яцков, канд. физ.-мат. наук, доц.*

В работе представлены результаты сравнительного анализа алгоритмов равномерного приближения и проекции, нейронной сети прямого распространения, ансамблевого алгоритма на основе методов главных и независимых компонент, стохастического вложения соседей с t-распределением, равномерного приближения и проекции, многомерного шкалирования, неотрицательного матричного разложения для классификации данных об экспрессии генов в заболевании рака груди. Выполнен анализ данных экспрессии микро- и информационной РНК молекул. Наиболее точным алгоритмом является нейронная сеть прямого распространения, точность классификации микроРНК – 69%, информационной РНК – 83%.

***Ключевые слова:*** экспрессия генов; микроРНК; информационная РНК; алгоритмы снижения размерности данных; ансамблевые алгоритмы; нейронные сети.

## ВВЕДЕНИЕ

Многомерные мультиомные данные являются стандартом в области исследования рака [1]. Новейшие геномные секвенаторы позволяют регистрировать большие наборы экспериментальных данных о нуклеотидном составе молекул ДНК/РНК, что усложняет визуализацию, интерпретацию и понимание результатов [2]. Алгоритмы снижения размерности данных устраняют указанные ограничения, при этом сохраняя важные свойства объекта исследования [3].

Наиболее перспективными методами снижения размерности данных являются: главных компонент (далее используется аббревиатура PCA от англ. principal component analysis), независимых компонент (ICA от англ. independent component analysis), стохастического вложения соседей с t-распределением (tSNE от англ. t-distributed stochastic neighbor embedding), равномерного приближения и проекции (UMAP от англ. uniform approximation and projection), многомерного шкалирования (MDS от англ. multidimensional scaling), неотрицательного матричного разложения (NMF от англ. non-negative matrix factorization), архитектура искусственных нейронных сетей автоэнкодер, а также варианты ансамблевых методов, основанных на стеккинге [1,4].

Рак груди – один из наиболее распространенных видов рака, являющийся гетерогенным заболеванием в отношении молекулярных изменений, клеточного состава и клинического исхода. Это разнообразие создает проблему при классификации опухолей, которые клинически полезны с точки зрения прогноза, что создает необходимость разработки предсказательных моделей заболевания по геномным данным. Большинство молекулярных исследований рака молочной железы сосредоточено на платформах с высоким содержанием информации, чаще всего на профилировании экспрессии мРНК или анализе числа копий ДНК, а в последнее время – на массовом параллельном секвенировании.

Цель работы – установить наиболее эффективный метод снижения размерности данных, регистрируемых в экспериментах геномного секвенирования по исследованию рака груди BRCA (от англ. Breast Cancer).

## АЛГОРИТМЫ СНИЖЕНИЯ РАЗМЕРНОСТИ ДАННЫХ

Рассмотрены нейронные сети автоэнкодеры. Автоэнкодер – нейронная сеть, состоящая из однослойного кодера с функцией активации гиперболический тангенс (Tanh от англ. TanhyPerbolic), которая сжимает данные до более низких измерений, и однослойного декодера с функцией активации Tanh, восстанавливающей исходный набор данных. Выходной слой кодера содержит сжатое представление исходных данных.

Стеккинг – технология ансамблевых алгоритмов, суть которой состоит в комбинировании результатов работы набора алгоритмов, в частности PCA, ICA, tSNE, UMAP, MDS, NMF [5]. После комбинирования результатов работы алгоритмов снижения размерности данных интегрированный набор проходит проверку на коррелированность признаков. Удаляются признаки с корреляцией выше 0.75. Полученный набор данных анализируется с помощью наилучшего, по результатам опубликованных работ [5], алгоритм снижения размерности данных UMAP [6].

## ЭКСПЕРИМЕНТАЛЬНЫЕ ДАННЫЕ

Исследованы данные об экспрессии генов опухолей рака груди, представленные наборами микроРНК и информационной РНК для различных групп пациентов. Исследование является практически значимым в области диагностики рака груди. Наборы данных

различаются по числу генов и пациентов, степени зашумленности, что позволяет точно оценить эффективность алгоритмов. Наборы данных содержат пять подтипов рака груди BRCA, LumA, LumB, Basal, Her2, и данные Normal для эталонной группы здоровых людей (таблица 1). Подтип заболевания используется для оценки эффективности алгоритмов снижения размерности данных при разделении пациентов на кластеры.

*Таблица 1*

**Количество пациетов и генов в наборах данных BRCA**

| Вид молекулы | Число пациентов | | | | | Число генов |
|---|---|---|---|---|---|---|
| | LumA | LumB | Basal | Her2 | Normal | |
| иРНК | 579 | 219 | 191 | 82 | 141 | 20126 |
| микроРНК | 407 | 139 | 133 | 58 | 110 | 625 |

## ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Выполнен анализ данных о микроРНК и информационной РНК молекулах с использованием методов снижения размерности данных и стэккинга. Эффективность алгоритмов оценена по трём критериям качества: 1) отношение средних внутрикластерных и межкластерных расстояний $Q_1$, 2) отношение суммы квадратов внутрикластерных и межкластерных расстояний $Q_2$, 3) точность классификации наборов данных в сниженном пространстве с использованием алгоритма случайного леса (с англ. – «Random forest») $acc_{RF}$ [7].

