ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
2024. Т. 6, № 3 (18). С. 61–70 http://siit.ugatu.su
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 6.60.606 DOI 10.54708/2658-5014-SIIT-2024-no3-p61 EDN RCXENH

СЕТЕВОЙ АНАЛИЗ ПРОФИЛЕЙ ЭКСПРЕССИИ ГЕНОВ

О. О. МИРАСОВ • Г. Р. ШАХМАМЕТОВА

Аннотация. В статье рассматривается предлагаемый процесс анализа генетических материалов, а именно профилей экспрессии генов, для выявления генов, связанных с возникновением заболевания. Рассматривается пайплайн сбора, загрузки и предобработки данных в программу. Представлена математическая постановка задачи, а именно: указаны алгоритмы обработки данных в виде метода главных компонент, линейного дискриминантного анализа, случайного леса, обобщенной линейной модели регрессии, модели Lasso, а также причины использования метода построения графов SCUDO. Приводится блок-схема последовательного применения моделей для достижения результата.

Ключевые слова: методы машинного обучения; профили экспрессии генов; рак; ДНК; геномные данные.

ВВЕДЕНИЕ

Сетевой анализ – мощный подход для изучения сложных взаимодействий и взаимоотношений между генами, белками и другими молекулами. Этот метод особенно актуален для исследований профилей экспрессии генов. Ключевые аспекты сетевого анализа включают в себя:

– Сети совместной экспрессии генов: создаются путем сопоставления профилей экспрессии генов в образцах для идентификации групп генов со схожими паттернами экспрессии. Для этой цели обычно используются такие инструменты, как WGCNA (анализ сети взвешенной коэкспрессии генов).

– Сети белок-белкового взаимодействия (PPI): объединение данных об экспрессии генов с известными взаимодействиями белков для идентификации ключевых регуляторных белков и их взаимодействий. Такие базы данных, как STRING и BioGRID, предоставляют ценные ресурсы для построения сетей PPI.

– Регуляторные сети: анализ взаимодействии между факторами транскрипции и их генами-мишенями. Эти сети помогают идентифицировать ключевые регуляторы изменений экспрессии генов в конкретных условиях.

– Сети путей: объединенный анализ путей с сетевым анализом используется, чтобы понять, как различные пути взаимодействуют и вносят свой вклад в процесс заболевания.

В данной работе предложен метод анализа профилей экспрессии генов при протоковой аденокарциноме поджелудочной железы.

ПОСТАНОВКА ЗАДАЧИ

Профилирование экспрессии генов – это мера активности (экспресии) генов в клетке. На текущий момент технологии позволяют высчитывать экспрессию всего генома в целом.

Собранные в лаборатории клетки при помощи микрочипов ДНК (DNA microarrays) или секвенирования РНК (RNA-Seq) проходят процесс, преобразующий «сырые» данные в закодированные последовательности, представляющие собой комбинации транскрипта клетки в виде последовательности нуклеотидных оснований и уникального молекулярного идентификатора (УМИ), призванного не допустить биологического или технического дублирования молекул (Рис. 1).

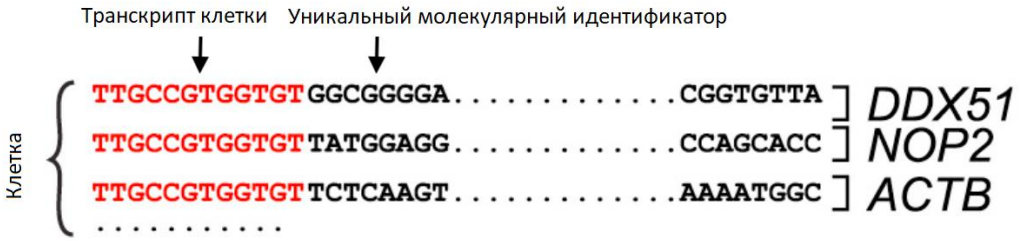

Рис. 1 Пример кодирования клетки.

Далее полученные данные используются для создания матрицы счетов, представляющей собой информацию о силе экспрессии (сколько раз последовательность ДНК превращалась в конечный продукт – белок или РНК) всех генов в каждой клетке. Информация поступает в виде файла, обычно формата BCL или FASTQ (Рис. 2), и с помощью специальных библиотек проходит процесс обработки, состоящий из:

- форматирования кодов и фильтрации шумов;
- демультиплексации образцов;
- картрирования кДНК;
- свертывания УМИ и оценки качества.

Рис. 2 Пример файла формата FASTQ.

