МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ
Кафедра системного анализа и компьютерного моделирования

ДЕМИДИК Мария Александровна

КЛАССИФИКАЦИЯ ТИПОВ И СТАДИЙ ОНКОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ ПО БИОМЕДИЦИНСКИМ ГЕНОМНЫМ ДАННЫМ

Аннотация к магистерской диссертации

специальность 1-98 80 01 «Информационная безопасность»

Научный руководитель:
кандидат физико-математических
наук, доцент Н.Н. Яцков

Консультант:
кандидат физико-математических
наук, П.В. Назаров

Допущена к защите
«___»_____ 2022 г.
Зав. кафедрой системного анализа и
компьютерного моделирования
кандидат физико-математических наук,
доцент В.В. Скакун

Минск, 2022

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В магистерской диссертации 58 страниц, 26 рисунков, 1 таблица, 57 источников, 2 приложения.

Ключевые слова: сокращение размерности данных, выделение биомаркеров, классификация, экспрессия генов, экспрессия экзонов, биоинформатика, анализ данных, рак легких.

Развитие онкологических заболеваний связано с изменением молекулярных механизмов, которые все еще недостаточно хорошо изучены, особенно на ранних стадиях заболевания. Совершенствование методов информатики и их использование в биологии дает возможность выявлять изменения генома, которые способствует возникновению и развитию разных типов рака, что позволяет лучше понимать причины возникновения заболевания.

Поставлена задача разработки и программной реализации модели предсказания типов и стадий онкологических заболеваний по биомедицинским геномным данным, связывание результатов классификации с качеством предсказания клинических групп.

Предметом исследования выступают данные экспрессии генов и экзонов больных немелкеклеточным раком легких.

Точность классификации подтипов рака легких разработанной моделью на основе полуограниченной машины Больцмана составила 96,1%, а точность определения стадии 69,1%. Выделены информативные признаки на уровне экспрессии генов и экзонов, которые могут являться биомаркерами ранней стадии развития онкологического заболевания.

Задача решалась с помощью языка программирования R в операционной системе Windows 10.

АГУЛЬНАЯ ХАРАКТАРЫСТЫКА ПРАЦЫ

У магістарскай дысертацыі 58 старонак, 26 малюнкаў, 1 табліца, 57 крыніц, 2 дадатка.

Ключавыя словы: скарачэнне памернасці дадзеных, вылучэнне біямаркераў, класіфікацыя, экспрэсія генаў, экспрэсія экзонаў, біяінфарматыка, аналіз дадзеных, рак лёгкіх.

Развіццё анкалагічных захворванняў злучана са зменай малекулярных механізмаў, якія ўсё яшчэ нядосыць добра вывучаны, асабліва на ранніх стадыях захворвання. Удасканаленне метадаў інфарматыкі і іх выкарыстанне ў біялогіі дае магчымасць выяўляць змены геному, якія спрыяюць узнікненню і развіццю розных тыпаў раку, што дазваляе лепш разумець прычыны ўзнікнення захворвання.

Пастаўлена задача распрацоўкі і праграмнай рэалізацыі мадэлі прадказання тыпаў і стадый анкалагічных захворванняў па біямедыцынскіх геномных дадзеных, звязванне вынікаў класіфікацыі з якасцю прадказання клінічных груп.

Прадметам даследавання выступаюць дадзеныя экспрэсіі генаў і экзонаў хворых немелкаклеткавым ракам лёгкіх.

Дакладнасць класіфікацыі падтыпаў раку лёгкіх распрацаванай мадэллю на аснове паўабмежаванай машыны Больцмана склала 96,1%, а дакладнасць вызначэння стадыі 69,1%. Вылучаны інфарматыўныя прыкметы на ўзроўні экспрэсіі генаў і экзонаў, якія могуць з'яўляцца біямаркерамі ранняй стадыі развіцця анкалагічнага захворвання.

Задача вырашалася з дапамогай мовы праграмавання R у аперацыйнай сістэме Windows 10.

GENERAL WORK DESCRIPTION

There are 58 pages, 26 figures, 1 table, 57 sources, 2 appendixes in this master's thesis.

Key words: feature selection, biomarker extraction, classification, gene expression, exon expression, bioinformatics, data analysis, lung cancer.

The development of cancer diseases is associated with a change in molecular mechanisms the understanding of which is still a challenging task, especially in the early stages of the disease. Improving informatics methods and their use in biology makes it possible to detect changes in the genome that contribute to the development of different types of oncology, which leads to revealing the causes of the disease.

The goal of the master's thesis is to develop and implement a model for predicting the types and stages of cancer diseases using biomedical genomic data, as well as to determine the quality of predicting clinical groups .

The subject of the study is gene and exons expression of the non-small cell lung cancer.

The classification accuracy of lung cancer subtypes by the developed model based on the semi-restricted Boltzmann machine is 96.1%, and the accuracy of staging classification is 69.1%. Informative features, which can be biomarkers of the early stage of cancer development, are selected from gene and exon expression.

All the analysis is performed under the R statistical environment and the Windows 10 operating system.