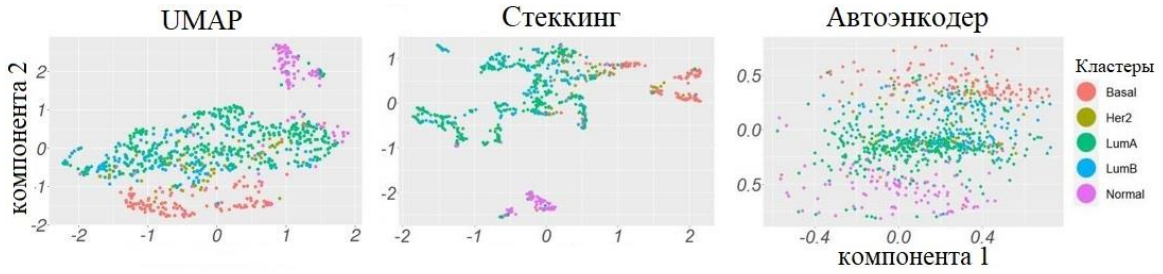
## РЕЗУЛЬТАТЫ

Реализованы ансамблевый алгоритм стеккинга на основе результатов методов PCA, ICA, tSNE, UMAP, MDS, NMF и сети автоэнкодер. Результаты сравнительного анализа алгоритмов для двух наборов данных представлены в таблице 2, диаграммы наиболее информативных компонент. Наилучшие результаты для данных микроРНК по критериям $Q_1$ и $Q_2$ получены в результате стеккинга: $Q_1$ = 32 и $Q_2$ = 29, по критерию точности – в результате работы автоэнкодера: $acc_{RF}$ = 69%, что ненамного ниже, чем в исходном пространстве признаков (74%). В результате стеккинга установлено более четкое разделение подтипов рака Basal, Normal и Her2, в сравнении с другими алгоритмами.

*Таблица 2*

**Оценки критериев качества работы алгоритмов на экспериментальных наборах данных**

| Набор BRCA | $Q_1$ | | | $Q_2$ | | | $acc_{RF}$, % | | |
|---|---|---|---|---|---|---|---|---|---|
| | UMAP | Стекк инг | Автоэн кодер | UMAP | Стекк инг | Автоэн кодер | UMAP | Стекк инг | Авто энко дер |
| иРНК | 25 | **25** | 32 | 15 | **14** | 24 | 79 | 78 | **83** |
| микро РНК | 38 | **32** | 45 | 37 | **29** | 50 | 69 | 67 | **69** |




Диаграммы разброса для кластеров данных miRNA в координатах двух наиболее информативных компонент, вычисленных методами UMAP, стеккинга и автоэнкодером

## ЗАКЛЮЧЕНИЕ

Выполнен сравнительный анализ методов снижения размерности данных для исследования экспрессии генов данных рака груди. Оптимальным алгоритмом является метод ансамблевый метод на основе стеккинга, точность которого на данных об экспрессии микроРНК – 69%, экспрессии информационной РНК – 83%.

### Библиографические ссылки


1. *Cantini, L., Zakeri, P., Hernandez, C. et al.* Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer / *L. Cantini, P. Zakeri, C. Hernandez* // Nature Communications, 2021. № 12. P. 124.
2. *Grinev V. V., Yatskou M. M., Skakun V. V., Chepeleva M. K., Nazarov P. V.* ORFhunteR: an accurate approach for the automatic identification and annotation of open reading frames in human mRNA molecules // Software Impacts. – 2022.
3. *Яцков, Н. Н.* Комплексный анализ данных при исследовании сложных бимолекулярных систем / *Н. Н. Яцков, В. В. Апанасович* // Информатика. – 2021. – Т. 18, № 1. – С. 105–122.
4. *Espadoto, M., Hirata, N. S. T., & Telea, A. C.* Self-supervised Dimensionality Reduction with Neural Networks and Pseudo-labeling. In C. Hurter, H. Purchase, J. Braz, & K. Bouatouch, IVAPP, 2021. P. 27-37.

5. *Gorbunova, A.* Comparative study of dimensionality reduction methods for multi-omics data / Anastasiya Gorbunova, Mikalai Yatskou, Petr Nazarov // Tumor Heterogeneity, Plasticity and Therapy. 5-6 May 2021. P. 77.
6. *McInnes, L. & Healy, J.* Umap: Uniform manifold approximation and projection for dimension reduction. arXiv – 2018. – № 18. – P. 1802.
7. *Горбунова, А.А.* Разработка и программная реализация алгоритмов снижения размерности данных биофизических экспериментов // 78-я научная конференция студентов и аспирантов Белорусского государственного университета: материалы конф. В 3 ч. Ч. 1, Минск, 2021 г. / Белорус. гос. ун-т; редкол.: В. Г. Сафонов (пред.) [и др.]. – Минск: БГУ, 2021.