Полученная в результате описанных выше действий матрица (Рис. 3) называется матрицей экспрессий и имеет размерность *M* строк (количество считанных генов) на *N* столбцов (количество клеток), на пересечении которых находится число – значение экспрессии гена в конкретной клетке.

	Клетка 1	Клетка 2		Клетка N
Ген 1	3	2	.	13
Ген 2	2	3	.	1
Ген 3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
Ген M	25	0	.	0

Рис. 3 Структура матрицы экспрессий.

После этого сформированная матрица подвергается контролю качества, необходимому для подтверждения корректности формирования матрицы, нормализации, коррекции (разбиение на партии) и выборе переменных генов.

Дальнейший анализ обычно состоит из:

- кластеризации клеток;
- идентификации маркеров;
- аннотации кластеров;
- дифференциальной экспрессии;
- вывода траекторий и динамики генов;
- композиционного анализа.

Весь процесс от предобработки данных до вариантов анализа представлен на Рис. 4.

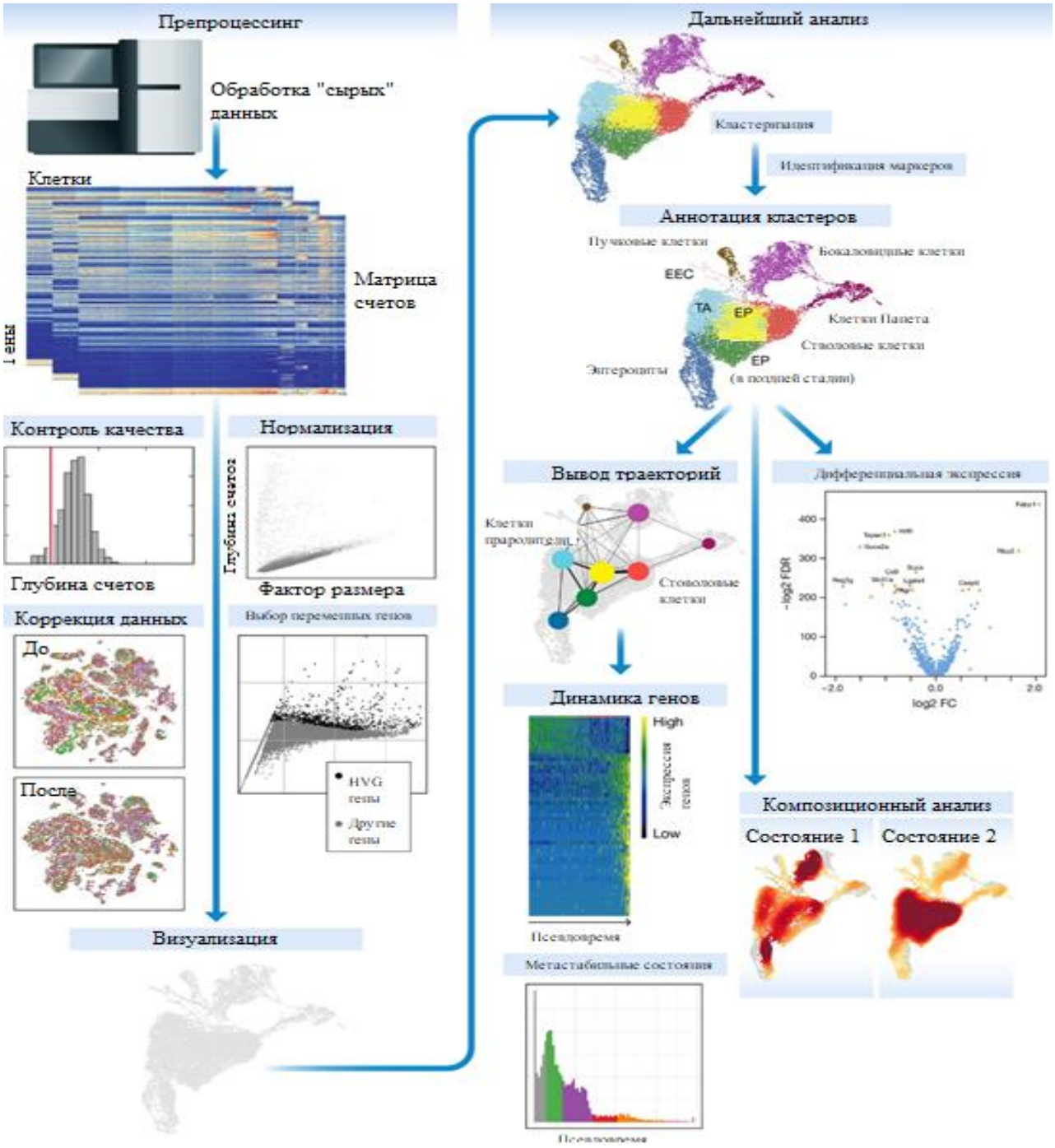

Рис. 4 Процесс работы с секвенированными данными РНК одиночной клетки.

