МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Кафедра системного анализа и компьютерного моделирования

КРИЦКИЙ Андрей Олегович

ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗИ МЕТАБОЛИЧЕСКОЙ АКТИВНОСТИ И ЭКСПРЕССИИ ГЕНОВ В РАКОВЫХ КЛЕТКАХ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Аннотация (реферат) дипломной работы

Научный руководитель:

Доцент, кандидат технических наук.

Е. В. Лисица

Научный консультант:
М. К. Чепелева

Допущена к защите
“_____”_____2024 г.
Заведующий кафедрой системного анализа и компьютерного моделирования
кандидат физико-математических наук, доцент Н. Н. Яцков

Минск, 2024

РЕФЕРАТ

Структура и объём дипломной работы: дипломная работа состоит из введения, 3 глав, заключения, списка использованных источников и 1 приложения. Объем дипломной работы составляет 47 страниц. В работе использовано 24 изображений, 2 таблицы. В процессе написания работы было использовано 29 источников.

Ключевые слова: рак, экспрессия генов, метаболит, метаболизм, секвенирование, машинное обучение, Python.

Объектом исследования дипломной работы стали: экспериментальные данные секвенирования одиночных клеток с помощью методов машинного обучения для предсказания уровней метаболитов в клетка.

Актуальность данной работы заключается в необходимости установления взаимосвязи экспрессии генов и уровнями метаболизмов в клетках.

Цель данной работы – исследование данных экспрессии генов, для предсказания уровней метаболитов с помощью методов машинного обучения.

Для достижения поставленной цели были решены следующие задачи:

1) Изучена литература по секвенированию одиночных клеток.

2) Изучены библиотеки машинного обучения такие как: pandas, numpy,matplotlib, seaborn. Отдельные методы библиотек: sklearn и plotly.

3) Выполнена программная реализация алгоритма исследования экспериментальных данных экспрессии генов и уровня метаболитов.

Методы исследования: изучение теоретической информации о методах секвенирования генов, изучение проблемы интеграции метаболомики и транскриптомики. Изучение методов обработки данных, метод главных компонент, метод независимых компонент. А также моделей машинного обучения: линейной регрессии, случайного леса и градиентного бустинга. Анализ и изучение библиотек анализа данных для работы с данными экспрессии генов, а также библиотек машинного обучения и доступных технологий для выполнения программной реализации алгоритма.

Рэферат

Структура і аб'ём дыпломнай працы: дыпломная праца складаецца з увядзення, 3 кіраўнікоў, заключэння, спісу выкарыстаных крыніц і 1 дадатку. Аб'ём дыпломнай працы складае 47 старонак. У рабоце выкарыстана 24 відарысы, 2 табліцы. У працэсе напісання працы было выкарыстана 29 крыніц.

Ключавыя словы: рак, экспрэсія генаў, метабаліт, метабалізм, секвеніраванне, машыннае навучанне, Python.

Аб'ектам даследавання дыпломнай працы сталі: эксперыментальныя дадзеныя секвеніравання адзіночных клетак з дапамогай метадаў машыннага навучання для прадказання узроўняў метабалітаў у клетка.

Актуальнасць дадзенай працы заключаецца ў неабходнасці ўстанаўлення ўзаемасувязі экспрэсіі генаў і ўзроўнямі метабалізмаў у клетках.

Мэта дадзенай працы - даследаванне дадзеных экспрэсіі генаў, для прадказання узроўняў метабалітаў з дапамогай метадаў машыннага навучання.Для дасягнення пастаўленай мэты былі вырашаны наступныя задачы:

1) Вывучана літаратура па секвеніраванні адзіночных клетак.
2) Вывучаны бібліятэкі машыннага навучання такія як: pandas, numpy, matplotlib, seaborn. Асобныя метады бібліятэк: sklearn і plotly.
3) Выканана праграмная рэалізацыя алгарытму даследавання эксперыментальных дадзеных экспрэсіі генаў і ўзроўню метабалітаў.

Метады даследавання: вывучэнне тэарэтычнай інфармацыі аб метадах секвеніравання генаў, вывучэнне праблемы інтэграцыі метабаломікі і транскрыптомікі. Вывучэнне метадаў апрацоўкі даных, метад галоўных кампанент, метад незалежных кампанент. А таксама мадэляў машыннага навучання: лінейнай рэгрэсіі, выпадковага лесу і градыентнага бустынгу. Аналіз і вывучэнне бібліятэк аналізу даных для работы з данымі экспрэсіі генаў, а таксама бібліятэк машыннага навучання і даступных тэхналогій для выканання праграмнай рэалізацыі алгарытму.

ABSTRACT

Structure and scope of the thesis: the thesis work of the head of administration, 3 chapters, managers, teachers, sources used and 1 appendix. The volume of the thesis is 47 pages. The work used 24 images, 2 tables. In the process of writing the work, 29 sources were used.

Keywords: cancer, gene expression, metabolite, metabolism, sequencing, machine learning, Python.

Object of study thesis: Experimental single-cell sequencing data using machine learning methods to predict metabolite levels in a cell.

Relevance of this work lies in the need to establish the relationship between gene expression and metabolic levels in cells.

The purpose of this work is exploring gene expression data to predict metabolite levels using machine learning methods.

To achieve this goal, the following tasks were solved:

1) Study the literature on single cell sequencing

2) Studied machine learning libraries such as: pandas, numpy, matplotlib, seaborn. Selected library methods: sklearn and plotly.

3) Perform a software implementation of an algorithm for predicting metabolites based on gene expression data.

Research methods: studying theoretical information about gene sequencing methods, studying the problem of integrating metabolomics and transcriptomics. Study of data processing methods, principal component method, independent component method. As well as machine learning models: linear regression, random forest and gradient boosting. Analysis and study of data analysis libraries for working with gene expression data, as well as machine learning libraries and available technologies for software implementation of the algorithm.