В данном исследовании мы предлагаем взять различные сетевые модели, сравнить их эффективность на конкретном наборе данных и затем использовать лучшую модель для анализа путей и поиска генов, связанных с протоковой аденокарциномой поджелудочной железы.

Взятый для анализа датасет [1] содержит профили экспрессии генов с микрочипа ДНК 45 совпадающих пар опухоли поджелудочной железы и прилегающих неопухолевых тканей от 45 пациентов с аденокарциномой протоков поджелудочной железы. Исходная матрица экспрессий содержит данные о 90 клетках и 28869 генах.

На вход программы подается матрица экспрессий (Рис. 5), в которой в строках представлены гены и в столбцах – имена клеток. На пересечении столбцов и строк находится значение, соответствующее экспрессии гена в клетке.

	Prog_013	Prog_019	Prog_031	Prog_037	Prog_008	Prog_014	Prog_020	Prog_026	Prog_032	Prog_038	Prog_002
GNAI3	568	234	2	4	22	97	69	6	282	332	446
PBSN	0	0	0	0	0	0	0	0	0	0	0
CDC45	3	20	1	0	2443	210	4	4	145	257	254
H19	1	0	0	0	0	0	1	0	0	0	2
SCML2	0	0	0	0	0	16	1	0	0	0	1
APOH	0	0	0	0	0	0	0	0	0	0	0
NARF	1	414	0	52	352	288	245	0	87	128	27
CAV2	0	0	0	0	1	1	0	0	0	0	0
KLF6	35	709	2	9	2	12	6	7	2	19	1
SCMH1	309	11	3	88	152	2	2	1	0	22	0
COX5A	271	809	58	17	864	781	607	39	289	633	639
TBX2	0	0	1	0	0	0	0	0	0	0	0
TBX4	2	1	1	0	0	0	0	1	0	2	0
ZFY2	1	0	0	0	0	0	0	0	0	0	0
NGFR	0	0	1	0	1	2	0	1	1	0	0
WNT3	1	0	0	0	0	0	0	0	0	0	0
WNT9A	0	0	0	0	0	0	0	0	0	0	0
FER	3	0	0	0	1	0	0	0	0	0	0

Рис. 5 Пример матрицы экспрессий.

После сравнения моделей на выходе программа выдает сети путей, готовые для анализа и поиска раковых генов.

Математическая постановка задачи

Метод главных компонент (МГК)

Метод главных компонент используется для уменьшения размерности датасета, состоящего из большого количества взаимосвязанных переменных, сохраняя при этом как можно больше вариаций, присутствующих в наборе данных.

Определим матрицу A, которая имеет вид

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{x_{11}}\ \boldsymbol{x_{12}} \dots \boldsymbol{x_{1n}} \\ \boldsymbol{x_{21}}\ \boldsymbol{x_{22}} \dots \boldsymbol{x_{2n}} \\ \dots \\ \boldsymbol{x_{m1}}\ \boldsymbol{x_{m2}} \dots \boldsymbol{x_{mn}} \end{pmatrix}.$$

Выполним нормировку (стандартизацию) данных по столбцам.

$$\boldsymbol{x_{ij}} = \frac{x_{ij} - \bar{x}^j}{s^j}, \tag{1}$$

где $\bar{x}^j$ и s^j – оценка математического ожидания и среднеквадратическое отклонение по j-му столбцу (i = 1, …, m, j = 1, …, n).

Затем посчитаем матрицу ковариации. Ввиду проведенной стандартизации данных матрица ковариации будет корреляционной матрицей исходных данных порядка $p \times p$:

$$\boldsymbol{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X_i} - \bar{\boldsymbol{x}})(\boldsymbol{Y_i} - \bar{\boldsymbol{y}}) . \tag{2}$$

Вычисляем собственные числа и собственные вектора корреляционной матрицы, определяющие направления главных компонент, воспользовавшись алгоритмом метода Якоби. В результате получаем ковариационную матрицу главных компонент

$$\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}.$$

Проведем снижение размерности. Диагональные элементы ковариационной матрицы показывают дисперсию по исходной системе координат, а её собственные значения – по новой. Разделим дисперсию каждой компоненты на сумму всех дисперсий и получим долю дисперсии, связанную с каждой компонентой.

Иерархическая кластеризации методом WPGMA

Алгоритм WPGMA (группы взвешенных пар со средним арифметическим) строит корневое дерево (дендрограмму), которое отражает структуру, присутствующую в матрице попарных расстояний (или матрице подобия). На каждом шаге два ближайших кластера, скажем, объединяются в кластер более высокого уровня. Тогда его расстояние до другого кластера – это среднее арифметическое средних расстояний между членами кластера и k и i, и k и j:

$$d_{(i \cup j),k} = \frac{d_{i,k} + d_{j,k}}{2}. \tag{3}$$

Алгоритм WPGMA создает корневые дендрограммы и требует предположения о постоянной скорости: он создает ультраметрическое дерево, в котором расстояния от корня до каждой вершины ветвей равны. Это предположение об ультраметричности называется молекулярными часами, когда кончики включают данные ДНК, РНК и белка.

Случайный лес

Предположим, что обучающий набор микрочипов

$$D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \tag{4}$$

выбран случайно из (возможно, неизвестной) вероятности распределения $(x_i, y_i) \sim (X, Y)$. Цель: построить классификатор, который прогнозирует y на основе x на наборе данных примеров D. Дано: ансамбль (возможно, слабых) классификаторов

$$h = \{h_1(x), \ldots, h_K(x)\}. \tag{5}$$

Если каждый из $h_K(x)$ представляет собой дерево решений, то ансамбль – случайный лес. Определяем параметры $h_K(x)$ дерева решений для классификатора (к этим параметрам относится структура дерева, переменные разбиваются, в каком узле и т. д.).

$$\Theta = \{\Theta_{k1}, \ldots, \Theta_{kp}\}. \tag{6}$$

Иногда пишут

$$h_K(x) = h(x|\Theta_k). \tag{7}$$

Таким образом, дерево решений k приводит к классификатору.

Линейный дискриминантный анализ

Предположим, что у нас есть два класса и *d*-мерные образцы, такие как x_1, x_2, ..., x_n, где n_1 образцов из класса (c_1) и n_2 из класса (c_2). Если x_i – точка данных, то ее проекцию на линию, представленную единичным вектором v, можно записать как $v^T x_i$. Пусть u_1 и u_2 являются средними значениями выборок класса c_1 и c_2 соответственно до проецирования, а $\hat{u}$ обозначает среднее значение выборок класса после проецирования и может быть рассчитано по формуле

$$\widehat{u_1} = \frac{1}{n_1}\sum_{x_i \in c1}^{n_1} v^T xi = v^T u_1. \quad (8)$$

Аналогично

$$\widehat{u_2} = v^T u_2. \quad (9)$$

Теперь в LDA нам нужно нормализовать $u_1 - u_2$. Пусть $y_i = v^T x_i$ будет прогнозируемой выборкой, тогда разброс для выборок c_1 составляет:

$$s_1^2 = \sum_{y_i \in c_1}(y_1 - u_1)^2. \quad (10)$$

Аналогично

$$s_2^2 = \sum_{y_i \in c_1}(y_1 - u_2)^2. \quad (11)$$

Теперь нам нужно защитить наши данные на линии, имеющей максимальное направление v:

$$J(v) = \frac{\widehat{u_1} - \widehat{u_2}}{s_1^2 + s_2^2}. \quad (12)$$

Для максимизации приведенного выше уравнения нам нужно найти вектор проекции, который максимизирует разницу средств уменьшения разбросов обоих классов. Теперь матрица разброса s_1 и s_2 классов c_1 и c_2 равна:

$$s_1 = \sum_{x_i \in c_1}(x_i - u_1)(x_i - u_1)^T; \quad (13)$$

$$s_2 = \sum_{x_i i \in c_2}(x_i - u_2)(x_i - u_2)^T. \quad (14)$$

Теперь, чтобы максимизировать приведенное выше уравнение, нужно продифференцировать его по v:

$$Mv = \lambda v, \quad (15)$$

где $\lambda = \frac{v^T s_b v}{v^T s_w v}$, $M = s_w^{-1} s_b$. Здесь в качестве максимального значения $J(v)$ мы будем использовать значение, соответствующее наибольшему собственному значению. Это предоставит нам лучшее решение для LDA.

Линейная регрессия

Определим модель зависимости как

$$y_i = w_1 + w_2 x_i + \varepsilon_i; \quad (16)$$

Согласно методу наименьших квадратов, искомый вектор параметров $w = (w_1, w_2)^T$ есть решение нормального уравнения

$$w = (A^T A)^{-1} A^T y, \quad (17)$$

где y – вектор, состоящий из значений зависимой переменной, $y = (y_1, \dots, y_m)$. Столбцы матрицы A есть подстановки значений свободной переменной $x_i^0 \to a_{i1}$ и $x_i^1 \to a_{i2}, i = 1, \dots, m$. Матрица имеет вид

$$A = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix}.$$

Зависимая переменная восстанавливается по полученным весам и заданным значениям свободной переменной

$$y_i^* = w_1 + w_2 x_i \,, \tag{18}$$

иначе

$$y^* = Aw \,. \tag{19}$$

Отрицательная биномиальная регрессия

В отрицательной биномиальной регрессии среднее значение y определяется временем экспозиции t и набором k регрессирующих переменных. Выражение, связывающее эти величины, имеет вид

$$\mu_i = \exp\left(\ln(t_i) + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}\right) . \tag{20}$$

Заметим, что $X_1 \equiv 1$ и β_1 называется точкой пересечения. Коэффициенты регрессии $\beta_1, \beta_2, .., \beta_k$ – неизвестные параметры, которые оцениваются по набору данных. Используя эти обозначения, запишем фундаментальную модель отрицательной биномиальной регрессии для наблюдения i:

$$\Pr\left(Y_i = y_i \middle| \mu_i, \alpha\right) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} . \tag{21}$$

Алгоритм анализа

В работе представляется пайплайн нахождения топ-генов протоковой аденокарциномы поджелудочной железы (рис. 6).

Перед анализом предлагается провести предобработку в виде лог-трансформации, метода главных компонент и кластеризации. На предобработанных данных предлагается построить несколько моделей для дальнейшего выбора лучшей из них, на основе которой будут выбраны наиболее значимые гены. Использованные модели:

- случайный лес [2];
- линейный дискриминантный анализ [3];
- обобщенная линейная модель [4];
- модель Lasso [5];
- SCUDO [6].

Среди моделей по заранее определенным метрикам выбирается модель с лучшими показателями, которая в дальнейшем используется для определения наиболее топ-генов по принципу значимости (importance). Затем для каждого топ-гена реализуется метод GOST и находится сетевой путь среди баз KEGG, GO, Reactome.

Заключение

Предложенный процесс обработки данных может быть использован для анализа профилей экспрессии генов при различных заболеваниях. Внедрение в процесс анализа нескольких моделей позволяет улучшить качество материала и подбора лучшей модели для каждого конкретного набора данных. Несмотря на проблемы с качеством данных и модификацией методов машинного обучения для обработки геномных данных, этот подход имеет все шансы помочь в определении генов, связанных с анализируемым заболеванием.

Дальнейшее развитие этой темы планируется в направлении практической реализации рассмотренных методов.

При анализе предметной области использованы материалы исследований [7–18] по интеллектуальной обработке и анализу биомедицинских данных, ведущихся на кафедре вычислительной математики и кибернетики Уфимского университета науки и технологий.

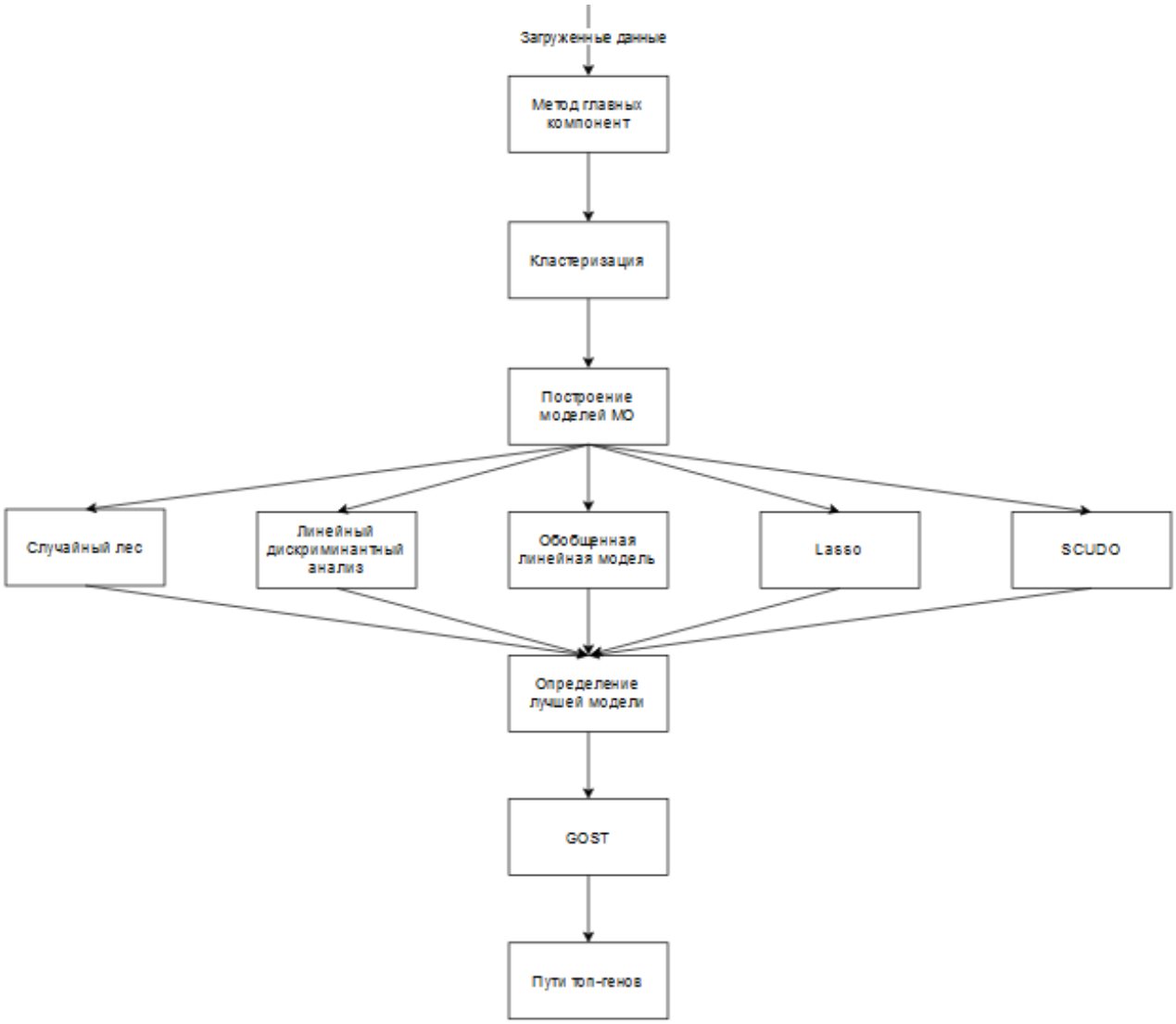

Рис. 6 Блок-схема алгоритма обработки данных.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Microarray gene-expression profiles of 45 matching pairs of pancreatic tumor and adjacent non-tumor tissues from 45 patients with pancreatic ductal adenocarcinoma. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi (дата обращения 29.05.2024)
2. Díaz-Uriarte R., Alvarez de Andrés S. Gene selection and classification of microarray data using random forest // BMC Bioinformatics. 7. 3 (2006). DOI 10.1186/1471-2105-7-3.
3. Huang D., Quan Y., He M., et al. Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data // J Exp Clin Cancer Res. 28. 149 (2009). DOI 10.1186/1756-9966-28-149.
4. Xu X. et al. Gsw-fi: a GLM model incorporating shrinkage and double-weighted strategies for identifying cancer driver genes with functional impact // BMC Bioinformatics. 25. 99 (2024). DOI 10.1186/s12859-024-05707-8.
5. Huang H. H., Rao H., Miao R., et al. A novel meta-analysis based on data augmentation and elastic data shared lasso regularization for gene expression // BMC Bioinformatics. 23 (Suppl 10). 353 (2022). DOI 10.1186/s12859-022-04887-5.
6. Lauria M., Moyseos P., Priami C. SCUDO: a tool for signature-based clustering of expression profiles // Nucleic Acids Research. 43(W1). 2015.W188–W192. DOI 10.1093/nar/gkv449.
7. Шапошникова А. С., Богданов М. Р. Определение сердечного ритма плода по неинвазивному ЭКГ с применением различных фильтров // СИИТ. 2023. Т. 5. № 6(15). С. 32-37. DOI 10.54708/2658-5014-SIIT-2023-no5-p32. EDN WBBOVK. [[Shaposhnikova A. S., Bogdanov M. R. “Determination of fetal heart rate by non-invasive ECG using various filters” // SIIT. 2023. Vol. 5, No. 6(15), pp. 32-37. DOI 10.54708/2658-5014-SIIT-2023-no5-p32. EDN WBBOVK. (In Russian).]]
8. Шахмаметова Г. Р., Ахметшин А. А. Обзор современного состояния исследований в области применения машинного обучения в обработке ПГИА данных // Высшая школа: научные исследования: Мат-лы Межвузовского международного конгресса, Москва, 26 мая 2023 г. Т. 2. М: Инфинити, 2023. С. 127-140. EDN USBHTE. [[Shakhmametova G. R., Akhmetshin A. A. “Review of the current state of research in the field of machine learning application in processing PGIA data” // Higher school:

scientific research: Proceedings of the Interuniversity International Congress, Moscow, May 26, 2023. Vol. 2. Moscow: Infinity, 2023, pp. 127-140. EDN USBHTE. (In Russian).]]

9. Yusupova N., Zulkarneev R., Rizvanov D., et al. Classification of interaction participants in the formation of the trajectory of diagnosis and treatment of bronchopulmonary diseases to design agents of a multi-agent system // Software Engineering Application in Systems Design: Proceedings of 6th Computational Methods in Systems and Software. 2022. Czech Republic. October 10–15. 2022. Vol. 596. Switzerland: Springer Nature Switzerland AG. Part of Springer Nature, 2023, pp. 723-731. DOI 10.1007/978-3-031-21435-6_62. EDN GWZYNZ.
10. Shakhmametova G. R., Evgrafov A. A., Zulkarneev R. Kh. Development of data storage and user interface in the clinical decision support system // Software Engineering Research in System Science: Proceedings of 12th Computer Science On-line Conference. 2023. Zlin, Czech Republic, 01–30 апреля 2023 года. Vol. 722-1. Springer Nature Switzerland AG: Springer Nature Switzerland AG, 2023, pp. 808-816. DOI 10.1007/978-3-031-35311-6_75. EDN WICSIR.
11. Зиновьев М. С., Нургаянова О. С. Оценка индивидуального риска развития сахарного диабета второго типа и возможных осложнений // СИИТ. 2023. Т. 5. № 4(13). С. 101-110. DOI 10.54708/2658-5014-SIIT-2023-no5-p101. EDN HIIXFH. [[Zinoviev M. S., Nurgayanova O. S. “Assessment of individual risk of developing type 2 diabetes mellitus and possible complications” // SIIT. 2023. Vol. 5, No. 4(13), pp. 101-110. DOI 10.54708/2658-5014-SIIT-2023-no5-p101. EDN HIIXFH. (In Russian).]]
12. Юсупова Н. И., Нургаянова О. С., Зулкарнеев Р. Х. Формализация этапов риск-анализа в СППР с учетом оценок клинических рисков при бронхолегочных заболеваниях // СИИТ. 2023. Т. 5. № 1(10). С. 11-24. DOI 10.54708/2658-5014-SIIT-2023-no1-p11. EDN KHIIHT. [[Yusupova N. I., Nurgayanova O. S., Zulkarneev R. Kh. “Formalization of risk analysis stages in decision support system taking into account clinical risk assessments for bronchopulmonary diseases” // SIIT. 2023. Vol. 5, No. 1(10), pp. 11-24. DOI 10.54708/2658-5014-SIIT-2023-no1-p11. EDN KHIIHT. (In Russian).]]
13. Шахмаметова Г. Р., Береговая С. П., Христодуло А. Д. Анализ эндокринологических данных на основе моделей классификации // СИИТ. 2022. Т. 4. № 2(9). С. 30-36. DOI 10.54708/26585014_2022_42930. EDN LBZVZL. [[Shakhmametova G. R., Beregovaya S. P., Khristodullo A. D. “Analysis of endocrinological data based on classification models” // SIIT. 2022. Vol. 4, No. 2(9), pp. 30-36. DOI 10.54708/26585014_2022_42930. EDN LBZVZL. (In Russian).]]
14. Насыров Р. В. Причинный подход к построению бионических вычислений на основе рекурсивных моделей анализа данных // СИИТ. 2022. Т. 4. № 1(8). С. 27-36. DOI 10.54708/26585014_2022_41827. EDN UOMMOU. [[Nasyrov R. V. “Causal approach to the construction of bionic computations based on recursive models of data analysis” // SIIT. 2022. Vol. 4, No. 1(8), pp. 27-36. DOI 10.54708/26585014_2022_41827. EDN UOMMOU. (In Russian).]]
15. Shakhmametova G., Yusupova N., Zulkarneev R., Khudoba Ye. Concept map for clinical recommendations data and knowledge structuring // Proceedings of the 8th International Conference on Applied Innovations in IT. Koethen. Germany. March 10. 2020. Vol. 8. Issue 1. Koethen, Germany: Anhalt University of Applied Sciences, 2020, pp. 71-76. EDN PXBUZZ.
16. Юсупова Н. И., Гаянова М. М., Богданов М. Р., Юсупова, Н. И. Извлечение информации об использовании информационных технологий для поддержки принятия решений в медицинской диагностике // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. 2022. Т. 22. № 1. С. 14-27. DOI 10.14529/ctcr220102. EDN HSTNZS. [[Yusupova N. I., Gayanova M. M., Bogdanov M. R. Yusupova, N. I. “Extracting information on the use of information technologies to support decision-making in medical diagnostics” // Bulletin of the South Ural State University. Series: Computer technologies, control, radio electronics. 2022. Vol. 22, No. 1, pp. 14-27. DOI 10.14529/ctcr220102. EDN HSTNZS. (In Russian).]]
17. Шахмаметова Г. Р., Христодуло А. Д., Береговая С. П. Анализ эндокринологических данных на основе моделей классификации // СИИТ. 2022. Т. 4. № 2(9). С. 30-36. DOI 10.54708/26585014_2022_42930. EDN LBZVZL. [[Shakhmametova G. R., Khristodullo A. D., Beregovaya S. P. “Analysis of endocrinological data based on classification models” // SIIT. 2022. Vol. 4, No. 2(9), pp. 30-36. DOI 10.54708/26585014_2022_42930. EDN LBZVZL. (In Russian).]]
18. Мирасов О. О. Анализ современного состояния исследований в области обработки геномных данных методами машинного обучения // Мавлютовские чтения: Мат-лы XV Всероссийской молодежной научной конференции: В 7 тт. Уфа, 26–28 октября 2021 г. Т. 4. Уфа: УГАТУ, 2021. С. 262-268. EDN DRAWPI. [[Mirasov O. O. “Analysis of the current state of research in the field of genomic data processing using machine learning methods” / scientific supervisor G. R. Shakhmametova // Mavlyutov Readings: materials of the XV All-Russian youth scientific conference, Ufa, October 26-28, 2021. Volume 4, pp. 262-268. EDN DRAWPI. (In Russian).]]

Поступила в редакцию 7 июня 2024 г.

МЕТАДАННЫЕ / METADATA

Title: Network analysis of gene expression profiles.

Abstract: The article discusses the proposed process of analyzing genetic materials, namely gene expression profiles, to identify genes associated with the occurrence of the disease. The pipeline for collecting, loading and preprocessing data into the program is considered. A mathematical formulation of the problem is presented, namely: data processing algorithms are indicated in the form of the principal component method, linear discriminant analysis, random forest, generalized linear regression model, Lasso model, as well as the reasons for using the SCUDO graph construction method. A block diagram of the sequential application of models to achieve results is provided.

Key words: machine learning methods; gene expression profiles; cancer; DNA; genomic data.

Язык статьи / Language: русский / Russian.

Об авторах / About the authors:

МИРАСОВ Олег Олегович
Уфимский университет науки и технологий, Россия.
Магистрант ин-та информатики, математики и робототехники. Дипл. программист-математик (Уфимск. гос. авиац. техн. ун-т, 2022). Иссл. в обл. биоинформатики
E-mail: helgu76@gmail.com

ШАХМАМЕТОВА Гюзель Радиковна
Уфимский университет науки и технологий, Россия.
Зав. каф. вычислительной математики и кибернетики. Дипл. инж. по инф. системам (Уфимск. авиац. ин-т, 1992). Д-р техн. наук по сист. анализу, управлению и обработке информации (Уфимск. гос. авиац. техн. ун-т, 2013). Иссл. в обл. интеллект. поддержки принятия решений, распознавания образов, обработки биомедицинских данных методами машинного обучения и искусственного интеллекта.
E-mail: shakhgouzel@mail.ru
ORCID: http://orcid.org/0000-0002-7742-793X

MIRASOV Oleg Olegovich
Ufa University of Science and Technologies, Russia.
Master's student, Institute of Informatics, Mathematics, and Robotics. Graduated programmer-mathematician (Ufa State Aviation Technical University, 2022).
E-mail: helgu76@gmail.com

SHAKHMAMETOVA Gyuzel Radikovna
Ufa University of Science and Technologies, Russia.
Head of the Department of Computational Mathematics and Cybernetics. Dipl. Eng. on information systems (Ufa Aviation Institute, 1992). Dr. Tech. Sciences in system analysis, management and information processing (Ufa State Aviation Technical University, 2013). Research in the field of intelligent decision support, pattern recognition, processing of biomedical data using machine learning and artificial intelligence methods..
E-mail: shakhgouzel@mail.ru
ORCID: http://orcid.org/0000-0002-7742-793X